

## Chapter 1

# IDENTIFYING IMPORTANT EXPLANATORY VARIABLES FOR TIME-VARYING OUTCOMES

Oliver Bembom, Maya L. Petersen, Mark J. van der Laan

*Division of Biostatistics*

*University of California, Berkeley*

bembom@berkeley.edu, mayaliv@berkeley.edu, laan@stat.berkeley.edu

**Abstract** This chapter describes a systematic and targeted approach for estimating the impact of each of a large number of baseline covariates on an outcome that is measured repeatedly over time. These variable importance estimates can be adjusted for a user-specified set of confounders and lend themselves in a straightforward way to obtaining confidence intervals and  $p$ -values. Hence, they can in particular be used to identify a subset of baseline covariates that are the most important explanatory variables for the time-varying outcome of interest. We illustrate the methodology in a data analysis aimed at finding mutations of the human immunodeficiency virus that predict how well a patient responds to a drug regimen containing the two antiretroviral drugs lamivudine and stavudine. The most significant mutation we identify, 184IV, has previously been characterized as conferring high-level resistance to lamivudine. Our analysis furthermore points to a second mutation, 75AIMTS, that has been linked to moderate resistance to both lamivudine and stavudine.

**Keywords:** Time series, variable importance, multiple testing, HIV drug resistance

## 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 which follows a cohort of HIV-positive patients over time, 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 identi-

fyng 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 the time course of 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.

To tackle this problem, we first need to define precisely what we mean by “the impact of a mutation on the time course of viral load”. For this purpose, let us denote the collection of candidate mutations by  $A = (A_1, \dots, A_p)$ , with  $A_j = 1$  if a specific amino acid substitution is present at the given position and  $A_j = 0$  otherwise. Let  $Y(t)$  denote a patient’s viral load measured at time  $t$ . Suppose we also measure a number of clinical covariates  $C = (C_1, \dots, C_q)$  at baseline that tend to be associated with the occurrence of particular mutations and that independently affect a patient’s virologic response.

The simplest way of assessing the impact of a particular mutation  $A_j$  on  $Y(t)$  would now be to compare the virologic response among patients with  $A_j = 1$  to that among patients with  $A_j = 0$ . 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 and the clinical covariates  $C$ . The mutation  $A_j$  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  $A_j$  on  $Y(t)$  is *confounded* by the clinical baseline covariates  $C$ , our results do not generalize to a new population in which  $A_j$  and  $C$  are related to each other in a different way.

We might thus be interested in estimating the impact of  $A_j$  on  $Y(t)$  that is not due to associations of  $A_j$  with any of the baseline covariates  $C$ . 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  $A_j$ , *holding their clinical covariates  $C$  fixed at their current values*, as opposed to the scenario in which we give none

of the patients this mutation, holding again  $C$  fixed? Any observed difference could then not be due to differences of the two populations with regard to  $C$  and would thus be more likely to generalize to a new population in which  $A_j$  and  $C$  may be related to each other differently.

To appreciate the difference between the estimates obtained in this way and those described earlier, consider the ideal experiment that would correspond to these earlier estimates. If we simply compare patients with  $A_j = 1$  to those with  $A_j = 0$ , we would be asking: What difference in virologic response would we observe if we gave every patient in our study population the mutation  $A_j$ , *allowing their clinical covariates to take on values that are typical for patients with  $A_j = 1$* , as opposed to the scenario in which we give none of the patients this mutation, again allowing  $C$  to take on typical values? If we now encounter a patient from a new population, the typical values of  $C$  for patients with  $A_j = 1$  that we observed in our original study population may not correspond to typical values of  $C$  for such patients in this new population.

In the hypothetical experiment in which we control for  $C$  by holding it fixed at its observed values, any other covariates that are not included in  $C$  are implicitly allowed to take on values typical for the value of  $A_j$  we are considering. In particular, some of the remaining mutations may be strongly correlated with  $A_j$  so that they would be likely to change their values if we assigned every patient  $A_j = 1$  or  $A_j = 0$ . If these other mutations are now themselves independently related to  $Y(t)$ , our estimates for the impact of  $A_j$  on  $Y(t)$  may not translate well to a new population in which the mutations tend to occur in somewhat different patterns. Only if we adjust for all confounders of the relationship between  $A_j$  and  $Y(t)$ , i.e. all covariates that are associated with  $A_j$  and that are functionally related to  $Y(t)$ , can we be sure that our estimates will be applicable to a new population of patients. If we do so, we are in fact estimating the *causal* impact of  $A_j$  on  $Y(t)$ , rather than a mere association between  $A_j$  and  $Y(t)$ .

We note, however, that estimates of the impact of  $A_j$  on  $Y(t)$  may be interesting and meaningful even if we are not in the ideal situation of being able to adjust for all relevant confounders. We can still identify mutations that are strongly associated with  $Y(t)$  and that may thus allow us to predict a new patient's virologic response to a particular drug regimen, assuming that this patient comes from a population in which the unmeasured confounders are associated with  $A_j$  in a manner that is not too dissimilar from that observed in our study population. Depending on the nature of the unmeasured confounders, this assumption may not be at all unreasonable.

In this chapter, we describe an approach that allows us to estimate such measures of variable importance for any set of baseline covariates we may wish to adjust for. Mathematically speaking, these methods allow us to estimate the parameter

$$\Psi_j(t) = E\left[E[Y(t) \mid A_j = 1, W_j] - E[Y(t) \mid A_j = 0, W_j]\right] \quad (1.1)$$

for each  $j$  and each  $t$ , where  $W_j = (W_j^1, \dots, W_j^m)$  is the desired set of adjustment variables. The estimates we obtain rely on a minimum number of assumptions and in some cases are as precise as possible. Furthermore, the approach provides an honest measure of the statistical significance of each estimate. For a more rigorous treatment, the interested reader is referred to the article by van der Laan, 2006b.

## 2. Basic Concepts

In this section, we describe how the variable importance parameter  $\Psi_j(t)$  is estimated in practice. The central step consists of transforming the recorded data for each observation into a quantity whose expectation equals  $\Psi_j(t)$ . Conceptually, these quantities can be thought of as giving a measure of the impact of  $A_j$  on  $Y(t)$  as derived from a single observation. We can then estimate the entire function  $\Psi_j(t)$  relating variable importance to time by fitting a statistical model for how the expectation of the transformed quantities depends on time, i.e. by *regressing* them on time.

