UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Causal Inference and Prediction in Health Studies: Environmental Exposures and Schistosomiasis, HIV-1 Genotypic Susceptibility Scores and Virologic Suppression, and Risk of Hospital Readmission for Heart Failure Patients

Permalink

https://escholarship.org/uc/item/9z03362s

Author

Sudat, Sylvia

Publication Date

2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Causal Inference and Prediction in Health Studies: Environmental Exposures and Schistosomiasis, HIV-1 Genotypic Susceptibility Scores and Virologic Suppression, and Risk of Hospital Readmission for Heart Failure Patients

By

Sylvia Elise Keuter Sudat

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Alan Hubbard, Chair
Professor Sandrine Dudoit
Professor John Colford

Fall 2012

Causal Inference and Prediction in Health Studies: Environmental Exposures and Schistosomiasis, HIV-1 Genotypic Susceptibility Scores and Virologic Suppression, and Risk of Hospital Readmission for Heart Failure Patients

© 2012

by Sylvia Elise Keuter Sudat

Abstract

Causal Inference and Prediction in Health Studies: Environmental Exposures and Schistosomiasis, HIV-1 Genotypic Susceptibility Scores and Virologic Suppression, and Risk of Hospital Readmission for Heart Failure Patients

by

Sylvia Elise Keuter Sudat

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Associate Professor Alan Hubbard, Chair

Causal inference-inspired semi-parametric methods of measuring variable importance are well designed to answer questions of interest in health settings. Unlike traditional regression approaches, such variable importance measures are based on causal parameters that have straightforward real-world definitions, regardless of the approach used to estimate them. Parameters of regression models, in contrast, are not at all straightforward to interpret in real-world settings, because their definition relies completely on the correctness of the pre-specified model. Prediction-focused machine learning methods can avoid the issues of model pre-specification, but still do not provide estimates of variable importance that can be easily interpreted; the set of predictors chosen can also be highly variable. Semi-parametric methods combine the best of both approaches, and are able to utilize data-adaptive estimation algorithms while still returning a parameter estimate that is meaningful and can be simply understood.

In this dissertation, semi-parametric methods to assess variable importance are applied to three real-world health applications: the relationship between types of water contact and the prevalence of schistosomiasis infection in rural China; HIV-1 treatment regimen genotype susceptibility scores and their relationship with the rate of virologic suppression; and the impact of a telemanagement program on and the association of multiple risk factors with the rates of hospital readmission for heart failure patients. Emphasized are (1) the choice of parameter of interest as motivated by the research question, (2) estimator choice based on a consideration of theoretical properties and performance under non-ideal conditions, and (3) the use during the estimation process of machine learning algorithms and algorithms that utilize multiple candidate models. Four different causal parameters are defined and described, and multiple estimators are considered.

Each data analysis presents different opportunities to investigate aspects of causal inference-based semi-parametric methods. In the schistosomiasis analysis, a traditional regression approach is compared with semi-parametric methods. Estimator performance is compared in the HIV analysis, particularly in the context of the observed extreme violations of the experimental treatment assignment (ETA) assumption. The G-computation estimator, the inverse-probability-of-censoring-weighted (IPCW), its double-robust counterpart (DR-IPCW), and the targeted

maximum likelihood estimator (TMLE), are included in this comparison. The heart failure analysis addresses differences in causal parameter definition for a community-level treatment, and the related assumptions that must be added to the typical theoretical framework. Also included in this analysis is a comparison of super learning with traditional regression in terms of predictive performance.

Table of Contents

Acknowledgements

Special thanks to my adviser Alan Hubbard for his support, guidance, and suggestions as I worked through the process that led to this dissertation. Thanks also to Sandrine Dudoit and Jack Colford for being on my dissertation committee and to Mark J. van der Laan, under whose direction I began the HIV analysis. I must also humbly acknowledge my husband and family; without their practical and moral support, the completion of this dissertation would not have been possible.

Chapter 1

Introduction

The ultimate goal of any data analysis is to be able to provide as the end result a real-world answer to the research question that originally motivated the analysis. The initial steps in an analysis should flow naturally from the research question to the parameter of interest, or the quantity that can best be used to answer the research question, and finally to the best approach to estimate this parameter or quantity. While this ordering seems intuitive, it is surprisingly easy to start with the data available and skip directly to modeling or estimation, often without thoroughly considering whether the end result will achieve the research goals and adequately address the question of interest. For example, logistic regression models are reflexively used when the outcome of interest is binary, and little consideration is given to the question of whether the parameters estimates provided by such models will truly answer the research question or even provide interpretable results outside the confining limits of the pre-specified regression model.

A particular risk factor's predictive value and independent association with an outcome of interest are often targeted in the same analysis. These two goals are quite different, and require different analysis approaches. Advances in computing power have made the use of data-adaptive algorithms and algorithms that encompass multiple candidate estimators possible for most investigators. These approaches are very attractive because they are flexible and allow the researcher to maximize his or her ability to learn from the data, and to minimize the need to pre-specify a prediction model; this is of particular value when little or nothing is known about the true form of the data-generating distribution, as is the case in most public health settings. Such models cannot be easily used to answer the question of a risk factor's association with an outcome, however, and their parameters can be difficult to interpret. This does not reduce their value in prediction, but emphasizes the need for a different approach when causal inferences or measures of association with an outcome are desired.

Traditionally, researchers have turned to traditional regression and other pre-specified models to investigate questions of association with an outcome, with no ability to incorporate the abovementioned improvements in prediction into the estimation process. Pre-specified models are not only unable to incorporate additional information about the data-generating distribution that data-adaptive estimation approaches could provide, but their parameters may not even truly answer the questions of interest. Their pre-specification supposes that the researcher possesses accurate knowledge about the true form of the data-generating distribution, which is rarely (or never) the case in complex health systems. If it can reasonably be assumed that most such models are misspecified, it then becomes unclear what the parameters from these models might mean in a real-world context. If their meaning is unclear, it raises the additional question of how they can be optimally positioned to answer any proposed research question.

Semi-parametric methods have emerged as a way to combat many of the issues with traditional approaches. The parameters of interest can initially be defined in a causal framework, and the assumptions required in support of a causal interpretation identified. The statistical parameter, or the parameter that is estimable from the data, can then be defined. Since causal assumptions are frequently unjustifiable in many health settings, it is important to be sure that the statistical parameter retains subject-matter value. Note that the definition of the parameter of interest is completely separate from the estimation approach, and its interpretation is therefore not confined to the appropriateness of any model or estimator that may be chosen.

Once the parameter of interest has been identified, it is then possible to proceed to the question of how to best estimate the parameter. Many potential estimators are available, such as the G-computation estimator, the inverse-probability-of-censoring-weighted (IPCW) estimator, the double-robust IPCW estimator (DR-IPCW), and targeted maximum likelihood estimator (TMLE) (van der Laan & Rose, 2011; van der Laan & Robins, 2003; Robins, 1986; Robins, 2000; van der Laan & Rubin, 2006). Each estimator may have differing theoretical properties, and these should be considered when choosing between them. Estimator performance under non-ideal conditions, particularly conditions that may be present in the data to be analyzed, should also be carefully considered. One common non-ideal condition is practical violation of the experimental treatment assignment (ETA) or positivity assumption, in which data sparsity in the face of many potential explanatory variables results in underrepresentation of one or more outcomes in certain strata of the data (van der Laan & Robins, 2003; Messer, Oakes, & Mason, 2010). If such an issue is present in the data, it should be kept in mind when the estimator of choice is selected.

Because the target parameter's definition is completely separate from the choice of estimator and estimation process, it is possible to employ any prediction method available to best model the data-generating distribution. Super learning, an approach that combines multiple candidate estimators into a single prediction model using cross-validation, is particularly appealing (Sinisi S. E., Polley, Petersen, Rhee, & van der Laan, 2007; van der Laan, Polley, & Hubbard, 2007). Such an approach gives the researcher the freedom to incorporate any potential model thought to be predictive of the outcome of interest without having to arbitrarily choose between them. If nothing is known about the true form of the data-generating distribution, this can be respected by including many candidate models; alternately, if something is known, any potential model based on the subject matter can be included as well. The super learning approach can provide as the final prediction model a convex combination of candidate models, weighted using cross-validation, or can use cross-validation to select a particular candidate model as the best choice.

This dissertation walks through three applications of semi-parametric parameter estimation and data-adaptive prediction methods to traditional health applications. In each application, the parameter of interest is chosen in response to research questions, and not simply in response to the type of data being analyzed. Different parameters of interest are considered, as are different estimators of those parameters of interest.

Chapter 2 discusses an analysis of the relationship between types of water contact and the prevalence of schistosomiasis infection in rural China. A traditional regression analysis is compared with a semi-parametric approach using a so-called population intervention parameter,

as estimated by the inverse probability of censoring-weighted (IPCW) estimator (Ahern, Hubbard, & Galea, 2009; Greenland & Drescher, 1993; Fleischer, Fernald, & Hubbard, 2007; Hubbard & van der Laan, 2008). Recursive partitioning, regression, and classification trees are used to estimate the data-generating distribution (Breiman, Friedman, Olshen, & Stone, 1984). Also discussed is the difference between measures of variable importance returned by prediction-focused methods and methods derived with causal inference in mind.

Chapter 3 considers HIV-1 treatment regimen genotype susceptibility scores and their relationship with the rate of virologic suppression, both in terms of association and prediction. A population intervention parameter is again employed to investigate association with the outcome. G-computation, IPCW, DR-IPCW, and TMLE are compared as estimators of the parameter of interest, and super learning is applied to estimate the data-generating distribution. Estimator bias induced by observed extreme practical violations in the experimental treatment assignment (ETA) or positivity assumption is investigated using a parametric bootstrap diagnostic (Wang, Petersen, Bangsberg, & van der Laan, 2006; Petersen, Porter, Gruber, Wang, & van der Laan, 2012). Influence curve-derived standard errors for IPCW, DR-IPCW, and TMLE are also compared to standard errors obtained using the nonparametric bootstrap.

Chapter 4 describes an evaluation of a heart failure intervention designed to reduce rates of hospital readmission, and the estimation of the independent association of a variety of risk factors with the same readmission outcomes. Two parameters of interest are considered, and estimated using super learning and TMLE. Causal parameters in the setting of a community-level intervention are discussed, and the required additional assumptions described (van der Laan M. J., 2010). Also described is the method of using cross-validation to compute influence curve-based standard errors for effect estimates. Predictive performance of a readmission risk score used by the heart failure program is also investigated, and super learner prediction models are compared with simple main terms logistic regression prediction models.

Chapter 2

Using Variable Importance Measures from Causal Inference to Rank Risk Factors of Schistosomiasis Infection in a Rural Setting in China

This chapter has been published as a jointly-authored article in *Epidemiologic Perspectives & Innovations* (Sudat, Carlton, Seto, Spear, & Hubbard, 2010).

2.1 Background

Schistosomiasis is a parasitic disease affecting an estimated 200 million people in 76 countries (WHO, 2006). Humans become infected with schistosomiasis following contact with water containing cercaria, the larval stage of the parasite. Infection can lead to liver fibrosis and portal hypertension, and may cause anemia (King, Dickman, & Tisch, 2005; Leenstra, et al., 2006; Ross, et al., 2002).

Recent studies have shown that the distribution of human schistosomiasis infections can be explained in part by spatial variability in water contact, particularly with respect to differences in cercarial density. For example, clusters of *Schistosoma hematobium* infections in rural Kenya were identified near water bodies with high numbers of cercaria-shedding snails (Clennon, et al., 2004). Also, in contrast to water contact measures that ignore spatial variability in cercarial density, measures of water contact that adjust for estimated cercarial density at the site of contact have shown strong correlations with human infection intensity (Li Y. , et al., 2000; Seto, Lee, Liang, & Zhong, 2007).

Less attention has been paid to temporal variability in infection risk and to the variability in infection risk from specific water contact activities. While diurnal variations in the infectivity of cercaria have been recognized for decades, little is known about the variability in infection risk throughout the transmission season (Nojima, Santos, Blas, & Kamiya, 1980). Li *et al.* observed two annual peaks in *S. japonicum* infection prevalence in the lower Yantzee basin (Li Y. S., Sleigh, Ross, Williams, Tanner, & McManus, 2000). In the irrigated hillsides of southwest China, temporal fluctuations in both hydrology and snail populations have been documented, and may yield corresponding variation in infection risk throughout the transmission season (Remais, Hubbard, Wu, & Spear, 2007; Remais, Liang, & Spear, 2008). Specific water contact activities may also affect infection risk, due perhaps to the location in which these activities are performed and the parts of the body exposed. Several specific water contact activities have been associated with the prevalence of *S. hematobium* infection in Zanzibar and *S. mansoni* infection in Cote d'Ivoire (Matthys, et al., 2007; Rudge, et al., 2008). However, neither analysis accounted for the duration or timing of water contact, and such relationships have not yet been examined for *S. japonicum*.