We will describe three different transformations of the observed data that are suitable for our purposes. These transformations themselves involve parameters that are generally not known by the researcher and hence must be estimated from the observed data. Since these parameters are not of interest in themselves, we refer to them as *nuisance parameters*. The first of these nuisance parameters is the so-called treatment mechanism. The variable  $A_j$  whose effect on  $Y(t)$  we would like to estimate is often referred to as the treatment variable. The treatment mechanism  $g_j$  now gives the probability of observing a given treatment  $A_j = a$  for a subject with a particular covariate profile  $W_j$ :

$$g_j(a, W_j) \equiv P(A_j = a \mid W_j) \quad (1.2)$$

The second nuisance parameter consists of a regression of  $Y(t)$  on  $A_j$  and  $W_j$ :

$$Q_j(a, t, W_j) \equiv E[Y(t) \mid A_j = a, W_j] \quad (1.3)$$

To estimate  $g_j$  and  $Q_j$  we ideally do not want to rely on the assumption of a particular functional form. For example, we would like to avoid an

assumption such as that the expectation of  $Y(t)$  given  $A_j$  and  $W_j$  can be written as

$$E[Y(t) | A_j, W_j] = \beta_{j0} + \beta_{j1}A_j + \beta_{j2}t + \beta_{j3}W_j^1 + \dots + \beta_{j(m+2)}W_j^m \quad (1.4)$$

for some coefficient vector  $\beta_j = (\beta_{j0}, \dots, \beta_{j(m+2)})$ . On the basis of biological knowledge alone it is very difficult to arrive at an appropriate functional form, and poorly specified functional forms can lead to severely biased estimates of variable importance. Thus, the functional form of the nuisance parameter models should be chosen based on the information that is contained in the dataset, i.e. data-adaptively. One popular approach to this model selection problem is the D/S/A algorithm (Sinisi and van der Laan, 2004) that relies on deletion, substitution, and addition moves to search through a large space of possible functional forms. Another well-known model selection technique that searches through a much smaller space of candidate functional forms and thus requires somewhat less computing time has been introduced by Kooperberg et al., 1997.

In modelling  $Q$ , we have to bear in mind that repeated viral load measurements on the same patient will be correlated. The usual generalized linear models were formulated for outcomes that are independent of each other (McCullagh and Nelder, 1989). In the presence of correlated outcomes, these models provide estimates that, while still reliable, are no longer as precise as possible. Furthermore, the  $p$ -values they provide for testing whether or not certain regression coefficients are equal to zero cannot be trusted since they tend to be too small.

Generalized estimating equations address both of these issues and have thus been a popular tool for modelling correlated outcomes (Liang and Zeger, 1986; Zeger and Liang, 1986). Unlike generalized estimating equations, the D/S/A algorithm for data-adaptive model selection does not explicitly take into account the correlation between outcomes  $Y(t)$  measured on the same subject. Although the model fits obtained by this algorithm are not as precise as possible, the estimates are still reliable and useful for the purpose of model selection. Furthermore, the regression of  $Y(t)$  on  $A_j$  and  $W_j$  is only a nuisance parameter, needed to estimate the variable importance of  $A_j$ , but is not of primary interest in itself. Thus, we are not interested in testing whether some coefficients in the selected model might be equal to zero or not. The only adjustment that is necessary when using the D/S/A algorithm in this situation is to supply an ID variable that can be used to identify the independent experimental units. This allows the algorithm to carry out an honest cross-validation procedure by assigning measurements from the same subject to the same validation sample. The approach by Kooperberg et

al. can be used for modelling the treatment mechanism, but cannot be extended in a straightforward way for model selection in the context of correlated outcomes.

Given estimates  $g_{j,n}$  and  $Q_{j,n}$  of the nuisance parameters  $g_j$  and  $Q_j$ , we can now generate three different types of transformations of the observed data. Let  $T_{i,k}$  denote the time at which the  $k$ th measurement of the outcome  $Y$  was obtained for subject  $i$ . For each outcome measurement  $Y_{i,k}(T_{i,k})$  for subject  $i$ , we will obtain one transformed observation  $D_{i,k}^j$ . The regression-based transformation only makes use of the nuisance parameter  $Q_j$  and is given by

$$D_{i,k}^j(T_{i,k}) = Q_{j,n}(1, T_{i,k}, W_{j,i}) - Q_{j,n}(0, T_{i,k}, W_{j,i}) \quad (1.5)$$

The inverse-probability-of-treatment-weighted (IPTW) transformation only makes use of the nuisance parameter  $g_j$  and is given by

$$D_{i,k}^j(T_{i,k}) = \left\{ \frac{I(A_{j,i} = 1)}{g_{j,n}(1, W_{j,i})} - \frac{I(A_{j,i} = 0)}{g_{j,n}(0, W_{j,i})} \right\} Y_{i,k}(T_{i,k}) \quad (1.6)$$

where  $I(\cdot)$  is the indicator function that equals one if the condition in parentheses is true and zero otherwise. The double robust transformation, finally, makes use of both nuisance parameters and is given by

$$\begin{aligned} D_{i,k}^j(T_{i,k}) = & Q_{j,n}(1, T_{i,k}, W_{j,i}) - Q_{j,n}(0, T_{i,k}, W_{j,i}) + \\ & \left\{ \frac{I(A_{j,i} = 1)}{g_{j,n}(1, W_{j,i})} \left( Y_{i,k}(T_{i,k}) - Q_{j,n}(1, T_{i,k}, W_{j,i}) \right) - \right. \\ & \left. \frac{I(A_{j,i} = 0)}{g_{j,n}(0, W_{j,i})} \left( Y_{i,k}(T_{i,k}) - Q_{j,n}(0, T_{i,k}, W_{j,i}) \right) \right\} \quad (1.7) \end{aligned}$$

Since both the regression-based and IPTW transformation of the data only rely on one of the two estimated nuisance parameters, variable importance estimates based on these transformations will only be reliable if the relevant nuisance parameter is estimated well. The double robust transformation, however, relies on both nuisance parameters and has the remarkable property that it yields correct estimates of variable importance if either one of these two nuisance parameters is estimated well.

We now obtain three different estimates of the variable importance parameter  $\Psi_j(t)$  by regressing each of these transformed observations on time  $t$ . As above, the use of data-adaptive model selection approaches such as the D/S/A algorithm for fitting these regressions can help to minimize the reliance on assumptions about how variable importance varies

as a function of time. The same approach can be used to obtain estimates of the variable importance of a covariate  $A_j$  conditional on some subset  $V$  of the adjustment covariates  $W_j$ . For example, a researcher may be interested in the impact of a given mutation on the time course of viral load conditional on the viral load at baseline. Mathematically speaking, this corresponds to the parameter

$$\Psi_j(t, v) = E\left[E[Y(t) \mid A_j = 1, W_j] - E[Y(t) \mid A_j = 0, W_j] \mid V = v\right] \quad (1.8)$$

Such parameters are straightforward to estimate once we have created the transformed observations described above by simply regressing them on both  $t$  and  $V$  rather than on  $t$  alone.