The two studies of *S. mansoni* and *S. hematobium* mentioned above examined numerous risk factors for infection using traditional correlation and multivariate regression techniques. The multivariable regression approach, while common, imposes an arbitrary model that limits the interpretation of results (Robins & Ritov, 1997). For example, parameters from such models rarely have simply understood definitions within the context of the subject matter; they only have meaning within the context of the arbitrarily specified model. Multivariable regression models can also return misleading inference, because the assumption of an arbitrary model does not allow for model misspecification, and thus incorrectly estimates variability (Hubbard, et al., 2010).

In contrast to multivariable regression, semi-parametric variable importance measures inspired by parameters from the causal inference literature have the virtue of (1) using machine learning algorithms to determine flexibly how to adjust for potential confounding variables without requiring arbitrary model pre-specification and (2) returning a simple and interpretable measure of variable importance that under assumptions can also yield estimates of the effect of intervention (Ahern, Hubbard, & Galea, 2009). Such parameters have been referred to as population intervention parameters (Ahern, Hubbard, & Galea, 2009; Greenland & Drescher, 1993; Fleischer, Fernald, & Hubbard, 2007; Hubbard & van der Laan, 2008). This alternative to a traditional regression analysis is well suited to the exploratory analysis of high-dimensional data, where one desires to investigate the independent association of one variable and an outcome in the presence of many correlated variables.

We analyzed data from a retrospective study in which 1011 individuals reported their water contact during the 2000 *S. japonicum* infection season in rural China; infection status in 2000 was also recorded for these individuals. Water contact was calculated using the estimated duration of water contact and the estimated body surface area in contact with water during the specific water contact activity. We aimed to explore the relative importance of different types of water contact, defined by both water contact activity and by the month in which the water contact occurred, on the probability of schistosomiasis infection. We analyzed these data in three ways: first, by applying a prediction (machine learning) algorithm; second, by using a simple multivariable regression; and third, by assessing variable importance using a causal inference-inspired population parameter. We discuss the results of each method, as well as the limitations of interpretation within the context of the method used.

2.2 Methods

2.2.1 Data Collection

This research was conducted in Xichang County located in the southwest of Sichuan Province, China. The region is hilly with irrigated agriculture and historically high schistosomiasis infection prevalence. Twenty villages ranging in size from approximately 100 to 300 residents were selected to participate in a cross-sectional study to characterize determinants of schistosomiasis infection (Spear, et al., 2004). In November 2000, all residents in the 20 villages were asked to participate in schistosomiasis infection surveys and in an interview to assess basic demographic characteristics including age, occupation and educational attainment. Participation rates were high: an estimated 90% of residents participated in these surveys. This research was conducted in close collaboration with the Xichang County Anti-Schistosomiasis Station and the Institute of Parasitic Diseases at the Sichuan Center for Disease Control. All participants provided verbal informed consent and human data collection protocols were approved by the Berkeley Committee for the Protection of Human Subjects and the Sichuan Institutional Review Board.

A 25% random sample of residents, stratified by village and occupation, was interviewed in person in November 2000 about their water contact patterns throughout the schistosomiasis transmission season. Participants were asked about eight different activities that involve contact

with irrigation, pond or stream water each month from April through October: washing clothes or vegetables, washing agricultural tools, washing hands and feet, playing or swimming, irrigation ditch cleaning and water diverting, planting rice, harvesting rice and fishing. These water contact activities will be referred to subsequently as laundry, tool washing, bathing, swimming, ditch digging, rice planting, rice harvesting, and fishing, respectively. Participants were asked how often they performed each activity each month and for how many minutes each time, providing an estimate of water contact frequency and duration. Each activity was assigned an exposure intensity weight in order to account for differences in body surface area exposed. Field studies in the selected villages were conducted to observe which body parts were typically wetted for each water contact activity, and burn charts were used to estimate the percent of total body surface area accounted for in each exposed body part [21]. Water contact intensities were assigned as follows: laundry (0.05), tool washing (0.03), bathing (0.12), swimming (0.20), ditch digging (0.05), rice planting (0.05), rice harvesting (0.05) and fishing (0.32). Total body surface area for adults was estimated to be 1.626m^2, and for children age 14 and under: 1.130 m^2 (Mosteller, 1987). For each activity *i* in month *k*, water exposure in minutes-meters^2 was calculated:

$$WC_{ik} = Frequency_{ik} \times Duration_{ik} \times Intensity_i \times BodySurfaceArea.$$

An individual's water contact for each month was calculated by summing water exposure for all activities that month. Likewise, an individual's total water exposure for each activity was calculated by summing the activity-specific water exposure over the seven months. The total water contact over the entire period was also calculated. Because it was determined that only one infected individual had any water contact associated with rice harvesting, rice harvesting was excluded from the set of activity variables. This type of water contact was not excluded from the monthly water contact variables, or from the total water contact variables.

At the same time as the water contact surveys, and corresponding with the end of the transmission season, schistosomiasis infection surveys were conducted using two different stool examination techniques. Participants submitted stool samples from three different days and each sample was examined using the miracidial hatch test according to Chinese Ministry of Health protocols (The Office of Endemic Disease Control MoH, 2000). The Kato-Katz thick smear procedure was also used; three 41.5mg slides were prepared from homogenized stool samples and examined for *S. japonicum* eggs (Katz, Chaves, & Pellegrino, 1972). Any person with a positive miracidial hatch test or at least one *S. japonicum* egg detected through Kato-Katz was classified as infected. All infected individuals were referred to local health officials for treatment with praziquantel.

2.2.2 Statistical Analyses

Prediction Algorithm

In our first analysis, we used a machine-learning algorithm to choose the "best" set of infection predictors. This algorithm formed recursive partitioning, regression, and classification trees, as implemented in the R function *rpart* (R version 2.10.0, Copyright (C) 2009; Breiman, Friedman, Olshen, & Stone, 1984; Therneau & Atkinson, 1997). The algorithm was allowed to choose

among all of the possible water contact variables, as defined above: activity type, water contact month, and total water contact. Since the activities are sums over all months, the months are sums over all activities, and the total is the sum of all water contact over the entire study period, including these variables together would not make sense in an approach attempting to determine associations between the variables and the outcome (as in the analyses conducted later in the paper). However, from the prediction standpoint, the only concern is the accuracy of prediction; it makes the most sense, therefore, to include as many variables as possible in the potential prediction algorithm, which is why we included all variables. We note that *rpart* is just one of many machine learning algorithms that could be used, including algorithms that combine results from several learners (Sinisi S. E., Polley, Petersen, Rhee, & van der Laan, 2007). This approach generalizes to any such routines.

In an attempt to assess the relative "importance" of the variables in predicting the outcome, we applied a Monte Carlo re-sampling approach (nonparametric bootstrap) (Efron, 1982). The study individuals were randomly re-sampled with replacement (meaning that one subject could be sampled more than once, but that all samples were of the same size), and the *rpart* tree was recalculated. This bootstrapping method is a commonly used way of simulating re-sampling from the target population, and can help to examine how small changes in the data can affect the prediction model chosen. We performed this re-sampling approach 5000 times, and tabulated the number of times each variable was chosen by *rpart* in the prediction model. Multiple splits on a given variable within the same *rpart* fit were counted only once on each iteration.

Multiple Regression

Turning away from the prediction-focused approach, our second analysis was a main-effects log-linear regression, in which we also included age category (<18, 18-29, 30-29, 40-49, 50+) and village indicator variables as possible confounders. Here we separated the activity types from the months into two separate models, and excluded total water contact from both models. We could not use log-linear binomial models because they generated predicted probabilities that exceeded one, so we used instead Poisson log-linear models.

$$\begin{aligned}\text{Model 1}: &\log\left[E\left(Y \mid W_{activity}, V\right)\right] \\ &= \alpha + \beta_{activity} W_{activity} + \gamma V \\ \text{Model 2}: &\log\left[E\left(Y \mid W_{month}, V\right)\right] \\ &= \alpha + \beta_{month} W_{month} + \gamma V\end{aligned}$$

In both models, Y is the (binary) outcome, V is the vector of village and age category indicators, and γ is the vector of coefficients associated with V. In Model 1, $W_{activity}$ is the vector of activity type water contact variables, and $\beta_{activity}$ is the vector of activity type coefficients; in Model 2, W_{month} is the vector of monthly water contact variables, and β_{month} is the vector of month coefficients. Because we did not wish to rely upon the Poisson assumption for estimating our standard errors and deriving inference, we instead calculated robust standard errors using the Huber/White sandwich estimator (Huber, 1967; White, 1980). Regression estimates were obtained using the *glm* command in Stata (Stata 10).

Variable Importance

Our third (semi-parametric) approach estimated a so-called *variable importance* (VI) parameter which compares the current distribution of the outcome to its distribution under a theoretical experiment where the variable of interest is set to the lowest risk. In our data, this is equivalent to comparing the observed infection prevalence distribution to the distribution of infection in a theoretical experiment in which the entire study population never experienced a particular type of water contact.

Assume the current variable of interest is *A*, the outcome is *Y*, and the confounders – in this case, all other water contact variables except *A* – are *W*, and *V* are the additional confounders (age category and village). Our VI estimate is inspired by the following causal parameter:

$$\frac{E(Y_0)}{E(Y)}.$$

Y_a represents the outcome if – possibly contrary to fact – everyone had exposure $A = a$. Outcomes defined in such a way have been referred to as *counterfactuals* (Rubin D. B., 1978). In the case of our binary outcome variable, *E(Y)* is estimated as the current disease prevalence in our target population, which is estimated as the average of the observed *Y* values.

If *Y* is binary (yes/no) – as it is in our case – this parameter can be interpreted as the proportional change, relative to current rates, in the prevalence of schistosomiasis in our target population if everyone were unexposed to the particular risk. This parameter is akin to the attributable risk, and its magnitude is both a function of the adjusted association of *A* and *Y* and of the prevalence of exposure. For example, removing exposure would have little effect on the value of this causal parameter if the exposure in question were very rare, even if it were strongly related to the disease outcome. Conversely, removing a common exposure that only modestly increased the risk of disease could have a much larger impact on the parameter's value.

With regards to the distribution of the data alone – that is, without assuming the necessary identifiability conditions for making causal inference (no unmeasured confounders and independence of counterfactual outcomes, or the so-called stable unit treatment value assumption – SUTVA) (Rubin D. B., 1986) – our VI measure is an estimate of the following:

$$VI = \frac{E_{W,V} E(Y \mid A = 0, W, V)}{E(Y)}.$$

The numerator is interpreted as the mean predicted value of *Y* assuming one sets the exposure to 0 (*A*=0 means unexposed) but keeps the other variables at their observed values. $E_{W,V}$ in the numerator denotes that this mean predicted value of *Y* is also taken over all *W* and *V*.

The denominator was estimated by simply taking the mean of the *Y* values. To estimate the numerator, we used the so-called inverse-probability-of-censoring-weighted (IPCW) estimator:

$$\hat{E}_{W,V}\hat{E}(Y \mid A=0,W,V) = \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=0)Y_i}{\hat{P}(A_i=0 \mid W_i,V_i)}.$$

Here $\hat{P}(A_i = 0 \mid W_i, V_i)$ is an estimate of the probability that A=0 given the values of the covariates W_i and V_i for subject i. The form of this estimator makes obvious another assumption, which has been called positivity or experimental treatment assignment (ETA) assumption, which in this case says that $P(A=0 \mid W,V) > 0$ in the data-generating distribution (Cole & Hernan, 2008; Mortimer, Neugebauer, van der Laan, & Tager, 2005; Messer, Oakes, & Mason, 2010).

The IPCW estimator is a type of weighted average of the Y values, in which the weights are proportional to the probability of being unexposed ($A_i=0$) given the other covariates (W_i and V_i). The IPCW estimator relatively up-weights the disease outcomes of unexposed individuals with covariates underrepresented within the unexposed group, which has the effect of adjusting for confounding bias. Because $P(A_i|W_i,V_i)$ is unknown in this case, we used a machine-learning algorithm (*rpart*) to estimate a model for this probability.

A VI estimate was calculated for each variable of interest. Specifically, we define the VI estimate for each water contact activity as follows:

$$VI_{activity} = \frac{E_{W_{activity},V}E(Y \mid A=0,W_{activity},V)}{E(Y)},$$

where A represents the water contact activity type for which a VI estimate is being calculated, $W_{activity}$ represents the remaining water contact activity type variables, and V represents the age category and village covariates. The VI estimate for each month is defined equivalently, with W_{month} in place of $W_{activity}$. As in the logistic regression analysis, total water contact was excluded; it would not be meaningful to estimate $E_{W,V}E(Y \mid A=0,W,V)$ for A=total water contact, since none of the other water contact variables could be nonzero if total water contact were equal to zero.