Plots of the estimated variable importance  $\Psi_{j,n}(t)$  as a function of time can now be used to explore how the impact of  $A_j$  on  $Y(t)$  changes over time. Such plots can furthermore be used as inputs for a clustering algorithm to identify treatment variables  $A_j$  whose impact on  $Y(t)$  develops according to a similar dynamic over time. Alternatively, we may be interested in testing the hypotheses that the importance of treatment variable  $A_j$  is zero at some set of time points  $t_1, \dots, t_d$ , with the goal of identifying treatment variables  $A_j^*$  and time points  $t^*$  for which we have strong evidence against such hypotheses. For this purpose, we use the following bootstrap approach to first obtain separate  $p$ -values  $p_{j,1}, \dots, p_{j,d}$  for the respective hypotheses that  $\Psi_j(t_1), \dots, \Psi_j(t_d)$  equal zero. We draw a large number of samples of size  $n$  with replacement from the pool of  $n$  subjects in our dataset to obtain bootstrap datasets that contain all outcome measurements  $Y(t)$  for the selected subjects. For each of these bootstrap samples we now repeat the entire estimation process as outlined above to arrive at bootstrap estimates  $\Psi_{j,n}^\#(t_1), \dots, \Psi_{j,n}^\#(t_d)$  of the desired variable importance measures. If we have used a computationally involved model selection technique like the D/S/A algorithm to regress  $A_j$  on  $W_j$ ,  $Y(t)$  on  $A_j$  and  $W_j$ , or  $D(t)$  on  $t$ , we may avoid the model selection step as part of this bootstrap process and simply refit the regressions according to the selected functional form. This approach saves a significant amount of time and generally leads to  $p$ -values that are only slightly optimistic.

We can now take the variance of these bootstrap estimates as an estimate of the variance of  $\Psi_{j,n}(t_k)$  under the null hypothesis that  $\Psi_j(t_k) = 0$  and form  $t$ -statistics by dividing  $\Psi_{j,n}(t_k)$  by the square root of this estimated variance. Under the null hypotheses, these test statistics will be closely approximated by a standard normal distribution once we have a reasonable sample size. The desired  $p$ -value can thus be obtained as the probability that the absolute value of a standard normal variate exceeds the absolute value of the observed test statistic. Each of these  $p$ -values

gives an estimate of the proportion of times we would reject a true null hypothesis if the experiment and corresponding hypothesis test were to be performed over and over again. The  $p$ -values are formulated for individual hypothesis tests and thus do not take into account that we are testing several hypotheses simultaneously.

A large number of methods exist for obtaining  $p$ -values that are interpretable in this context of multiple testing. (Dudoit and van der Laan, 2006; Lehmann and Romano, 2005; Westfall and Young, 1993). Among the most straightforward methods are those that simply transform the raw  $p$ -values obtained from the individual hypothesis tests into a set of adjusted  $p$ -values. The well-known Bonferroni adjustment, for example, simply multiplies each  $p$ -value by the number of comparisons that are made (Bland and Altman, 1995). The adjusted  $p$ -values obtained in this manner estimate the proportion of times we would falsely reject at least one true null hypothesis if we repeatedly carried out a test according to which we reject all hypotheses with adjusted  $p$ -values smaller than some cut-off. Another popular method by Benjamini and Hochberg instead produces adjusted  $p$ -values that estimate the false discovery rate, i.e. the expected proportion of true null hypotheses among all hypotheses that are rejected (Benjamini and Hochberg, 1995).

### 3. Advantages and Disadvantages

#### Advantages

- The methodology described in this chapter starts with a clear definition of what is meant by the impact of a treatment variable  $A_j$  on the outcome process  $Y(t)$ . With this definition in hand, the corresponding parameter can then be estimated separately and directly for each candidate treatment variable. In contrast to other approaches, we do not have to derive these variable importance estimates from a regression of  $Y(t)$  on the complete set of treatment variables  $A_1, \dots, A_p$  and potential confounders that was fitted with the goal of accurately predicting  $Y(t)$  rather than with the goal of estimating the importance of a particular  $A_j$ .
- The definition of variable importance described here is quite flexible since the user can decide which variables are to be included in the adjustment covariates  $W_j$ . On one extreme, we can estimate unadjusted variable importance by leaving  $W_j$  empty. This would allow us to identify variables that are useful for predicting  $Y(t)$  *in our study population*, but not necessarily in a new population of subjects. On the other extreme, we may be able to adjust for all

relevant confounders of the relationship between  $A_j$  and  $Y(t)$ , in which case we will obtain estimates of the *causal* effect of  $A_j$  on  $Y(t)$ . In addition, the treatment variables whose impact on  $Y(t)$  we would like to estimate can be any extractions of the available baseline covariates. In particular, they can consist of continuous rather than binary variables as in the mutation example considered here. A slightly more complicated example of an interesting extraction is given by cross-product terms like  $A_1 \times A_2$  or linear combinations of baseline covariates. Furthermore, we may be interested in the importance of a multivariate treatment variable  $\mathbf{A}_j = (A_{j1}, A_{j2})$ , which would allow us, for example, to study the impact of the simultaneous presence of two mutations on virologic response.

- The targeted estimation of the impact of  $A_j$  on  $Y(t)$  allows us in a straightforward manner to obtain measures of statistical significance such as confidence intervals or  $p$ -values. Other methods generally do not provide these and thus do not allow us to distinguish between variables whose importance is genuinely different from zero and those whose importance is in fact zero but is estimated to be non-zero due to sampling variation.
- The methodology described here aims to be as robust as possible, i.e. it aims to rely on as few assumptions as possible. In particular, it avoids the assumption of knowing the functional form of the nuisance parameters  $g$  and  $Q$  *a priori*. Methods that rely on this assumption can yield severely biased estimates of variable importance if the functional form is guessed incorrectly. This is particularly important in the context of modern genomics and proteomics applications in which the number of variables that might be included in these models is very large, making it virtually impossible to guess the correct model.
- We have described three different transformations of the observed data that each give a different estimate of variable importance. This is valuable since these three transformations differ in how they rely on the two nuisance parameters  $g$  and  $Q$  so that they can be expected to succeed in different situations. In the setting of a clinical trial, for example, the treatment mechanism  $g$  is generally known or straightforward to estimate so that variable importance estimates based on the IPTW transformation can be expected to be very reliable.

- Under certain assumptions the variable importance estimates based on the double robust transformation are as precise as possible, meaning, for example, that it would be impossible to obtain more narrow confidence intervals. Specifically, this is the case if both nuisance parameters are estimated reliably and at a fast enough rate (van der Laan, 2006b).

## Disadvantages

- Data-adaptive model selection techniques such as the D/S/A algorithm that search through a large space of candidate functional forms are computationally intensive and do not scale well with a growing number of candidate variables to choose from. Hence, it may be necessary to first reduce the number of these candidate variables by, for example, eliminating those that are not associated with  $Y(t)$  in univariate regression models. Such variables are unlikely to confound the relationship between the treatment variables and  $Y(t)$  and also would add little to the precision with which we could estimate variable importance measures. Alternatively, we may resort to a less exhaustive model selection algorithm such as the one introduced by Kooperberg et al., at least for the estimation of  $g$  where repeated-measures regression is not needed.
- The computational burden of the D/S/A algorithms often makes it difficult to carry out a completely honest bootstrap simulation to estimate the variance of our point estimates. This would require that we repeat the data-adaptive model selection procedure for each bootstrap sample. Due to time constraints, however, we are often forced to treat the selected functional forms for  $g$  and  $Q$  as given for the purposes of the bootstrap by simply refitting the selected models for each bootstrap sample. This approach ignores the extra variability of our estimates that is introduced by the model selection procedure and thus tends to underestimate their variance somewhat. The  $p$ -values obtained in this way are still useful, however, for the purpose of ranking the treatment variables in order of statistical significance. In practice, the variance estimates obtained by ignoring the data-adaptive model selection process are also often not too different from those obtained from a completely honest bootstrap.
- The methodology described here is relatively new and thus has not yet been implemented in the form of a publicly available software package. As described in more detail below, however, the individual steps that are required are fairly straightforward to carry out

in a modern statistical computing environment like R (R Development Core Team, 2005).

#### 4. Caveats and Pitfalls

The IPTW transformation of the observed data relies crucially on an estimate of the probability that a given subject would have received his or her observed treatment. It weights each observation by the inverse of this probability, thus downweighting observations that were likely to have received their observed treatment and upweighting those that were instead unlikely to have been observed with the treatment we recorded for them. This essentially creates a new sample in which treatment assignment is independent of the baseline covariates, making it straightforward to estimate the impact of treatment  $A_j$  on the outcome, *controlling for*  $W_j$ , by simply comparing the two groups with  $A_j = 0$  and  $A_j = 1$ . This approach breaks down if for certain values of  $W_j$  we never observe one of the two treatment values  $A_j = 0$  or  $A_j = 1$ . In that case we cannot use weighting to create a new sample in which  $A_j$  is independent of  $W_j$  since the new sample will still not contain any observations with that value of  $W_j$  and the missing value of  $A_j$ . Variable importance estimates based on the IPTW transformation thus also rely on the so-called Experimental Treatment Assignment (ETA) assumption which states that there are no values of  $W_j$  for which treatment is assigned in a deterministic fashion. In fact, IPTW-based estimates also perform poorly if the ETA assumption is practically violated, i.e. if for some values of  $W_j$ , treatment is assigned in a *nearly* deterministic fashion (Neugebauer and van der Laan, 2005).

We can examine the extent to which the ETA assumption is violated in a number of *ad hoc* ways. We may, for example, look at the proportion of observations for which the probability of having received their observed treatment is very close to zero or one, say less than 0.05 or greater than 0.95. Such observations would hint at values of  $W_j$  for which there is very little experimentation with respect to treatment assignment. To look more closely at the relationship between  $W_j$  and these fitted probabilities we may also plot the probabilities against the linear combination of  $W_j$  that was chosen for the treatment model, or examine observed counts of assigned treatments  $A_j$  within deciles of that linear combination.

Current research in this area is investigating variable importance measures that are based on slightly different ideal experiments than those described above (van der Laan, 2006a). Instead of considering hypothetical scenarios in which each member of the population is assigned a particular treatment like  $A_j = 1$ , these efforts focus on so-called dy-

dynamic treatment rules that assign  $A_j = 1$  to those members for which this assignment is sensible but  $A_j = 0$  to the remaining ones. If for certain treatment histories, for example, it is impossible or very unlikely to observe a particular mutation in the virus of a patient, such rules would never assign such a patient to have this mutation. Treatment rules that are realistic in this sense then no longer rely on the ETA assumption.

The double robust and regression-based transformations rely on an estimate of the regression  $E[Y(t) \mid A_j, W_j]$ . As mentioned above, we would like to avoid the assumption that the functional form for the dependence of  $Y(t)$  on  $A_j$  and  $W_j$  is known *a priori* by using data-adaptive model selection techniques. The models selected in this way, however, often contain neither the treatment variable  $A_j$  nor any interaction terms between  $A_j$  and time, especially in genomics or proteomics applications with a large number of candidate explanatory variables to choose from. Such models are unsatisfactory since they do not allow us to examine the impact of  $A_j$  on  $Y(t)$  and the change of this impact over time. If we use the regression-based transformation, such models will in fact directly translate into an estimate of zero variable importance for  $A_j$ .

To explicitly acknowledge that we are interested in estimating the effect of  $A_j$  on  $Y(t)$  over time, we might hence fit two separate data-adaptive regression models, one among subjects with  $A_j = 0$  and one among subjects with  $A_j = 1$ . This is problematic, however, for the following reason. Suppose the adjustment variables  $W_j$  contain an important confounder that is very strongly correlated with  $A_j$  and that has an independent effect on  $Y(t)$ . Then clearly this variable should be included in a model predicting  $Y(t)$  from  $A_j$  and  $W_j$  to adequately control for confounding by  $W_j$ . Within groups defined by  $A_j$ , this variable will show very little variation, however, and thus will contribute little to the accurate prediction of  $Y(t)$ . Model selection procedures are thus unlikely to include this variable in the chosen regression model.

We therefore recommend the following two-step approach: First, fit a data-adaptive regression model for the expectation of  $Y(t)$  given  $W_j$  alone, *excluding  $A_j$  from the set of candidate explanatory variables*. Then fit a second data-adaptive regression model that is forced to contain all the terms of the first model along with the terms  $A_j$  and  $A_j \times t$ . The first step guarantees that no important confounders are omitted due to strong correlations with  $A_j$ . The second step then allows the model selection algorithm to add interaction terms between  $A_j$ ,  $A_j \times t$ , and the baseline covariates selected for the first model.

In our description of the data structure we have assumed that all subjects are followed up for the entire duration of the study. This assumption is often not met, with subjects dropping out of the study for

various reasons such as moving away or being switched to a new drug regimen due to poor response to the current regimen. The methodology we have described so far will still give reliable estimates of variable importance if such loss to follow-up is not related to what a subject's future outcomes would have been in the two hypothetical scenarios in which each subject is either given the treatment or not. This is, for example, reasonable in the case of subjects moving away since the decision to move away is probably not influenced by what the future outcomes  $Y(t)$  might have been. If patients are switched to different drugs due to poor response, however, we will systematically be missing patients that would have had a poor outcome, had it been observed. In the presence of such informative censoring, the estimation procedure described above provides estimates of the importance of  $A_j$  among the non-representative subgroup for whom the outcome was measured, which generally differs from its importance in the entire study population.