To derive our inference, we estimated standard errors using the non-parametric bootstrap with 5000 iterations. Specifically, participants were re-sampled with replacement, producing 5000 bootstrap samples of size 1011. For each of these 5000 samples, VI estimates were calculated, including a re-calculation of $\hat{P}(A_i = 0 \mid W_i, V_i)$. The standard deviation across these 5000 estimates was then calculated and used for inference. Because the model for $P(A_i = 0 \mid W_i, V_i)$ was not pre-specified, this method of calculating the standard error will account for both sampling variability (by re-sampling) and the variability introduced by model uncertainty with regards to $P(A_i = 0 \mid W_i, V_i)$ (by allowing for changes in the model for $\hat{P}(A_i = 0 \mid W_i, V_i)$ at each iteration).

2.3 Results

Figure 2.1 shows the full data *rpart* tree formed by allowing the machine learning algorithm to choose splits from the pool of all water contact variables. April, May, June, tool washing, ditch digging, bathing, and rice picking were the water contact variables chosen for classification.

When the data were re-sampled with replacement, Table 2.1 lists the number and percentage of times (out of 5000) each variable was chosen for classification in a given *rpart* tree. The covariates are ordered according to the number of times they were chosen to be part of each *rpart* tree, from largest to smallest. This method identified April (92%), June (92%) and total water contact (86%) as the most frequently chosen predictors of infection status within the bootstrapping algorithm. The six variables chosen for classification in the original full data tree (Figure 1) are among the top seven identified most frequently for use in the bootstrap sample *rpart* trees. However, total water contact, chosen 86% of the time in the bootstrap samples, was not part of the original full data tree.

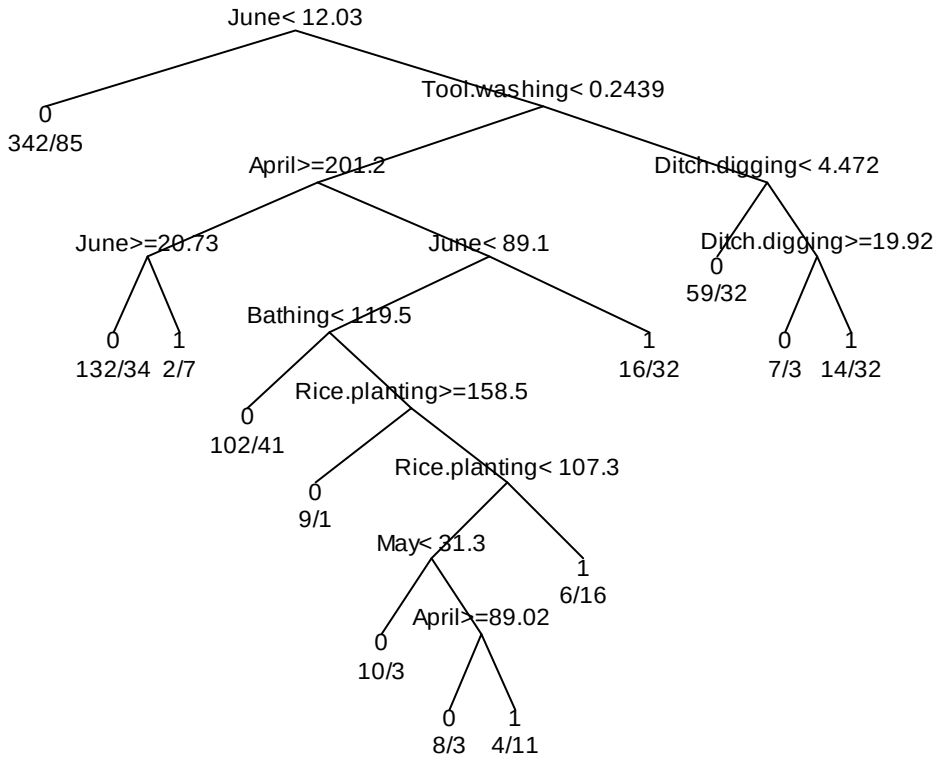

Figure 2.1 Full data *rpart* classification tree.

Table 2.1 - Number of times out of 5000 that each water contact type was chosen by *rpart* to form a data-adaptive classification tree.

Water contact type	Number of times chosen	Percentage
April	4608	92.2%
June	4602	92.0%
Total	4283	85.7%
Tool washing	4067	81.3%
Ditch digging	3825	76.5%
Rice planting	3677	73.5%
May	3652	73.0%
September	3326	66.5%
July	3181	63.6%
Bathing	3073	61.5%
October	2787	55.7%
Swimming	2481	49.6%
Laundry	2133	42.7%
August	1892	37.8%
Fishing	69	1.4%

Tables 2.2 and 2.3 show results from the log-linear regression models, along with the prevalence of each type of water contact in our sample. The correlations between the various water contact variables range from -0.02 (between April and August) to 0.68 (between July and August) for the monthly variables and from -0.15 (between swimming and bathing) and 0.28 (between rice picking and bathing) for the activity variables. The reported relative risks were calculated as $e^{\hat{\beta}_i \overline{X}_i}$, where $\hat{\beta}_i$ is the estimated regression coefficient and $\overline{X}_i$ is the mean water contact across all subjects for water contact variable *i*. This relative risk therefore reports the risk of having the mean value for water contact variable *i* versus the risk of having no water contact of type *i*. As previously mentioned, the month and activity variables were separated into two different models, which is why the results are reported separately. The estimates in Tables 2.2 and 2.3 are also adjusted for age category and village. We do not report relative risks associated with age category and village because the effects of these covariates were not the focus of this study.

Table 2.2 – Relative risk estimates for water contact by month.

Month	Prevalence	Relative Risk	95% CI	Std. error	p-value
June	0.75	1.03	(0.98, 1.09)	0.03	0.20
October	0.58	0.95	(0.89, 1.03)	0.04	0.22
May	0.75	1.04	(0.97, 1.13)	0.04	0.25
April	0.73	1.04	(0.95, 1.14)	0.05	0.45
August	0.76	1.03	(0.94, 1.13)	0.05	0.51
September	0.70	0.98	(0.89, 1.07)	0.05	0.64
July	0.77	0.98	(0.91, 1.06)	0.04	0.68

These estimates are based on a main-effects log-linear regression, and are also adjusted for age category and village. The relative risks reflect the difference in risk of infection between exposure at the mean value for that month and zero exposure.

Table 2.3 – Relative risk estimates for water contact by activity.

Month	Prevalence	Relative Risk	95% CI	Std. error	p-value
Tool washing	0.20	1.03	(1.01, 1.05)	0.01	<0.01
Laundry	0.22	1.02	(1.00, 1.05)	0.01	0.08
Swimming	0.21	1.02	(1.00, 1.05)	0.01	0.10
Ditch digging	0.48	0.99	(0.98, 1.00)	0.01	0.16
Fishing	0.02	1.01	(1.00, 1.02)	0.01	0.17
Bathing	0.49	0.98	(0.92, 1.04)	0.03	0.46
Rice planting	0.65	1.03	(0.94, 1.13)	0.05	0.52

These estimates are based on a main-effects log-linear regression, and are also adjusted for age category and village. The relative risks reflect the difference in risk of infection between exposure at the mean value for that month and zero exposure.

In the log-linear regression framework, none of the monthly water contact variables were found to have strong associations with the outcome. All month-specific relative risk estimates are very close to one and have 95% confidence intervals that include one. This implies that the risk of having a positive stool sample when these variables are at their mean values is indistinguishable from the risk when there is zero water exposure during these months. Similarly, the relative risks associated with the water contact activity types are also all very close to one, and almost all have 95% confidence intervals that include one. The tool washing-specific relative risk has a 95% confidence interval that does not cross one; the estimated relative risk is still extremely close to one, however, implying almost no detected difference in risk. These results are of course only interpretable in the context of the regression models used.

Tables 2.4 and 2.5 show VI estimates for the two sets of water contact variables. As in the log-linear regression framework, the monthly water contact variables were analyzed separately from the water contact activity variables. As previously explained, the VI estimates were adjusted for age category and village by including these variables in the estimation of $P(A_i|W_i,V_i)$. (In similarity with the regression analysis, we did not calculate VI estimates for age category and village.) Confidence intervals and *p*-values based on the bootstrap-derived standard errors are also reported. In contrast to the log-linear regression results, which identified no detectable adjusted associations with the outcome among the monthly water contact variables, July's VI estimate indicates a strong adjusted association. If one interprets this VI estimate as an estimate of $\frac{E(Y_0)}{E(Y)}$, it implies that eliminating water contact in July would reduce the prevalence of schistosomiasis measured in the study by 84%, or from 0.3 to 0.05. The 95% confidence interval for this estimate indicates a range of 78% to 89%. The prevalence of exposure in July is 0.77, which along with August is the highest of any month. The VI estimates for all other months are near one and have 95% confidence intervals that include one (many of which are quite broad). No other month, therefore, has a detectable association with the outcome.

In terms of VI, no other type of water contact had as large an impact on infection risk as July water contact. Tool washing and rice planting were the only two activities with a discernable impact on infection risk – all other activity types (Table 4) have VI estimates near one and 95% confidence intervals that include one. Both of the VI estimates associated with tool washing and rice planting, in contrast, have 95% confidence intervals that do not cross one. Interpreting the

VI results once again as estimates of $\frac{E(Y_0)}{E(Y)}$ would imply an estimated 12% reduction in the prevalence of schistosomiasis by eliminating tool washing and an estimated 29% reduction by eliminating rice planting. The associated 95% confidence intervals for these estimates imply a range of 3% to 20% for tool washing and 4% to 47% for rice planting. As shown in Table 5, the prevalence of water exposure due to tool washing in our study population was 0.20, while the prevalence of water exposure due to rice planting was 0.65.

Table 2.4 - Variable importance estimates for water contact by month.

Month	Prevalence	VI estimate	95% CI	Std. Error	p-value
July	0.77	0.16	(0.11, 0.22)	0.18	<0.01
August	0.76	1.70	(0.48, 6.02)	0.66	0.42
May	0.75	1.18	(0.32, 4.30)	0.66	0.81
October	0.58	1.05	(0.60, 1.84)	0.28	0.86
June	0.75	0.97	(0.27, 3.56)	0.66	0.97
September	0.70	1.01	(0.41, 2.49)	0.46	0.98
April	0.73	1.00	(0.40, 2.50)	0.46	1.00

The prevalence of water contact for each month in our study population is also shown.

Table 2.5 - Variable importance estimates for water contact by activity type.

Month	Prevalence	VI estimate	95% CI	Std. Error	p-value
Tool washing	0.20	0.88	(0.80, 0.97)	0.05	0.01
Rice planting	0.65	0.71	(0.53, 0.96)	0.15	0.03
Swimming	0.21	0.96	(0.87, 1.06)	0.05	0.38
Ditch digging	0.48	0.94	(0.80, 1.10)	0.08	0.42
Bathing	0.49	1.09	(0.88, 1.35)	0.11	0.42
Laundry	0.22	0.97	(0.89, 1.06)	0.04	0.45
Fishing	0.02	1.00	(0.98, 1.02)	0.01	0.83

The prevalence of water contact for each month in our study population is also shown.

2.4 Conclusions

The three analysis approaches used here are all attempts to answer the same research question: what is the best estimate of the contribution of one explanatory variable to the mean outcome in the presence of other correlated explanatory variables? We specifically hoped to see how various types of water contact affected the probability of a positive stool sample, adjusting for other types of water contact, age, and village.

The use of machine learning algorithms for model selection is attractive, particularly because the model does not have to be pre-specified; this means estimating the association parameters while acknowledging that very little is typically known about the form of the model. A comparison of Figure 1 and Table 1, however, provides an example of how simply determining whether or not a variable is chosen by a machine learning algorithm (such as *rpart*) is not a particularly robust procedure for defining the importance of a variable. Given a finite sample size and highly correlated predictors – as we have in our data – small changes in the data often result in large changes in the variables chosen as predictors. This can occur even as the fidelity of prediction is nearly unchanged; there are often several sets of variables in various functional forms that can provide nearly identical accuracy of prediction. This issue is partially what inspired the idea of

bagging or bootstrapping these machine learning algorithms, such as in the case of random forests (Breiman, 2001). For example, our full data tree could lead us to conclude that total water contact is less predictive of a positive stool sample than the specific activity and month variables chosen to be part of the tree. Table 2.1, however, would lead us to conclude that total water contact is one of the top three most predictive variables – and therefore more "important" than four out of the six variables identified in the full data tree. Due to this instability, machine learning algorithms alone provide sub-optimal information for determining the importance of variables.