This problem can be addressed by weighting each observation by the inverse of the probability that the subject was not censored by the time the measurement was made. Such inverse-probability-of-censoring weights work analogously to those used as part of the IPTW transformation in that they artificially create a sample in which censoring is independent of any confounders we would like to adjust for. In practice the needed probabilities can be estimated by modelling the probability of being censored at each time point  $t$  given  $A_j$  and  $W_j$ , using, for example, a Cox proportional hazards model if we treat time as a continuous variable or a pooled logistic regression model if outcome measurements were made at pre-determined intervals for each subject.

An analogous approach can also be taken if the treatment variable  $A_j$  or some of the outcome measurements  $Y(t)$  are missing for some of the subjects. This could, for example, be the case if some patients never had their genotype measured. As before, we need to address this type of missingness if it is related to what a subject's future outcomes would have been in the two hypothetical scenarios in which each subject is either given the treatment or not. In this case we would use a logistic regression model to estimate the probability that the variable of interest is recorded for a particular subject given what we have observed so far on this person. The observations with available measurements are then weighted by the inverse of these estimated probabilities.

## 5. Alternatives

Many applications in statistics and biology have been concerned with estimating the impact of a number of treatment variables  $A_j$  on an

outcome process  $Y(t)$ . Several methods are commonly used for this purpose. Researchers who wish to estimate unadjusted variable importance frequently use generalized estimating equations to regress  $Y(t)$  on  $A_j$  according to simple models such as

$$E[Y(t) | A_j] = \beta_0 + \beta_1 A_j + \beta_2 t + \beta_3 A_j \times t \quad (1.9)$$

The null hypothesis of  $A_j$  having no impact on  $Y(t)$  is then equivalent to the hypothesis that both  $\beta_1$  and  $\beta_3$  are equal to zero, which is straightforward to evaluate using the standard error estimates provided by generalized estimating equations. Conclusions drawn in this way, however, rely on the assumption that the expectation of  $Y(t)$  given  $A_j$  can in fact be written according to such a simple functional form. If this is not the case, variable importance estimates may be severely biased. This problem becomes even more pressing once we wish to adjust for a number of baseline covariates  $W_j$ . If the number of such baseline covariates is large and we include each of them as a simple main-effects term in the model, we will be virtually guaranteed to have mis-specified the functional form according to which the expectation of  $Y(t)$  depends on  $A_j$  and  $W_j$ .

The bias incurred due to such model mis-specification has motivated the use of data-adaptive model selection techniques like the D/S/A algorithm or classification and regression trees (Breiman et al., 1984). In general, researchers include all treatment variables of interest along with the set of potential confounders in the pool of candidate explanatory variables from which the model selection algorithm is then allowed to select a subset of variables for inclusion in the final regression model. In spite of considering a large number of quite complex candidate models, such data-adaptive algorithms frequently end up selecting a model that contains only a relatively small number of covariates. Such models are disappointing for the purpose of estimating variable importance since they do not give us an explicit estimate of the importance of those covariates that are not selected by the algorithm. We can only conclude that these covariates have no impact on  $Y(t)$  at all.

This issue is commonly addressed through a resampling-based technique known as **bootstrap-aggregating** or bagging (Breiman, 1996) that is based on re-fitting the data-adaptive regression models on a large number of bootstrap samples drawn from the original data set and then averaging out the coefficient estimates for each variable across all these regression fits. Since different bootstrap samples typically result in the selection of a different set of variables, this approach allows us to obtain non-zero variable importance estimates for a much larger number of variables.

Neither variable importance estimates based on a single data-adaptive regression fit nor those based on bagging lend themselves to an assessment of their statistical significance. This major drawback can be ascribed to the fact that these methods are not designed for the specific purpose of estimating the impact of a number of treatment variables  $A_j$  on  $Y(t)$  but rather for the purpose of accurately predicting  $Y(t)$ . Measures of variable importance are only obtained in a secondary step, as a derivative of the estimated regression fit. This is in stark contrast to the methods described in this chapter that are targeted directly at estimating the importance of each separate treatment variable  $A_j$ . It is precisely this targeted nature of the variable importance estimates described here that allows us to assess their statistical significance in such a straightforward way. This observation underscores the need to separate the statistical problem of accurately predicting the outcome  $Y(t)$  from that of assessing the importance of each of the treatment variables  $A_j$ . In many instances we will be interested in the first problem, in which case data-adaptive regression fits obtained, for example, by classification and regression trees represent some of the most powerful tools currently available. If we are interested in estimating variable importance, however, the targeted methods described in this chapter offer many advantages that make them the approach of choice.

## 6. Case Study: HIV drug resistance mutations

In this section we apply the methodology described above to the task of identifying HIV mutations that modulate how well the virus can replicate in the presence of a particular combination of antiretroviral drugs, and thus how well a patient responds to that drug regimen. 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, 2002). 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.

In order to estimate the causal effect of a given mutation of interest, we would ideally also adjust for the presence of additional mutations. As with treatment history, this would help to ensure that the association we observe between a given mutation and the outcome is indeed causal, rather than due to the effect of other mutations that occur frequently with the mutation of interest. Estimation of such a causal effect is desirable not only from the point of view of mechanistic understanding, but also because it is not dependent on population characteristics such as past treatment patterns that would limit the extent to which it might translate to other HIV-infected populations.

Mutations conferring resistance to drugs of a class different from that targeted by the mutation of interest, thus affecting a distinct viral enzyme, can indeed be controlled for by simply including them in  $W_j$ . However, mutations conferring resistance to the same drug class, thus affecting the same viral enzyme, are often correlated to the extent that it is not possible to distinguish which mutation is causally responsible for a given effect. 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. Hence, certain mutations are essentially never observed in the absence of another mutation, making it next to impossible to disentangle the individual impacts of these two mutations on virologic response. The statistical consequence of this correlation or collinearity between individual mutations lies in considerable instability

of the variable importance estimates we might obtain if we included the other mutations in the same viral enzyme in the group of adjustment variables  $W_j$ . Attempting to do so would also cause a severe violation of the ETA assumption since the presence of one mutation might virtually guarantee the presence of the mutation whose importance we are trying to estimate. These considerations suggest that we should not adjust our variable importance estimates for the other mutations in the same viral enzyme.

The dataset we use is derived from the Stanford HIV drug resistance database, a patient sample drawn from 16 Kaiser Permanente Northern California clinics for which longitudinal data on HIV reverse transcriptase and protease sequences, antiretroviral treatment, and viral load were recorded. From this database, we identified episodes during which a patient who has failed a previous drug regimen is followed under a new regimen in which at least one of the drugs has been changed. We require that the patient has a baseline viral load measurement available that was taken no more than 24 weeks before initiation of the new treatment. For such records we obtained all viral load measurements that were taken in the 24 weeks following the treatment change. After about 24 weeks, clinicians may switch patients to yet another drug regimen if they do not appear to be responding well to the current salvage therapy regimen. By restricting ourselves to viral load measurements taken before this time point, we avoid having to adjust for the bias introduced into our variable importance estimates by this informative loss to follow-up.