The actual best set of predictor variables is a function of the type of model, the method for constructing candidate models, and the method used to choose the so-called tuning parameters. Our results here therefore do not generalize to all machine learning routines – such as, for example, the Deletion/Substitution/Addition algorithm, POLYCLASS or random forests (Breiman, 2001; Sinisi & van der Laan, 2004; Kooperberg, Bose, & Stone, 1997). Generally, as implied by the results displayed in Table 2.1 and Figure 2.1, prediction algorithms are not constructed to provide any easily interpretable estimates of each water contact variable's contribution to the probability of a positive stool sample, which is ultimately what we were trying to investigate. Machine learning algorithms can be applied most effectively to answering our question of interest when used within an estimation framework whose parameters are defined independently from the specific model chosen by a given algorithm (such as *rpart*). This semi-parametric approach, of which our VI analysis is an example, contrasts dramatically with estimating simple, parametric regression models and reporting the resulting coefficients as association parameters (such as the relative risks reported in Tables 2.2 and 2.3). Though such regression analyses can produce parameters with relatively straightforward public health interpretations, the interpretations only remain straightforward if the pre-specified regression model is correct; any interpretation of the estimates obtained must implicitly assert the truth of the model used, though there is very rarely any justification for a specific parametric model's *a priori* truth. In addition, the lack of data-adaptive procedures can sacrifice power by resulting in much larger residual variability than approaches that use the data to fit the models. Tables 2.2 and 2.3, for example, show that under the constraints of the regression model, even the coefficients with 95% confidence intervals that did not cross one yielded relative risks very close to one, suggesting little contribution to the variability of the outcome. Whether this is a true result, however, or merely reflective of a poorly chosen model, is impossible to assess. The regression approach, though common, is therefore a dangerous choice as a basis for making causal inferences. Interpretation of parameters (conditional relative risks) in the context of a misspecified model are also of dubious value, since it is difficult to know what such interpretations really mean. This is true of the innumerable regression approaches reflexively used throughout observational epidemiology and other empirical fields.

Though one data analysis cannot justify the global use of an analysis technique, at least there is some hope that our approach here has found potentially interesting associations. Specifically, the importance of July water contact in our VI results – not detected by the regression analysis – could suggest temporal variability in infection risk during the infection season. This could be due to a combination of factors, since infection risk depends not only on water contact intensity but also on cercarial concentration in that water. A summer peak in cercarial concentration was observed in a number of villages in this same area in 2001 using a mouse bioassay procedure

throughout the infection season (Spear, et al., 2004). The peak occurred in August, not July, but year-to-year variability in cercarial concentration can be expected due to seasonal fluctuations in snail populations and agricultural activities driven by changes in rainfall, temperature, and humidity. Temporal variability in infection risk can also be influenced by seasonal changes in activities known to be associated with infection, such as swimming, which may increase during summer months when school is not in session and ambient temperatures are high. In addition, prior work has documented seasonal fluctuations in hydrology which correspond to differences in infection patterns between schistosomiasis endemic regions within Sichuan province (Remais, Liang, & Spear, 2008). One must consider, however, that this dataset has a number of limitations. The retrospective nature of the water contact surveys calls into question the accuracy of recall – particularly given the relatively long period of time (seven months) during which study participants were asked to recount their water contact activities. The analysis also relies on the definition of water contact, which as previously described includes an estimate of the body surface area believed to be in contact with water during certain activities. We are additionally limited by the need to analyze the monthly water contact and water contact activity variables separately; while it would have been ideal to consider the 56 activity type-by-month variables, the number of covariates is simply too large in comparison with the sample size for any technique to single out individual contributions. We therefore chose to simplify the set of variables by considering activity separately from month, thus providing some power to detect adjusted associations.

While the results of this analysis are far from conclusive, they nonetheless suggest possibly fruitful areas for future research. If a high-risk period in the schistosomiasis infection season could be detected in something close to real time, new prevention options would be opened. Recent advances in detecting schistosome cercariae in water using PCR techniques could potentially provide such a tool (Hung & Remais, 2008). The notion of changing from a surveillance system that relies on episodic human infection surveys to one based on water monitoring has many attractions, including the likelihood of lower cost. Water monitoring is also an appealing option in areas where schistosomiasis re-emergence has occurred or is suspected (Liang, et al., 2007).

Though we compare here three specific analysis techniques, we note that many different machine learning algorithms (other than classification trees) are available, different regression models could be specified, and different approaches to estimating our VI parameter could be used (including G-computation and Targeted Maximum Likelihood) (Robins, 2000; van der Laan & Rubin, 2006). The general principals contrasting these methods remain the same, however, and are important in the larger issue of estimating the independent and potentially causal association of risk factors in data sets with large numbers of covariates. Prediction (machine learning) algorithms are very well-designed to provide optimal prediction and to balance the variance and bias in the predicted value (the estimate of $E(Y|A,W,V)$); they are not optimal for determining the contributions of individual variables directly. This is particularly obvious since small changes in the data can result in large changes in the variables chosen. In contrast, the standard regression model approach has a nicely interpretable parameter, but is entirely dependent upon the correctness of the model specified. The definition of the parameter itself is also generally tied to the form of the model – for example, adding a multiplicative interaction term into a regression model changes the meaning of the main effect term. Thus, the definition of a given parameter is

only useful if the model is correct, and that parameter's interpretation changes as other variables are added to or removed from the model. In reality, such models are never correct, and there is no mechanism for allowing them more flexibility (such as through machine learning algorithms) to reduce bias as sample size grows. These issues expose the need for a meaningful parameter, one whose estimation can capitalize on the virtues of the asymptotic bias-reduction of machine learning algorithms and whose definition is not dependent upon the model chosen by these algorithms. The VI parameter we use is an answer to this need. We employ a machine learning algorithm to estimate the parameter, but differences in the model chosen by the algorithm do not change the definition of the parameter.

The semi-parametric approach is evolving, and recent advances promise to increase the power of this combination of machine learning and causal inference methods. We do not necessarily advocate the details of the semi-parametric VI algorithm used here – we in fact used a relatively inefficient method, and more refined methods are available to target model selection towards optimizing the particular parameter of interest (van der Laan & Gruber, 2009). We simply argue that it is possible to devise estimation strategies that, given unavoidable assumptions, can converge to unbiased estimates of the causal effects defined as sample size grows. In addition to the aforementioned alternate approaches for estimating our VI parameter, one can also use so-called asymptotically linear estimators; these are normally distributed, and in many cases simple standard errors based on this normality can be derived if one wishes to avoid re-sampling-based techniques (i.e. the bootstrap).

Risk factor epidemiology has for too long relied upon inherently biased techniques, particularly for observational data. There is no longer any reason to do so; the bias-reduction flexibility of semi-parametric models can be combined with estimation of simple and frankly more meaningful parameters in public health. We suggest using techniques that (1) define parameters with convenient public health interpretations, (2) use flexible, data-adaptive routines that do not pre-suppose arbitrary and scientifically unjustifiable models, and (3) employ honest inference that accounts for all the aspects of variation, including model selection.

Chapter 3

HIV-1 Genotypic Resistance Test Interpretation Algorithms and Virologic Suppression: Variable Importance and Prediction

3.1 Introduction

Prediction of HIV-1 virologic suppression after a treatment change is an active area of AIDS research. Genotypic resistance testing is thought to be predictive of the virologic response to a given treatment regimen, and is recommended for use by physicians in treating their HIV-1 infected patients (Ormaasen, Sandvik, Asjø, Holberg-Petersen, Gaarder, & Bruun, 2004; Hirsch, et al., 2003; Vandamme, et al., 2011). Interpretation of genotypic resistance tests, however, is challenging, particularly due to the complex interactions between drug-resistance mutations (Cabrera, et al., 2004; Schmidt, Walter, Zeitler, & Korn, 2002).

Multiple algorithms exist to interpret genotypic resistance tests. These algorithms produce a so-called genotypic susceptibility score (GSS) for each drug in a given treatment regimen, according to the patient's baseline HIV-1 genotype. These individual GSS can then be combined, typically using summation, to produce a GSS for the entire treatment regimen, or regimen-specific GSS (rGSS).

Various studies have investigated the value of the rGSS in predicting virologic suppression, employing traditional regression and correlation techniques as well as machine learning techniques (random forests) (Altmann, et al., 2009; Revell, et al., 2011). Most studies have found the rGSS to be at least somewhat predictive of virologic response (Rhee, et al., 2009; Frentz, et al., 2010; De Luca, et al., 2003; Anderson, et al., 2008). At least one large study, however, found the performance of the rGSS to be close to chance, and far inferior to the performance of random forest models populated by other features of the treatment regimen and other baseline covariates, such as treatment history, baseline viral load, and baseline CD4 count (Revell, et al., 2011). Comparisons of different genotypic resistance test interpretation algorithms have also revealed differences predictive performance between algorithms (Helm, et al., 2007; Assoumou, et al., 2008). There is also some evidence that prediction using the rGSS can be improved by weighting according to drug potency (Zazzi, et al., 2009; Fox, et al., 2007).

This analysis investigates (1) the association between the rGSS and virologic suppression, adjusted for other possible explanatory variables, and (2) the value of the rGSS in predicting virologic suppression. This requires two statistical approaches, since association with an outcome of interest does not guarantee inclusion in an optimal prediction model; conversely, predictive value does not imply the degree of association, nor can such value be easily translated into a causal framework for richer interpretation.

Causal inference-inspired semi-parametric variable importance techniques are ideal for investigating association, particularly in the presence of many correlated explanatory variables. A so-called population intervention parameter is defined here as the parameter of interest (Ahern, Hubbard, & Galea, 2009). Estimator choice is considered, by applying four different estimators and comparing them, particularly in terms of variance and bias induced by violation of the positivity or experimental treatment assignment (ETA) assumption (Petersen, Porter, Gruber, Wang, & van der Laan, 2012).

Predictive models including the rGSS along with other explanatory variables are constructed using super learning, which can employ multiple candidate models without requiring arbitrary

choice of a particular model by the investigator. These prediction models are then compared to a super learner prediction model without the rGSS (i.e. composed of the other potential explanatory variables only). Finally, rGSS is considered as the sole predictor of the virologic outcome. Data from 734 HIV-1 treatment regimens are analyzed, all from patients who had a treatment change within 24 weeks of a baseline genotypic resistance test performed at Stanford University between September 1998 and December 2007. Four genotypic resistance test interpretation algorithms and three rGSS weighting schemes are compared.

3.2 Methods

3.2.1 Data

This analysis was conducted on a retrospective dataset containing information from 641 patients, drawn from a patient population from 16 clinics of the Northern California Kaiser Permanente Medical Care Program. These patients had plasma HIV-1 samples sent to Stanford University Hospital for genotypic resistance testing between September 1998 and December 2007. Patients considered eligible for this analysis possessed a valid treatment-change episode (TCE) – defined as a change in antiretroviral (ARV) treatment regimen meeting the following criteria:

1) Treatment change occurred within 24 weeks of a baseline genotypic resistance test.
2) The new ARV regimen was administered for at least four weeks.
3) There existed at least one plasma HIV-1 RNA level >1000 copies/mL in the eight weeks prior to the treatment change.
4) At least two plasma HIV-1 RNA levels were obtained between 4 and 36 weeks after the treatment change.
5) A complete list of all ARVs ever received by the patient was available.
6) All ARVs in the new regimen were interpretable by all four genotypic interpretation algorithms (see below).

TCEs with new treatment regimens containing ARVs that were only licensed in the last months of the study period (raltegravir, maraviroc, etravirine) were excluded. The final dataset consisted of 734 eligible TCEs for 641 patients (Rhee, et al., 2009).

The outcome of interest is virologic response to the salvage regimen, according to the plasma HIV-1 RNA levels obtained in the 4 to 36 week follow-up period after treatment change. The outcome was defined as a binary value, and set to one if any plasma HIV-1 RNA level was below the limit of quantification (BLQ; <75 copies/mL, Versant bDNA assay), and zero otherwise (Rhee, et al., 2009).

Genotypic Susceptibility Scores

HIV-1 genotypic susceptibility scores (GSS) were provided for each TCE, according to four interpretation algorithms: (1) Agence National de Recerche sur le SIDA (ANRS) version 2007.10 (Agence National de Recerche sur le SIDA (ANRS)); (2) HIV Drug Resistance Database (HIVdb) (Liu & Shafer, 2006); (3) Rega version 7.1.1 (Van Laethem, De Luca,

Antinori, Cingolani, & Vandamme, 2002); and (4) ViroSeq version 2.8 (Eshleman, et al., 2004). Each algorithm produced a value between zero and one for each ARV in the new treatment regimen, representing the degree of predicted susceptibility of HIV-1 to the ARV. A value of 1.0 denotes full susceptibility and a value of 0 denotes full resistance to the ARV. For each algorithm, the GSS for enfuvirtide (fusion inhibitor) was set to 1.0 if the drug had not been taken previously, and set to 0 otherwise (Rhee, et al., 2009).

To create a single GSS composite value for the entire salvage regimen, or a regimen-specific GSS (rGSS), three weighted sums of the individual GSS were computed:

1) No weighting: the weights for each GSS were set to one.
2) Boosted PI weighting: the weights for each boosted protease inhibitor (PI) were set to 1.5.
3) Comprehensive weighting: the weights for each nucleoside reverse-transcriptase inhibitor (NRTI) were set to 0.5, and the weights for each boosted PI were set to 2.

Non-nucleoside reverse transcriptase inhibitors (NNRTIs), the fusion inhibitor enfuvirtide, and nelfinavir (an unboosted PI) were always assigned weights equal to one. The notion of weighting was motivated by the observation that not all ARVs are equally potent. The comprehensive weighting scheme in particular originated with the Rega algorithm, which unlike the other three algorithms provided instructions for calculating a weighted rGSS (Rhee, et al., 2009).

The four algorithms and the three weighting schemes defined 12 rGSS variants.

Explanatory Variables

In addition to the rGSS, other potential explanatory variables included demographic variables, information about the patient's treatment history and clinical status at baseline, and features of the new treatment regimen known at baseline and not directly related to the calculation of the rGSS. Table 3.1 lists these variables by category.

Table 3.1 – Potential explanatory variables

Variable Category	Variable Description
Demographic	Gender, age at baseline, ethnicity
Treatment history	Number of drugs in different classes received prior to baseline
	Number of treatment regimens received prior to baseline
	Number of Highly Active Antiretroviral Therapy (HAART) and non-HAART regimens received prior to baseline
	Number of ARVs received prior to baseline
	History of previous virologic suppression
	Year therapy began, number of years of therapy prior to baseline
Clinical status	Baseline viral load (log copies/mL) and CD4 count (cells/ml)
	Year of baseline viral load and CD4 count
	Year of baseline genotype
New treatment regimen	Number of total ARVs and number of ARVs in each drug class in the new regimen
	Number of new ARVs in the new regimen
	Whether or not a new drug class was introduced in the new regimen
	Year new regimen began

3.2.2 Variable Importance

Data Structure

The observed data $O = (W,A,Y)$ consists of three elements: W, the possible confounders or adjustment set (baseline covariates); A the variable of interest (target variable); and Y, the (binary) outcome. In this application W is the set of explanatory variables listed in Table 3.1, A is the rGSS, and Y is the binary virologic response after treatment change.

This observed data O follows some unknown distribution P_0, which is in turn a component of $\mathcal{M}$. Individual observations $O_1,O_2,\ldots,O_n$ can be defined as i.i.d. observations of O:

$$O_i = (W_i, A_i, Y_i),\ i \in \{1, 2, ..., n\}.$$

Note that the independent unit in this analysis is not the patient but the treatment change episode (TCE).

Because variable importance with a binary target variable A requires the least assumptions, the rGSS was dichotomized. Four dichotomization schemes were considered:

1) $A = 0$ when rGSS < 1, $A = 1$ otherwise
2) $A = 0$ when $1 \leq$ rGSS < 2, $A = 1$ otherwise
3) $A = 0$ when $2 \leq$ rGSS < 3, $A = 1$ otherwise
4) $A = 0$ when rGSS ≥ 3, $A = 1$ otherwise

This approach resulted in four possible values for A for each of the 12 rGSS variants, or 48 target variables in total.

The number of drugs in the new regimen and number of drugs in different drug classes in the new regimen define the maximum value the rGSS can achieve. For example, a TCE could not have an unweighted rGSS greater than two if the number of drugs in the new regimen were equal to two. Because each drug-specific GSS ranges between zero and one, the maximum unweighted rGSS for a given TCE is always equal to the number of drugs in the new regimen. For the other two weighting schemes, the maximum rGSS are given below:

$$\max\left(\mathrm{rGSS}_C\right) = 2 \cdot \mathrm{C}_{PI} + 0.5 \cdot \mathrm{C}_{NRTI} + \mathrm{C}_{TOTAL} - \mathrm{C}_{PI} - \mathrm{C}_{NRTI}$$
$$\max\left(\mathrm{rGSS}_{BPI}\right) = 1.5 \cdot \mathrm{C}_{PI} + C_{TOTAL} - \mathrm{C}_{PI}.$$

The boosted PI- and comprehensively weighted rGSS are represented above as rGSS_C and rGSS_{BPI}; C_{PI}, C_{NRTI}, and C_{TOTAL} represent the number of PIs, number of NRTIs, and total number of drugs in the new regimen, respectively.

For each dichotomization and weighting scheme, TCEs were excluded if their maximum rGSS limit precluded the possibility of achieving $A = 0$. This resulted in reduced sample sizes for dichotomization scheme (4) for the unweighted and boosted PI-weighted rGSS from 734 to 690.

For the comprehensively weighted rGSS, sample sizes were reduced from 734 to 712 for dichotomization scheme (3), and from 734 to 528 for dichotomization scheme (4).

Model and Target Parameter

The observed data *O* can be considered a missing data structure on the hypothetical full data *X*,

$$X = (W, Y_0, Y) \sim P_X .$$

W and *Y* are still the observed baseline covariates and the observed outcome, respectively. Y_0 is the counterfactual outcome when $A = 0$, which is observed for O_i with $A_i = 0$, but is missing for O_i with $A_i = 1$.

In the world of the hypothetical full data *X*, both *Y* and Y_0 are always observed. In the context of *X*, the following causal parameter can be defined:

$$\psi(P_X) = E[Y_0] - E[Y].$$

This is the parameter of interest in this analysis, and can be interpreted in the full data framework as the change in the observed mean outcome (virologic response) when *A* (the dichotomized rGSS) is set to zero.

The statistical parameter, or the parameter that can be defined under the observed data distribution P_0, is defined as follows:

$$\psi(P_0) = E_W\left[E(Y \mid A = 0, W)\right] - E[Y].$$

Additional assumptions are required in order for the equivalence $\psi(P_0) = \psi(P_X)$ to hold.

Assuming exogenous variables $U = (U_W, U_A, U_Y)$, we can define the following nonparametric structural equation model (NPSEM) for the endogenous full data *X*: (Pearl, 2000)

$$W = f_W(U_W)$$
$$A = f_A(W, U_A)$$
$$Y = f_Y(W, A, U_Y)$$

This NPSEM implies the so-called “no unmeasured confounding” assumption, (van der Laan & Robins, 2003)

$$A \perp Y_0 \mid W .$$

We must also assume that the observed data *O* are a missing data structure on *X* (consistency assumption) (van der Laan & Robins, 2003). Finally, we require the positivity or experimental

treatment assignment (ETA) assumption (van der Laan & Robins, 2003; Messer, Oakes, & Mason, 2010):

$$\Pr\left(A=0\,|\,W\right)>0\,.$$

This means that every combination of covariates must have a positive probability of having *A*=0. Under these assumptions, the statistical parameter is equivalent to the causal parameter.

Parameter Estimation

The statistical parameter of interest $\psi\left(P_0\right)$ consists of two elements, $E_W\left[E\left(Y\,|\,A=0,W\right)\right]$ and $E\left[Y\right]$. $E\left[Y\right]$ can be nonparametrically estimated by the empirical mean, $\bar{Y}$:

$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i.$$

Various approaches could be employed to estimate $E_W\left[E\left(Y\,|\,A=0,W\right)\right]$. In this analysis, four estimators are considered. First is the likelihood-based G-computation estimator (Robins, 1986; Robins, 2000; van der Laan & Rubin, 2006). If we define $Q_n^0(0,W)$ as $\hat{E}\left[Y\,|\,A=0,W\right]$, the G-computation estimator of $\psi\left(P_0\right)$ is given by

$$\psi_n^{G-COMP}=\frac{1}{n}\sum_{i=1}^{n}Q_n^0(0,W_i)-\bar{Y}.$$

The second and third estimators are derived using estimating equation methodology, and are the inverse probability of censoring-weighted (IPCW) estimator and its double-robust counterpart (DR-IPCW): (van der Laan & Robins, 2003; Hubbard & van der Laan, 2008)

$$\psi_n^{IPCW}=\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=0)}{g_n(0\,|\,W_i)}Y_i-\bar{Y}$$

$$\psi_n^{DR-IPCW}=\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=0)}{g_n(0\,|\,W_i)}\left(Y_i-Q_n^0(0,W_i)\right)+Q_n^0(0,W_i)-\bar{Y}.$$

The estimated probability that A_i = 0 given W_i is represented above as $g_n(0\,|\,W_i)$.

The fourth and final estimator is the targeted maximum likelihood estimator (TMLE), which is a combination of estimating equation and likelihood approaches (van der Laan & Rubin, 2006; van der Laan & Rose, 2011). It is defined as follows:

$$\psi_n^{TMLE} = \frac{1}{n}\sum_{i=1}^{n} Q_n^1(0, W_i) - \bar{Y}, \text{ where}$$
$$\text{logit}\left[Q_n^1(A,W)\right] = \text{logit}\left[Q_n^0(A,W)\right] + \varepsilon_n h(A,W), \text{ and}$$
$$h(A,W) = \frac{I(A=0)}{g_n(0 \mid W)}.$$

The parameter ε is estimated by maximum likelihood. TMLE's targeting step, the addition of *h(A,W)* to Q^0 and estimation of ε, is designed to reduce estimator bias in relation to the targeted feature of P_0 (the parameter $\psi(P_0)$) instead of focusing on the entire distribution.

The estimator ψ_n^{IPCW} will be consistent if the so-called "treatment mechanism" *g* is correctly specified, and ψ_n^{G-COMP} will be consistent if Q^0 is correctly specified. TMLE and DR-IPCW are double-robust, meaning that ψ_n^{TMLE} and $\psi_n^{DR-IPCW}$ will be consistent if either *g* or Q^0 are correctly specified. Consistency in this context means that the estimator ψ_n converges in probability to the true parameter $\psi(P_0)$ as $n \to \infty$. TMLE and DR-IPCW are also locally efficient, meaning that they are asymptotically efficient if the working model contains the true models.

Estimation of Q^0 and g

Q^0 was estimated by super learning, as implemented in the R package *SuperLearner*. *SuperLearner* uses *V*-fold cross-validation to construct a convex combination of candidate estimators. It is most desirable to choose an estimation approach that (1) respects what is known about the true form of Q^0 (nothing), (2) considers multiple models and utilizes machine learning to come as close as "possible" to the true Q^0, and (3) avoids manual manipulation of the data in choosing the final model. Super learning meets all of these criteria, and in addition has the theoretical property of performing as well asymptotically as the so-called "oracle" selector which, in the context of a particular loss function, minimizes risk under the true data-generating distribution (Sinisi S. , Polley, Petersen, Rhee, & van der Laan, 2007; van der Laan, Polley, & Hubbard, 2007).

For this analysis, the function *SuperLearner* was used with 7-fold cross-validation. The library of candidate estimators for the super learner included the following: main terms logistic regression (R function *glm*); logistic regression with the dichotomized rGSS as the sole predictor; generalized additive models (as implemented in the R package *gam*); stepwise logistic regression, with all main terms as the maximum size model (as implemented in the R package step); and polychotomous regression and multiple classification (as implemented in the R package polspline, 5-fold cross-validation) (Kooperberg, Bose, & Stone, 1997). The dichotomized rGSS (*A*) was required to be present in the model chosen by each candidate estimator.

The function *polyclass* in the *polspline* package does not easily allow for covariates to be forced into the final model. For this reason, a workaround was constructed to allow use of *polyclass* and also ensure that the dichotomized rGSS would always be present in the model selected. First,

polyclass was fit on the entire dataset (or, in the case of the super learner, on the training dataset). Second, the predicted probability of *Y =1* was obtained, per the *polyclass* fit. Finally, this fit was used in a logistic regression model containing the dichotomized rGSS:

$$\text{logit}\left[E\left(Y \mid A,W\right)\right] = \gamma_0 + \gamma_1 A + \gamma_2 \text{logit}\left[Z_n\left(A,W\right)\right] \cdot A + \gamma_3 \text{logit}\left[Z_n\left(A,W\right)\right],$$

where $Z_n\left(A,W\right)$ represents the fitted probabilities from *polyclass*.