We would like to identify mutations that modulate virologic response to drug regimens that contain the two NRTI drugs lamivudine and stavudine and thus limit ourselves to patients whose salvage therapy regimen contains these drugs. To isolate mutations specific to these two drugs, we exclude patients who are also taking other NRTI drugs. Since mutations thought to confer resistance to NRTI drugs are unlikely to affect susceptibility to PI or NNRTI drugs, we make no requirements as to which drugs of these two classes might be included in the patient's regimen. However, we do control for these covariates in our analyses, as the presence of an NRTI mutation can be associated with the potency of the non-NRTI drugs in the regimen, which in turn can independently affect virologic response. We exclude patients that have never taken an NRTI drug before since they are virtually guaranteed not to have any of the mutations thought to confer resistance to NRTI drugs. Including this group of patients in our analysis would thus cause a severe violation of the ETA assumption. Based on these inclusion criteria, our dataset contains 855 viral load measurements from 288 individual treatment change episodes. These measurements were made on 278 in-

dividual patients, with a small number of them contributing more than one treatment change episode.

We are interested in assessing the impact on virologic response to lamivudine and stavudine for any mutation in the HIV reverse transcriptase gene that has previously been linked to resistance to NRTI drugs. Mutations are coded as  $A_j = 1$  if any of a number of amino acid substitutions potentially related to drug resistance is detected at the given position of the viral enzyme. The mutation 44AD, for example, is considered to be present if either alanine or aspartic acid are found at position 44 of the reverse transcriptase enzyme. For the sake of statistical precision, we only consider mutations that occur at least 15 times among the treatment change episodes we have identified, giving us a total of 14 mutations, whose impact on virologic response we would like to estimate.

We would like to define the outcome  $Y(t)$  as the change in log viral load at time  $t$  as compared to the baseline measurement made before the treatment change. In our dataset, viral loads below  $10^{1.7}$  are not detectable so that viral loads below this threshold are simply recorded as below the limit of detection. Since patients whose viral load becomes undetectable during the course of treatment are considered to respond as well as possible to the new drug regimen, we impute the change in log viral load for such patients by the maximal change in log viral load observed across the entire dataset,  $-4.2$ . We note that this outcome would not be suitable if our goal was to estimate or predict the true change in log viral load resulting from a mutation. However, here, our goal is to estimate the clinical importance of each mutation considered. The outcome definition used thus incorporates the two types of viral response considered a clinical success: 1) a large decrease in viral load, or 2) a final undetectable viral load.

The following variables are used to capture a patient's treatment history: duration of antiretroviral therapy; number of past regimens; history of past PI, NRTI, and NNRTI drug use; number of PI, NRTI, and NNRTI drugs failed in the past; and history of mono/dual therapy. We characterize the current drug regimen through the total number of drugs as well as the number of PI and NNRTI drugs included in that regimen. Furthermore, we have available information about the duration between baseline viral load measurement, sequencing of the virus, and initiation of the salvage therapy regimen. While we do not adjust each of our variable importance estimates for the presence or absence of the 13 other mutations thought to confer resistance to NRTI drugs, we do adjust them for the presence or absence of a number of mutations that have been linked to resistance to PI and NNRTI drugs. Lastly,

we also include Stanford susceptibility scores to the drugs in these two classes that are calculated on the basis of these mutations. This leaves us with a total of 80 baseline covariates that we consider as potential confounders of the relationship between mutations and change in viral load. To reduce the computational burden on the D/S/A algorithm, we reduce the number of baseline covariates to include in  $W_j$  by univariate repeated-measures regression of  $Y(t)$  on each candidate confounder, only keeping those covariates with adjusted  $p$ -values smaller than 0.05. After this initial dimension reduction, the remaining 16 variables considered include the following: the number of past regimens; the number of PI drugs failed in the past; the total number of drugs as well as the number of NNRTI drugs in the new regimen; susceptibility scores for the two NNRTI drugs delavirdine and efavirenz as well as the two PI drugs amprenavir and lopinavir; and four mutations each related to resistance to PI and NNRTI drugs.

We model the treatment mechanisms using the D/S/A algorithm, allowing the algorithm to search through models of up to 10 terms, possibly including products comprised of two candidate confounders. The variables most frequently selected for these treatment models include the number of PI drugs failed in the past, the two mutations 90M and 10FIRV that are related to resistance to PI drugs, as well as susceptibility scores for efavirenz and amprenavir. Judging by the percentage of fitted probabilities smaller than 0.05 or greater than 0.95, the majority of mutations appear to satisfy the ETA assumption, with most of these percentages being no greater than 5%. The three notable exceptions to this trend are given by the mutations 75AIMTS, 74IV, and 44AD for which 73%, 48%, and 38%, respectively, of all fitted treatment probabilities are smaller than 0.05 or greater than 0.95. The IPTW-based variable importance estimates for these three mutations may thus be unreliable.

To estimate the expectation of  $Y(t)$  given  $W_j$  and  $A_j$ , we first let the D/S/A algorithm choose an appropriate functional form for predicting  $Y(t)$  from  $W_j$ . The selected fit includes time  $t$ ; the susceptibility score for lopinavir; the number of PI drugs failed in the past; the total number of drugs as well as the number of NNRTI drugs in the current regimen; and the mutations 10FIRV, 84AV, and 90M. As described above, we would now like to fit a second data-adaptive regression model that is forced to contain all of these terms along with  $A_j$  and  $A_j \times t$ . In this case, however, we can in fact omit the term  $A_j$  based on the following consideration: For 50% of all subjects, baseline viral load was measured within five days, and for 80% of all subjects, it was measured within four weeks of initiation of the new regimen, suggesting that  $Y(0)$ , the change in log

viral load between treatment change and the baseline measurements, is close to zero for the majority of patients. This in turn implies that all variable importance measures should be close to zero at  $t = 0$ , which makes the term  $A_j$  unnecessary.

We described above how we can obtain estimates of the variable importance of each  $A_j$  at a chosen set of time points  $t_1, \dots, t_d$ . This will result in the simultaneous test of  $p \times d$  hypotheses of the form  $\Psi_j(t_k) = 0$ , where  $p$  is the number of treatment variables we are considering. In a first analysis aimed at identifying important explanatory variables for  $Y(t)$  rather than examining how their impact on  $Y(t)$  changes over time, however, it is often useful to obtain a single summary measure for the variable importance of each treatment variable. This reduces the number of simultaneous hypothesis tests that have to be performed and thus increases the chance of obtaining statistically significant results. In the present case, we can again make use of the assumption that variable importance measures at time  $t = 0$  should be close to zero by regressing the transformed observations  $D_{i,k}^j$  on time according to the simple model

$$E[D^j(T)] = \beta_j T \quad (1.10)$$

that does not include an intercept term. This functional form is likely to be too simplistic to fit the actual time course of variable importance very well, but we view it more as a means to obtain an interesting summary measure of this time course rather than as an accurate estimate of the time course itself. In particular, we can expect to find a positive coefficient  $\beta_j$  for those mutations that all in all lead to an increase in viral load and a corresponding negative coefficient for those mutations that all in all lead to a decrease in viral load.