The generalized additive model used smoothing splines with two target degrees of freedom for covariates with more than four unique values, and linear terms for all other covariates. These are the default specifications of for the *gam* function according to the R function *SuperLearner*.

It would have been possible to also estimate *g* using super learning. When the sample size is relatively modest in comparison with the number of potential explanatory variables, however, super learning can be overly aggressive and result in predicted probabilities when $A = 0$ close to zero, which are practical violations of the ETA assumption mentioned earlier. With this consideration in mind, *g* was estimated using forward stepwise logistic regression (R function step). Collaborative TMLE (C-TMLE) can also be used to combat this issue, but was not applied in this analysis (van der Laan & Gruber, 2009). Predicted probabilities of $A = 0$ were truncated when necessary at 0.01.

Data-adaptive restrictions on the adjustment sets for Q^0 and *g* were also performed, to account for data sparsity and in an attempt to reduce ETA violations. Specifically, binary explanatory variables were excluded from *W* in the estimation of Q^0 if less than 10 TCEs were observed to have any explanatory variable/outcome combination. Explanatory variables with less than 20 unique values were also excluded if one outcome was not observed for more than 30% of the possible unique values. Any explanatory variables excluded from Q^0 were also excluded from *g*. The adjustment set for *g* was additionally restricted using the same logic described above, but with *A* substituted for the outcome. An automated process was chosen over manual review and pre-specification of *W* because of both the large number of target variables (relevant only to the restriction of *g*) and so that the variability induced by the data-adaptive restriction of *W* could be mimicked in the nonparametric and parametric bootstraps to be subsequently described.

According to this method, five potential explanatory variables were excluded from Q^0 in all analyses: (1) the number of regimens received prior to baseline, (2) the number of HAART regimens received prior to baseline, (3) Asian ethnicity, (4) unknown ethnicity, and (5) the presence of a fusion inhibitor in the new treatment regimen.

Inference

Standard errors for all four estimators can be estimated using the nonparametric bootstrap. For G-computation, this is the only option. For IPCW, DR-IPCW, and TMLE, standard errors can also be approximated using the estimated influence curve (IC), assuming the sample size *n* is large enough that reliance upon asymptotic results is reasonable. For *n* large enough, ψ_n^{IPCW}, $\psi_n^{DR-IPCW}$, and ψ_n^{TMLE} will be approximately normal, with variance equal to the variance of the

appropriate IC, divided by $\sqrt{n}$. The influence curves of each estimator, as estimated under the empirical distribution P_n, are given below (van der Laan & Robins, 2003; van der Laan & Rose, 2011).

$$IC_n^{TMLE}(O) = \left(\frac{I(A=0)}{g_n(0 \mid W)} \right) \left(Y - Q_n^1(0,W) \right) + Q_n^1(0,W) - Y - \psi_n^{TMLE}$$

$$IC_n^{DR-IPCW}(O) = \left(\frac{I(A=0)}{g_n(0 \mid W)} \right) \left(Y - Q_n^0(0,W) \right) + Q_n^0(0,W) - Y - \psi_n^{DR-IPCW}$$

$$IC_n^{IPCW}(O) = \left(\frac{I(A=0)}{g_n(0 \mid W)} - 1 \right) Y - \psi_n^{IPCW}.$$

For comparison, standard errors were also calculated using the nonparametric bootstrap with 500 iterations for the 16 comprehensively weighted rGSS variables (4 interpretation algorithms, 4 dichotomization schemes). The comprehensive weighting scheme was chosen because the final sample sizes were smallest. To mimic the study design, the full dataset was re-sampled with replacement in each bootstrap sample, and the dataset was then restricted where necessary to prevent theoretical ETA violations. This means that the number of observations in each bootstrap sample could vary for the dichotomizations for which restriction of the full dataset was necessary. Bootstrap-derived standard errors were not calculated in all cases due to the large number (48) of rGSS variants.

3.2.3 Bias Diagnostic

Practical violations of the ETA assumption were observed during the parameter estimation process. It was therefore desirable to obtain an estimate of the possible bias induced by these violations in positivity. The parametric bootstrap diagnostic of Wang et al. (2006) can be used to diagnose the presence of bias induced by such ETA assumption violations (Wang, Petersen, Bangsberg, & van der Laan, 2006; Petersen, Porter, Gruber, Wang, & van der Laan, 2012). The diagnostic consists of three main steps:

Step 1: Estimate P_0. This involves estimating Q^0 and g, which was already done during the parameter estimation process (previous section). The same estimates were used again during the bias diagnostic process. The true target parameter $\psi\left(\hat{P}_0\right)$ under the bootstrap-generating distribution is defined as the maximum likelihood estimator applied to the observed data, which is equivalent to the G-computation estimator described in the previous section (ψ_n^{G-COMP}).

Step 2: Generate $P^{\#}$ by sampling from $\hat{P}_0$. In this step, bootstrap samples $P^{\#}$ are generated from $\hat{P}_0$. Each sample $P^{\#}$ consists of n i.i.d. observations $O^{\#} = \left(W^{\#}, A^{\#}, Y^{\#}\right) \sim \hat{P}_0$. For each bootstrap sample $P^{\#}$, $W^{\#}$ was generated first by sampling from the empirical distribution, i.e. by

sampling the rows of W with replacement. Next, g_n was applied to $W^{\#}$, and the resulting predicted probabilities used to generate $A^{\#}$ as Bernoulli random variables with probability $p_a = g_n\left(1 \mid W^{\#}\right)$. Finally, Q_n^0 was applied to $A^{\#}$ and $W^{\#}$ to generate $Y^{\#}$ as Bernoulli random variables with probability $p_y = Q_n^0\left(1 \mid A^{\#}, W^{\#}\right)$.

Step 3: Estimate $E_{\hat{P}_0}\left[\psi_n\left(P^{\#}\right)\right]$. The final step involves applying the entire estimation process to $P^{\#}$ as if it were the true dataset. $E_{\hat{P}_0}\left[\psi_n\left(P^{\#}\right)\right]$ is then estimated by taking the mean of the estimator ψ_n across bootstrap samples. This mean is compared with the true parameter as defined in Step 1 to calculate the ETA bias:

$$\text{Bias}_{ETA} = E_{\hat{P}_0}\left[\psi_n\left(P^{\#}\right)\right] - \psi\left(\hat{P}_0\right).$$

This parametric bootstrap diagnostic was applied with 500 iterations to each of the three estimators that utilize g (IPCW, DR-IPCW, TMLE), for the 10 rGSS dichotomizations for which the minimum value of $g_n\left(0 \mid W\right)$ was closest to zero. Time constraints prevented the application of the diagnostic in all 24 cases when truncation of $g_n\left(0 \mid W\right)$ at 0.01 was necessary.

3.2.4 Prediction

To investigate the predictive value of the rGSS, the super learner was used to fit prediction models including each of the 12 rGSS options (4 algorithms, 3 weighting schemes), along with the other explanatory variables. Prediction models were also constructed containing only the other explanatory variables (no rGSS), and, alternately, only the rGSS. Since the rGSS weighting schemes were motivated by differences in ARV potency, prediction models were also constructed for each algorithm containing sums of GSS by drug class, which allowed the super learner to dynamically choose the best weights by drug class. This "dynamic weighting" was included in models both with and without the other explanatory variables. Finally, as a simplest possible approach, the 12 rGSS variables were scaled so they ranged between 0 and 1, and this scaled value used as the predicted probability. Each rGSS was scaled to fall between 0 and 1 by dividing it by its maximum achievable value. For the unweighted rGSS, this maximum was the number of drugs in the new regimen; for the boosted PI- and comprehensively weighted rGSS, the maximum values were as described previously (as determined by the number of drugs in different classes present in the new regimen).

Super learning with 7-fold cross-validation was used in fitting all prediction models with the exception of those containing the rGSS only; for those models, logistic regression was used. The super learner library of candidate estimators included main terms logistic regression (*glm*), generalized additive models (*gam*), polychotomous regression and multiple classification (*polyclass*, 5-fold cross-validation), and stepwise logistic regression with all main terms as the maximum size model (*step*). The specifications for the generalized additive models implementation *gam* were the same as described previously. The function *polyclass* was used

without modification in this context, because the rGSS was not required to be present in the chosen prediction model.

Predicted probabilities of virologic suppression were estimated for each prediction model using 10-fold cross validation, and receiver operating characteristic (ROC) curves were constructed from the predicted probabilities (R package ROCR) (Sing, Sander, Beerenwinkel, & Lengauer, 2005). ROC curves depict graphically the tradeoff between the true positive rate (correct prediction of virologic suppression among those that achieved virologic suppression) and the false positive rate (incorrect prediction of virologic suppression among those that did not achieve virologic suppression). Because predicted probabilities were calculated within each validation sample, the resulting rate estimates will be unbiased for sample size $n(1-1/V)$, where V is the number of cross-validation folds. In this analysis, $n(1-1/V) = 734(1-1/10) \approx 660$.

3.3 Results

3.3.1 Variable Importance

Tables 3.2, 3.3, and 3.4 show the counts and percentages of the 734 TCEs defined as having $A = 0$ for each rGSS dichotomization scheme. In terms of the unweighted rGSS, one unit can be thought of as one fully active ARV (i.e. one ARV to which HIV-1 is predicted to be fully susceptible); for the boosted PI weighting scheme, one unit is equivalent to one two-thirds active boosted PI, or one fully active ARV of any other drug class; for the comprehensive weighting scheme, one unit is equivalent to one half-active boosted PI, two fully active NRTIs, or one fully active ARV of another drug class.

Table 3.2 - Unweighted rGSS. Counts and percentages of TCEs with $A = 0$ for each dichotomization.

Definition of $A = 0$	HIVdb		Rega		ViroSeq		ANRS		Eligible TCEs
rGSS < 1	105	*14.3%*	56	*7.6%*	68	*9.3%*	46	*6.3%*	734
1 ≤ rGSS < 2	254	*34.6%*	168	*22.9%*	181	*24.7%*	127	*17.3%*	734
2 ≤ rGSS < 3	226	*30.8%*	259	*35.3%*	233	*31.7%*	253	*34.5%*	734
rGSS ≥ 3	149	*21.6%*	251	*36.4%*	252	*36.5%*	308	*44.6%*	690

Table 3.3 - Boosted PI-weighted rGSS. Counts and percentages of TCEs with $A = 0$ for each dichotomization.

Definition of $A = 0$	HIVdb		Rega		ViroSeq		ANRS		Eligible TCEs
rGSS < 1	90	*12.3%*	42	*5.7%*	68	*9.3%*	46	*6.3%*	734
1 ≤ rGSS < 2	222	*30.2%*	121	*16.5%*	149	*20.3%*	111	*15.1%*	734
2 ≤ rGSS < 3	239	*32.6%*	256	*34.9%*	231	*31.5%*	242	*33.0%*	734
rGSS ≥ 3	183	*26.5%*	315	*45.7%*	286	*41.4%*	335	*48.6%*	690

Table 3.4 - Comprehensively weighted rGSS. Counts and percentages of TCEs with $A = 0$ for each dichotomization.

Definition of $A = 0$	HIVdb		Rega		ViroSeq		ANRS		Eligible TCEs
rGSS < 1	118	*16.1%*	69	*9.4%*	86	*11.7%*	80	*10.9%*	734
1 ≤ rGSS < 2	264	*36.0%*	172	*23.4%*	200	*27.2%*	170	*23.2%*	734
2 ≤ rGSS < 3	246	*34.6%*	284	*39.9%*	267	*37.5%*	255	*35.8%*	712
rGSS ≥ 3	106	*20.1%*	209	*39.6%*	181	*34.3%*	229	*43.4%*	528

The unweighted rGSS for ANRS, Rega, and ViroSeq ranged from 0 to 5, and from 0 to 4.25 for HIVdb. The mean unweighted rGSS was 1.9 for HIVdb, 2.2 for Rega, 2.1 for ViroSeq, and 2.3 for ANRS. The boosted PI-weighted rGSS ranged from 0 to 5.5 for ANRS, Rega, and ViroSeq, and from 0 to 4.5 for HIVdb. The mean boosted PI-weighted rGSS was 2.5 for Rega, 2.1 for HIVdb, 2.4 for ViroSeq, and 2.5 for ANRS. Finally, the comprehensively weighted rGSS ranged from 0 to 4.5 for ANRS, Rega, and ViroSeq, and from 0 to 4 for HIVdb. The mean comprehensively weighted rGSS was 2.2 for Rega, 1.8 for HIVdb, 2.0 for ViroSeq, and 2.1 for ANRS.

Tables 3.5, 3.6, and 3.7 show the variable importance estimates for the unweighted, boosted PI-weighted, and comprehensively weighted rGSS dichotomizations, by estimator and genotypic resistance test interpretation algorithm. In the counterfactual world, each estimated parameter would be interpreted as the change in the observed rate of virologic suppression if every regimen had an rGSS in the indicated range. For example, for $A = 0$ when the unweighted rGSS < 1 as calculated by the Stanford HIVdb algorithm (Table 3.5), the DR-IPCW estimate was -0.224, with a 95% confidence interval of (-0.37, -0.08). Interpreted in the causal framework, this estimate implies that were every TCE to have an unweighted rGSS of less than 1 (less than one fully active ARV) according to the HIVdb algorithm, the frequency of virologic suppression would go down from 64.7% to 42.3%, with a 95% confidence interval (CI) ranging from 28% to 57%.

The effect estimates for $A = 0$ when the unweighted rGSS < 1 are all negative, implying a deleterious effect of an unweighted rGSS less than one. The 95% CIs for the ANRS algorithm all cross zero, while the other 95% CIs do not. The largest negative effect estimate belongs to the Rega algorithm; the causal interpretation of the IPCW estimate (-0.432) would be that were all TCEs to have an unweighted rGSS < 1 as estimated by the Rega algorithm, the proportion of TCEs after which virologic suppression was achieved would go down from 64.7% to 21.5%, with a 95% CI ranging from 0% to 44%. This is a very broad 95% CI, and the estimated standard error is more than one quarter the size of the effect estimate. The largest negative effect estimate among the double robust estimators is -0.304 for TMLE, corresponding to a reduction in the rage of virologic suppression from 64.7% to 34.3%, with a 95% CI ranging from 22% to 48%. For the dichotomization $A = 0$ when $1 \leq$ unweighted rGSS < 2, the parameter estimates are also all negative, but most 95% CIs either include zero or almost include zero. The estimates with 95% CIs farthest from zero belong to the ANRS algorithm. For $A = 0$ when $2 \leq$ unweighted rGSS < 3, the parameter estimates are all positive, but once again, most 95% CIs either include zero or nearly include zero, and the estimates themselves are close to zero in most cases. The estimates farthest from zero belong to the HIVdb algorithm (across all estimators). Finally, for $A = 0$ when rGSS ≥ 3, the estimates are also all positive, and the 95% CIs for DR-IPCW and TMLE do not cross zero for any genotypic resistance test interpretation algorithm. The 95% CIs for IPCW do include or very nearly include zero for all algorithms. The largest positive effect estimate is for the ViroSeq algorithm – the causal interpretation of the TMLE estimate (0.126) would be that were all TCEs to have rGSS ≥ 3 according to the ViroSeq algorithm, the proportion of TCEs after which virologic suppression was achieved would go up from 64.7% to 77.3%, with a 95% CI ranging from 73% to 83%.

The estimates from the boosted PI-weighted rGSS (Table 3.6) are similar to the estimates from the unweighted rGSS. For $A = 0$ when rGSS < 1, the effect estimates are all negative, and most 95% CIs do not cross, zero, excepting the ANRS algorithm. No genotypic resistance test interpretation algorithm consistently yields the largest effect estimate across all estimators. The TMLE estimate farthest from zero (-0.276) corresponds to the Rega algorithm. This estimate's causal interpretation implies a reduction in virologic suppression from 64.7% to 31.7%, with a 95% CI ranging from 24% to 50%, were all TCEs in this analysis to have a boosted PI-weighted rGSS less than one, as interpreted by the Rega algorithm. For the dichotomization $A = 0$ when $1 \leq$ boosted PI-weighted rGSS < 2, the estimates are once again all negative, but with 95% CIs that cross zero in most cases. The exceptions correspond to the ANRS algorithm and for DR-IPCW also to the Rega algorithm. The DR-IPCW Rega estimate (-0.240, 95% CI -0.42 to -0.06) is more than two times larger than the DR-IPCW ANRS estimate (-0.106, 95% CI -0.18 to -0.03), and is the largest effect estimate across all algorithms and estimators for this dichotomization. The causal interpretation of the Rega estimate implies a reduction in the rate of virologic suppression from 64.7% to 40.7% (95% CI 23% to 59%), if causal assumptions hold. For $A = 0$ when $2 \leq$ boosted PI – weighted rGSS < 3, all estimates are positive but near zero for all but the HIVdb algorithm. The HIVdb estimates also have the only 95% CIs that do not include zero. The DR-IPCW and TMLE effect estimates for HIVdb, while still small, are more than three times larger than the next largest effect estimate for another genotypic resistance test interpretation algorithm; the IPCW estimate is more than two times larger than the next largest estimate. The DR-IPCW and TMLE estimates are 0.091 and 0.092, respectively, implying an increase in the rate of virologic suppression from 64.7% to 74%, if causal assumptions hold (95% CI for both estimators 69% to 80%). For the final dichotomization ($A = 0$ when boosted PI-weighted rGSS $\geq$ 3), all parameter estimates are again positive, and all 95% CIs for IPCW include zero, while none of the 95% CIs include zero for DR-IPCW and TMLE. The largest effect estimate with a 95% CI excluding zero is the TMLE estimate for the ViroSeq algorithm (0.127, or an increase in virologic suppression from 64.7% to 77.4%), and is very similar to the corresponding unweighted rGSS estimate (0.126). The 95% confidence intervals for the two estimates (boosted PI-weighted and unweighted) are almost the same (0.09 to 0.17, or 74% to 82% for boosted PI; 0.08 to 0.17, or 73% to 82% for unweighted).

For the comprehensive weighting scheme (Table 3.7), the first dichotomization ($A = 0$ when comprehensively weighted rGSS < 1) yields parameter estimates that are again all negative, and mostly significant at the 5% level (95% CIs do not include zero). 95% CIs for ViroSeq include zero for IPCW, DR-IPCW, and TMLE, as does the IPCW 95% CI for HIVdb. The genotypic resistance test interpretation algorithm with the largest parameter estimates overall (for this dichotomization) is ANRS. Its DR-IPCW estimate is -0.272, corresponding to a decrease in virologic suppression from 64.7% to 37.5%, with a 95% CI ranging from 24% to 52%, were all TCEs to have a comprehensively weighted rGSS less than one according to the ANRS algorithm (and assuming causal assumptions hold). For the second dichotomization ($A = 0$ when $1 \leq$ comprehensively weighted rGSS < 2), the estimates for ViroSeq and Rega are negative, with 95% CIs that do not include zero, with the one exception of the IPCW estimate for ViroSeq. The estimates for ANRS are also negative, but their 95% CIs do include zero for all estimators. The HIVdb-associated 95% confidence intervals include zero as well, and the parameter estimates themselves are very nearly zero, much smaller than for any other genotypic resistance test interpretation algorithm. For the third dichotomization ($A = 0$ when $2 \leq$ comprehensively

weighted rGSS < 3), the parameter estimates are all small and positive. The 95% CIs for the IPCW estimator all include zero. In contrast, for DR-IPCW, G-computation, and TMLE , only the ANRS 95% CIs include zero. The largest parameter estimates across estimators with 95% CIs that exclude zero for this dichotomization belong to the HIVdb algorithm. This aligns with the boosted PI results for the equivalent dichotomization, though the DR-IPCW and TMLE estimates are slightly smaller than the equivalent boosted PI estimates for HIVdb. Finally, for the fourth dichotomization (comprehensively weighted rGSS ≥ 3), all estimates are positive and all TMLE 95% CIs exclude zero, while all IPCW 95% confidence intervals either include or very nearly include zero. The G-computation and DR-IPCW 95% CIs include zero for HIVdb, and exclude zero for the other three genotypic resistance test interpretation algorithms. The largest effect estimate corresponds, once again, to the ViroSeq algorithm. Its TMLE estimate is 0.118, implying an increase in the rate of virologic suppression from 64.7% to 76.5%, with a 95% CI ranging from 72% to 81%, assuming causal assumptions hold. This is smaller than the equivalent estimates for the boosted PI-weighted and unweighted rGSS.

The parameter estimates for DR-IPCW and TMLE are of similar magnitude in most cases, while the estimates for G-computation and IPCW vary in comparison; the IPCW and G-computation estimates are in many cases quite similar to those of DR-IPCW and TMLE, and sometimes markedly larger or smaller. Standard errors overall are somewhat large in comparison with the parameter estimates, and in a number of instances are actually larger in magnitude than the estimates themselves. The standard errors for DR-IPCW and TMLE are mostly comparable to each other for each parameter definition. In the seven cases where a difference in standard error greater than 0.009 is noted between DR-IPCW and TMLE, the TMLE standard errors are always smaller. These seven occurrences were all attributable to rGSS dichotomizations where the minimum value of $g_n(0|W)$ was closest to zero. The G-computation standard errors are smaller than both their DR-IPCW and TMLE counterparts for rGSS < 1 and 1 ≤ rGSS < 2 in Table 3.7, and are comparable to the DR-IPCW and TMLE standard errors for the other two dichotomizations. The estimated standard errors for IPCW are consistently larger in all instances than those of the other estimates, often twice as large. This results in 95% CIs for IPCW that cross zero in a number of cases where those of the other estimators do not.

Tables 3.8, 3.9 and 3.10 show a comparison of influence curve- with bootstrap-derived standard errors and confidence intervals for the comprehensively weighted rGSS dichotomizations for IPCW, DR-IPCW, and TMLE, respectively. The sample sizes for the dichotomization $A = 0$ when 2 ≤ rGSS < 3 ranged from 698 to 727 across bootstrap samples, and for rGSS ≥ 3 the sample sizes ranged from 496 to 566. Table 3.8 shows that the bootstrap standard errors for IPCW are consistently smaller than the influence curve-derived standard errors – this is true in all but one case (ANRS, rGSS < 1). The differences are largest where the minimum value of $g_n(0|W)$ was observed to be closest to zero (rGSS < 1). In two instances the difference in estimated standard error resulted in a bootstrap-based 95% CI that did not include zero while the influence curve-based 95% CI did include zero, though the bootstrap-based 95% CIs come very close to zero (HIVdb and ViroSeq, rGSS < 1). For DR-IPCW and TMLE (Tables 3.9 and 3.10), bootstrap standard errors are mostly comparable to the influence curve-derived standard errors. For the DR-IPCW estimator (Table 3.9), differences between standard error estimates larger than 0.009 were observed in two cases, once with the influence curve-derived standard error larger (ViroSeq, rGSS < 1), once with the bootstrap standard error larger (HIVdb, rGSS ≥ 3). For

TMLE (Table 3.10), differences larger than 0.009 between standard error estimates were observed in four cases, with the bootstrap standard error larger in each case (Rega and ANRS, rGSS < 1; HIVdb and ViroSeq, rGSS ≥ 3). In no case did the TMLE or DR-IPCW bootstrap 95% CIs return different inference from the influence curve-derived 95% CIs.

Wald-type 95% CIs are also compared in Tables 3.8, 3.9, and 3.10 to 95% CIs obtained using quantiles of the bootstrap distribution. Overall, the two methods return 95% CIs that are quite similar. The largest differences between the two types of 95% CIs are for the dichotomization A=0 when rGSS < 1(ANRS for IPCW; Rega for DR-IPCW; ViroSeq for TMLE). Some of the 95% CIs for this dichotomization also include zero by one method and do not include zero by the other. In these cases the standard errors are also largest, meaning that the 95% CIs are broad and one limit is near zero. The Wald-type and quantile-based bootstrap 95% CIs were comparable for G-computation as well (data not shown), with no change in inference for any estimate when one method for calculating the 95% CI was used over the other.

Figures 3.1, 3.2 and 3.3 depict graphically the TMLE parameter estimates across the different rGSS dichotomizations and weighting schemes. In all three Figures, the estimated parameters across genotypic resistance test interpretation algorithms start out negative (corresponding to A = 0 when rGSS < 1), and become progressively more positive with each dichotomization.

The results of the parametric bootstrap bias diagnostic are shown in Table 3.11. Bias_{ETA} for IPCW is larger in all cases than Bias_{ETA} for either DR-IPCW or TMLE; in most cases, the estimated bias for the IPCW estimator is an order of magnitude larger than for the other two estimators. The largest absolute difference in Bias_{ETA} between DR-IPCW and TMLE is for HIVdb, boosted PI-weighted rGSS < 1, with DR-IPCW having the smallest estimated bias. For all estimators, Bias_{ETA} is small in comparison with the estimated standard errors.