As described above, we obtain unadjusted  $p$ -values for the hypotheses  $\beta_j = 0$  based on a bootstrap estimate of the variance of the estimated coefficient  $\beta_j$ . These  $p$ -values are adjusted using the Benjamini-Hochberg method for controlling the false discovery rate. Table 1.1 summarizes the estimated variable importance of each mutation 24 weeks after treatment change corresponding to the estimate obtained for  $\beta_j$ , along with adjusted  $p$ -values for the hypothesis that this variable importance is equal to zero. We present estimates based on each of the three different transformations of the data. The mutations are ranked in order of statistical significance according to the estimates based on the double robust transformation.

All three estimators identify the mutation 184IV as having the most significant impact on virologic response to treatment with lamivudine and stavudine. This mutation has in fact been shown to be responsible for high-level resistance to lamivudine based on extensive laboratory and

Table 1.1. Variable importance estimates based on the double robust (DR), inverse-probability-of-treatment-weighted (IPTW), and regression-based transformation. Estimates give the impact of a mutation on the change in log viral load after 24 weeks. Mutations marked with \* show a significant violation of the ETA assumption.

Mutation	DR	<i>p</i> -value	IPTW	<i>p</i> -value	Regression	<i>p</i> -value
184IV	0.8307	0.0025	0.7200	0.0364	0.6739	0.0004
75AIMTS*	0.5604	0.2611	0.5772	0.4638	0.6255	0.0917
41L	0.3362	0.4187	0.4821	0.1587	0.3017	0.1111
62V	0.6135	0.4187	0.9594	0.1587	0.5534	0.3367
118I	0.3050	0.4658	0.4605	0.2173	0.2631	0.3367
215FY	0.2421	0.5864	0.3877	0.2426	0.2151	0.3367
67EGN	0.1895	0.6566	0.2617	0.4638	0.2378	0.3367
69DN	0.1811	0.8015	0.3349	0.5810	0.1692	0.6081
74IV*	0.2295	0.8015	0.1466	0.6682	0.3208	0.4507
70RGE	0.0931	0.8783	0.1800	0.6221	0.1910	0.4932
210W	0.0800	0.8783	0.0929	0.6781	0.2114	0.3367
219ENQR	-0.0732	0.8783	-0.2537	0.5810	-0.0066	0.9766
44AD*	0.0417	0.9351	0.4866	0.1587	0.1502	0.6416
215CDEIVS	-0.0041	0.9920	0.2437	0.5810	0.1950	0.6967

clinical data (Boucher et al., 1993; Tisdale et al., 1993; Schurman et al., 1995). Analyses linking HIV mutations directly to *in vitro* drug susceptibility have furthermore identified this mutation as by far the most important mutation conferring resistance to lamivudine (Rhee et al., 2006).

The second most important mutation identified by the double robust and regression-based estimates is given by 75AIMTS. This mutation is ranked much lower based on the IPTW transformation, which as mentioned above, however, can be expected to give unreliable estimates in this case due to a violation of the ETA assumption. 75AIMTS has been shown to confer moderate resistance to the second drug in the regimen we consider, stavudine (Lacey and Larder, 1994). The analyses by Rhee et al. furthermore suggest that this mutation may also be related to drug resistance to lamivudine. The variable importance estimates for the remaining mutations do not approach statistical significance.

Rhee et al. identify 184IV, 69ins, 65R, and 75T as the most important mutations conferring resistance to lamivudine and 69ins, 151M, 77L, 65R, and 75MT as the most important mutations conferring resistance to stavudine. With the exception of 184IV and 75MT, these mutations are not part of our analysis since they are present in fewer than 15 treatment change episodes. Our results that identify 184IV and 75AIMTS as the only important drug resistance mutations for this combination of drugs

are hence in excellent agreement with these analyses based on *in vitro* susceptibility tests.

Once important explanatory variables for  $Y(t)$  have been identified, it may be of interest to examine in more detail how their impact on  $Y(t)$  changes over time. For this purpose, we can estimate the dependence of variable importance on time using data-adaptive or smoothing methods that are better suited to give accurate estimates of this time course than the simple model given in 1.10. Figure 1.1 shows estimates of the time course for 184IV and 75AIMTS based on the LOESS smoothing technique (Cleveland, 1979). These plots show that 184IV has a sizeable impact on virologic response within a few weeks of treatment initiation, with the effect stabilizing after about ten weeks. The impact of 75AIMTS on virologic response develops somewhat more slowly over time.

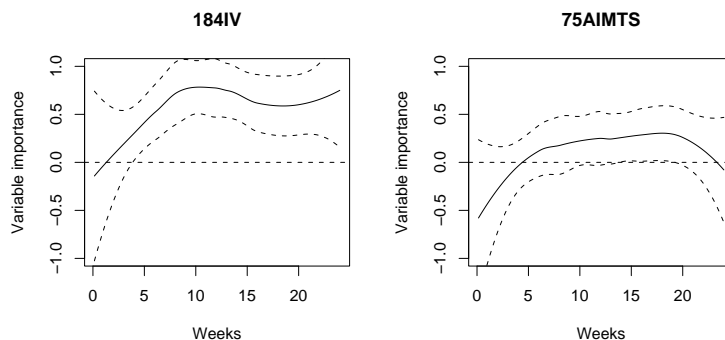


Figure 1.1. LOESS estimates of variable importance over time for the mutations 184IV and 75AIMTS with 95% pointwise confidence intervals.

## 7. Lessons Learned

This data analysis illustrates the importance of choosing an appropriate set of confounders  $W_j$  to adjust for when estimating the importance of each treatment variable  $A_j$ . In order for estimates to be more likely to translate to populations other than the one the sample was drawn from, one will generally want to adjust for as many of the known confounders of the relationship between  $A_j$  and  $Y(t)$  as possible. If some of these confounders are collinear with  $A_j$ , this will cause the variable importance measures to become very hard to estimate from the data

at hand, in which case we may be well-advised not to adjust for these collinear confounders.

The analysis further demonstrates that it is often preferable to obtain low-dimensional summary measures of variable importance time courses for the initial purpose of identifying important explanatory variables for  $Y(t)$ . At this stage, efforts to estimate variable importance measures at a chosen set of time points, for example, may unnecessarily increase the number of simultaneous hypothesis tests we have to perform and thus lower the chance of obtaining significant results. Once a subset of important explanatory variables has been identified, we may then investigate in more detail how their impact on  $Y(t)$  changes over time by using data-adaptive or smoothing methods that make fewer assumptions about the structure of this relationship.

The results obtained here also illustrate the importance of assessing the validity of the ETA assumption if estimates of variable importance are based on the IPTW transformation of the data. Seventy-three percent of all fitted treatment probabilities for the mutation 75AIMTS, for example, are either smaller than 0.05 or greater than 0.95, showing that for a majority of values of  $W_j$  it is essentially pre-determined whether a patient will have this mutation or not. In the absence of sufficient variability in the assignment of treatment for all values of  $W_j$ , variable importance estimates based on the IPTW transformation become very unreliable. In this case, the IPTW estimates rank 75AIMTS as only the seventh most important mutation conferring resistance to lamivudine and stavudine, while estimates based on the double robust and regression-based transformation identify it as the second most important drug resistance mutation, a ranking more likely to be correct given the current state of knowledge.

## 8. List of Tools and Resources

While the methodology described in this chapter has not yet been implemented in the form of a publicly available software package, the individual steps required as part of the analysis are fairly straightforward to carry out in a modern statistical computing environment like R (R Development Core Team, 2005). Within this environment, repeated-measures regression models based on generalized estimating equations can be fit using the `gee()` function found in the `gee` package. A function call of the form

```
gee(Y~X,id=ID,corstr='exchangeable')
```

is used to regress outcome measurements  $Y$ , made on individual subjects identified by the `ID` variable, on a covariate  $X$ , assuming an exchange-

able correlation structure among the measured outcomes. The D/S/A algorithm is implemented in the DSA package. The function call

```
DSA(X,Y,binind=1,IDlearn=ID,maxsize=10,maxorderint=2,
    maxsumofpow=2)
```

can be used to data-adaptively regress the binary outcome measurements  $Y$ , made on individual subjects identified by the  $ID$  variable, on a collection of candidate explanatory variables contained in the matrix  $X$ . The algorithm will search through models that contain up to 10 terms, including second-order interactions. The sum of powers of the variables contained in any one term cannot exceed two. The `pol spline` package implements the approach by Kooperberg et al., 1997, and offers the `polyclass()` and `polymars()` functions for data-adaptive regression of categorical and continuous outcomes, respectively, on a collection of candidate explanatory variables. A simple function call of the form

```
polyclass(Y,X)
```

is used to regress a categorical outcome  $Y$  on a collection of candidate explanatory variables contained in the matrix  $X$ . Note that these two functions are not suitable for modelling correlated outcomes so that they should only be used for fitting the treatment mechanism. The `multtest` package offers tools for obtaining valid  $p$ -values and confidence intervals in the context of simultaneous hypothesis tests. The Benjamini-Hochberg method for control of the false discovery rate can be carried out by a function call like

```
mt.rawp2adjp(rawp,proc="BH"))
```

where the vector `rawp` contains the unadjusted  $p$ -values.

The packages `gee`, `pol spline`, and `multtest` are available on the R website <http://www.r-project.org>. The DSA package will soon be posted on that website as well and can be accessed in the meantime at <http://www.stat.berkeley.edu/~laan/Software/>.

## 9. Conclusions

Given a list of 14 mutations thought to confer resistance to various NRTI drugs, the data analysis we describe here successfully identifies the mutation 184IV as most useful for predicting virologic response to lamivudine and stavudine. Extensive laboratory and clinical data have previously established 184IV as the most important mutation conferring resistance to lamivudine. The other mutation identified here, 75AIMTS, has been linked to moderate resistance to both lamivudine and stavudine. These results are also in excellent agreement with recent analyses

of *in vitro* susceptibility tests and thus illustrate the potential for the variable importance methodology described in this chapter to identify important explanatory variables for a time-varying outcome like viral load.

## Acknowledgments

We would like to thank Dr. Robert Shafer and Soo-Yon Rhee from the Stanford HIV drug resistance database for many helpful discussions, as well as kindly making available the dataset that was used in our case study analysis.

## References

- Benjamini, Y. and Hochberg, T. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 85:289–300.
- Bland, J.M. and Altman, D.G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310:170.
- Boucher, C.A.B., Cammack, P., Schipper, R., Rouse, P.L., and Cameron, J.M. (1993). High-level resistance to (-) enantiomeric 2′deoxy- 3′thiacytidine (3TC) in vitro due to one amino acid substitution in the catalytic site of human immunodeficiency virus type 1 reverse transcriptase. *Antimicrobial Agents and Chemotherapy*, 37:2231–2234.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees*. The Wadsworth Statistics/Probability series. Wadsworth International Group.
- Cleveland, W.S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.
- Dudoit, S. and van der Laan, M. J. (2006). *Multiple Testing Procedures and Applications to Genomics*. Springer. (In preparation).
- Kooperberg, C., Bose, S., and Stone, C.J. (1997). Polychotomous regression. *Journal of the American Statistical Association*, 92:117–127.
- Lacey, S.F. and Larder, B.A. (1994). Novel mutation (V75T) in human immunodeficiency virus type 1 reverse transcriptase confers resistance to 2′-3′dideoxy-2′,3′-dideoxythymidine in cell culture. *Antimicrobial Agents and Chemotherapy*, 38(6):1428–1432.
- Lehmann, E.L. and Romano, J. (2005). *Testing Statistical Hypotheses*. Springer, New York, 3rd edition.

- Liang, K. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models (Second edition)*. London: Chapman & Hall.
- Neugebauer, R. and van der Laan, M.J. (2005). Why prefer double robust estimates in causal inference? *Journal of Statistical Planning and Inference*, 129(1-2):405–426.
- R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rhee, S., Taylor, J., Wadhwa, G., Ravela, J., Ben-Hur, A., Brutlag, D., and Shafer, R.W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. (Submitted).
- Schurman, R., Nijhuis, M., van Leeuwen, R., Schipper, P., de Jong, D., Collis, P., Danner, S.A., Mulder, J., Loveday, C., and Christopherson, C. (1995). Rapid changes in human immunodeficiency virus type 1 RNA load and appearance of drug-resistant virus populations in persons treated with lamivudine (3TC). *Journal of Infectious Disease*, 171:1411–1419.
- Shafer, R.W. (2002). Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clinical Microbiology Reviews*, 15(2):247–277.
- Sinisi, S.E. and van der Laan, M.J. (2004). Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1).
- Tisdale, M., Kemp, S.D., Parry, N.R., and Larder, B.A. (1993). Rapid in vitro selection of human immunodeficiency virus 1 type 1 resistant to 3'-thiacytidine inhibitors due to a mutation in the YMDD region of reverse transcriptase. *Proceedings of the National Academy of Science*, 90:5653–5656.
- van der Laan, M.J. (2006a). Causal effects for intention to treat and realistic individualized treatment rules. Technical Report 203, Division of Biostatistics, University of California, Berkeley.
- van der Laan, M.J. (2006b). Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1).
- Westfall, P.H. and Young, S.S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley, New York.
- Zeger, S.L. and Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130.