Table 3.5 – Unweighted rGSS. Variable importance estimates by genotypic resistance test interpretation algorithm and estimator. The causal parameter of interest is the difference in probability of virologic suppression if all salvage regimens had rGSS in the indicated range versus observed values.

rGSS < 1										
	G-computation	IPCW			DR-IPCW			TMLE		
	Estimate	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	-0.271	-0.360	0.138	(-0.63, -0.09)	-0.224	0.074	(-0.37, -0.08)	-0.192	0.058	(-0.31, -0.08)
Rega	-0.250	-0.432	0.111	(-0.65, -0.21)	-0.275	0.06	(-0.39, -0.16)	-0.304	0.066	(-0.43, -0.17)
ViroSeq	-0.310	-0.405	0.143	(-0.69, -0.12)	-0.262	0.076	(-0.41, -0.11)	-0.232	0.060	(-0.35, -0.11)
ANRS	-0.302	-0.314	0.179	(-0.67, 0.04)	-0.195	0.128	(-0.45, 0.06)	-0.155	0.088	(-0.33, 0.02)
1 ≤ rGSS < 2										
	G-computation	IPCW			DR-IPCW			TMLE		
	Estimate	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	-0.011	-0.017	0.051	(-0.12, 0.08)	-0.012	0.027	(-0.06, 0.04)	-0.012	0.027	(-0.06, 0.04)
Rega	-0.080	-0.119	0.074	(-0.26, 0.03)	-0.101	0.047	(-0.19, -0.01)	-0.106	0.046	(-0.20, -0.02)
ViroSeq	-0.051	-0.136	0.058	(-0.25, -0.02)	-0.048	0.03	(-0.11, 0.01)	-0.046	0.030	(-0.11, 0.01)
ANRS	-0.095	-0.202	0.072	(-0.34, -0.06)	-0.114	0.038	(-0.19, -0.04)	-0.130	0.038	(-0.2, -0.06)
2 ≤ rGSS < 3										
	G-computation	IPCW			DR-IPCW			TMLE		
	Estimate	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	0.094	0.095	0.050	(0.00, 0.19)	0.102	0.025	(0.05, 0.15)	0.103	0.025	(0.05, 0.15)
Rega	0.060	0.071	0.044	(-0.02, 0.16)	0.061	0.024	(0.01, 0.11)	0.061	0.024	(0.01, 0.11)
ViroSeq	0.039	0.054	0.047	(-0.04, 0.15)	0.050	0.024	(0.00, 0.10)	0.050	0.024	(0.00, 0.10)
ANRS	0.016	0.013	0.044	(-0.07, 0.10)	0.021	0.023	(-0.02, 0.07)	0.023	0.023	(-0.02, 0.07)
rGSS ≥ 3										
	G-computation	IPCW			DR-IPCW			TMLE		
	Estimate	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	0.040	0.084	0.100	(-0.11, 0.28)	0.097	0.04	(0.02, 0.17)	0.103	0.032	(0.04, 0.16)
Rega	0.054	0.089	0.064	(-0.04, 0.21)	0.088	0.025	(0.04, 0.14)	0.088	0.024	(0.04, 0.13)
ViroSeq	0.081	0.091	0.058	(-0.02, 0.20)	0.116	0.025	(0.07, 0.17)	0.126	0.024	(0.08, 0.17)
ANRS	0.083	0.101	0.048	(0.01, 0.19)	0.111	0.022	(0.07, 0.15)	0.114	0.022	(0.07, 0.16)

Inference for G-computation (via the nonparametric bootstrap) was not obtained for these estimates. Inference for the other estimators was obtained using the influence curve.

Table 3.6 – Boosted PI-weighted rGSS. Variable importance estimates by genotypic resistance test interpretation algorithm and estimator. The causal parameter of interest is the difference in probability of virologic suppression if all salvage regimens had rGSS in the indicated range versus observed values.

rGSS < 1										
	G-computation	IPCW			DR-IPCW			TMLE		
	Estimate	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	-0.347	-0.400	0.143	(-0.68, -0.12)	-0.295	0.083	(-0.46, -0.13)	-0.262	0.062	(-0.38, -0.14)
Rega	-0.245	-0.422	0.115	(-0.65, -0.20)	-0.261	0.064	(-0.39, -0.14)	-0.276	0.067	(-0.41, -0.15)
ViroSeq	-0.312	-0.405	0.143	(-0.69, -0.12)	-0.266	0.076	(-0.42, -0.12)	-0.239	0.061	(-0.36, -0.12)
ANRS	-0.300	-0.314	0.179	(-0.67, 0.04)	-0.193	0.129	(-0.45, 0.06)	-0.153	0.089	(-0.33, 0.02)
1 ≤ rGSS < 2										
	G-computation	IPCW			DR-IPCW			TMLE		
	Estimate	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	-0.030	-0.046	0.073	(-0.19, 0.10)	-0.033	0.037	(-0.10, 0.04)	-0.032	0.037	(-0.10, 0.04)
Rega	-0.149	-0.146	0.176	(-0.49, 0.20)	-0.240	0.091	(-0.42, -0.06)	-0.191	0.093	(-0.37, -0.01)
ViroSeq	-0.066	-0.132	0.082	(-0.29, 0.03)	-0.056	0.036	(-0.13, 0.01)	-0.054	0.035	(-0.12, 0.01)
ANRS	-0.096	-0.239	0.081	(-0.40, -0.08)	-0.106	0.038	(-0.18, -0.03)	-0.116	0.039	(-0.19, -0.04)
2 ≤ rGSS < 3										
	G-computation	IPCW			DR-IPCW			TMLE		
	Estimate	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	0.100	0.108	0.048	(0.01, 0.20)	0.091	0.028	(0.04, 0.15)	0.092	0.027	(0.04, 0.15)
Rega	0.022	0.038	0.049	(-0.06, 0.13)	0.022	0.028	(-0.03, 0.08)	0.022	0.028	(-0.03, 0.08)
ViroSeq	0.027	0.020	0.047	(-0.07, 0.11)	0.022	0.026	(-0.03, 0.07)	0.022	0.026	(-0.03, 0.07)
ANRS	0.023	0.026	0.047	(-0.07, 0.12)	0.029	0.023	(-0.02, 0.07)	0.029	0.023	(-0.02, 0.07)
rGSS ≥ 3										
	G-computation	IPCW			DR-IPCW			TMLE		
	Estimate	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	0.059	0.111	0.092	(-0.07, 0.29)	0.110	0.037	(0.04, 0.18)	0.112	0.032	(0.05, 0.17)
Rega	0.093	0.070	0.041	(-0.01, 0.15)	0.103	0.019	(0.07, 0.14)	0.107	0.019	(0.07, 0.15)
ViroSeq	0.094	0.102	0.052	(0.00, 0.20)	0.119	0.022	(0.08, 0.16)	0.127	0.021	(0.09, 0.17)
ANRS	0.069	0.067	0.039	(-0.01, 0.14)	0.074	0.022	(0.03, 0.12)	0.074	0.023	(0.03, 0.12)

Inference for G-computation (via the nonparametric bootstrap) was not obtained for these estimates. Inference for the other estimators was obtained using the influence curve.

Table 3.7 – Comprehensively weighted rGSS. Variable importance estimates by genotypic resistance test interpretation algorithm and estimator. The causal parameter of interest is the difference in probability of virologic suppression if all regimens had rGSS in the indicated range versus observed values.

rGSS < 1												
	G-computation			IPCW			DR-IPCW			TMLE		
	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	-0.315	0.056	(-0.43, -0.21)	-0.286	0.150	(-0.58, 0.01)	-0.232	0.095	(-0.42, -0.05)	-0.239	0.091	(-0.42, -0.06)
Rega	-0.208	0.068	(-0.34, -0.07)	-0.331	0.157	(-0.64, -0.02)	-0.250	0.098	(-0.44, -0.06)	-0.239	0.101	(-0.44, -0.04)
ViroSeq	-0.241	0.062	(-0.36, -0.12)	-0.256	0.188	(-0.62, 0.11)	-0.166	0.124	(-0.41, 0.08)	-0.187	0.110	(-0.40, 0.03)
ANRS	-0.334	0.063	(-0.46, -0.21)	-0.377	0.102	(-0.58, -0.18)	-0.272	0.072	(-0.41, -0.13)	-0.233	0.064	(-0.36, -0.11)
1 ≤ rGSS < 2												
	G-computation			IPCW			DR-IPCW			TMLE		
	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	0.006	0.019	(-0.03, 0.04)	0.017	0.056	(-0.09, 0.13)	0.006	0.030	(-0.05, 0.06)	0.006	0.030	(-0.05, 0.07)
Rega	-0.108	0.040	(-0.19, -0.03)	-0.194	0.093	(-0.38, -0.01)	-0.180	0.056	(-0.29, -0.07)	-0.177	0.057	(-0.29, -0.06)
ViroSeq	-0.095	0.029	(-0.15, -0.04)	-0.100	0.086	(-0.27, 0.07)	-0.121	0.045	(-0.21, -0.03)	-0.118	0.047	(-0.21, -0.03)
ANRS	-0.046	0.033	(-0.11, 0.02)	-0.083	0.090	(-0.26, 0.09)	-0.076	0.048	(-0.17, 0.02)	-0.075	0.048	(-0.17, 0.02)
2 ≤ rGSS < 3												
	G-computation			IPCW			DR-IPCW			TMLE		
	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	0.087	0.024	(0.04, 0.13)	0.066	0.044	(-0.02, 0.15)	0.067	0.024	(0.02, 0.11)	0.064	0.023	(0.02, 0.11)
Rega	0.038	0.019	(0.00, 0.07)	0.038	0.041	(-0.04, 0.12)	0.039	0.021	(0.00, 0.08)	0.039	0.021	(0.00, 0.08)
ViroSeq	0.071	0.021	(0.03, 0.11)	0.080	0.043	(0.00, 0.16)	0.062	0.022	(0.02, 0.11)	0.062	0.022	(0.02, 0.10)
ANRS	0.030	0.021	(-0.01, 0.07)	0.009	0.046	(-0.08, 0.10)	0.007	0.028	(-0.05, 0.06)	0.008	0.028	(-0.05, 0.06)
rGSS ≥ 3												
	G-computation			IPCW			DR-IPCW			TMLE		
	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI	Estimate	SE	95% CI
HIVdb	0.043	0.042	(-0.04, 0.13)	0.023	0.091	(-0.16, 0.20)	0.078	0.038	(0.00, 0.15)	0.103	0.035	(0.03, 0.17)
Rega	0.094	0.028	(0.04, 0.15)	0.055	0.051	(-0.05, 0.15)	0.091	0.027	(0.04, 0.14)	0.090	0.026	(0.04, 0.14)
ViroSeq	0.113	0.030	(0.05, 0.17)	0.072	0.065	(-0.05, 0.20)	0.116	0.027	(0.06, 0.17)	0.118	0.027	(0.07, 0.17)
ANRS	0.108	0.025	(0.06, 0.16)	0.078	0.051	(-0.02, 0.18)	0.111	0.023	(0.07, 0.16)	0.112	0.023	(0.07, 0.16)

Inference was obtained using the nonparametric bootstrap for G-computation, and using the influence curve for the other three estimators.

Table 3.8 – Comparison of influence curve (IC)-derived inference with nonparametric bootstrap-derived inference. IPCW estimator, comprehensive weighting.

rGSS < 1						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	-0.286	0.150	(-0.58, 0.01)	0.141	(-0.56, -0.01)	(-0.51, 0.03)
Rega	-0.331	0.157	(-0.64, -0.02)	0.125	(-0.58, -0.09)	(-0.55, -0.05)
ViroSeq	-0.256	0.188	(-0.62, 0.11)	0.125	(-0.5, -0.01)	(-0.53, -0.06)
ANRS	-0.377	0.102	(-0.58, -0.18)	0.129	(-0.63, -0.12)	(-0.54, -0.06)
1 ≤ rGSS < 2						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.017	0.056	(-0.09, 0.13)	0.040	(-0.06, 0.1)	(-0.06, 0.1)
Rega	-0.194	0.093	(-0.38, -0.01)	0.086	(-0.36, -0.02)	(-0.34, -0.02)
ViroSeq	-0.100	0.086	(-0.27, 0.07)	0.085	(-0.27, 0.07)	(-0.23, 0.1)
ANRS	-0.083	0.090	(-0.26, 0.09)	0.099	(-0.28, 0.11)	(-0.21, 0.17)
2 ≤ rGSS < 3						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.066	0.044	(-0.02, 0.15)	0.029	(0.01, 0.12)	(0.02, 0.14)
Rega	0.038	0.041	(-0.04, 0.12)	0.026	(-0.01, 0.09)	(-0.02, 0.09)
ViroSeq	0.080	0.043	(0.00, 0.16)	0.028	(0.03, 0.13)	(0.03, 0.14)
ANRS	0.009	0.046	(-0.08, 0.10)	0.040	(-0.07, 0.09)	(-0.05, 0.11)
rGSS ≥ 3						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.023	0.091	(-0.16, 0.20)	0.074	(-0.12, 0.17)	(-0.11, 0.17)
Rega	0.055	0.051	(-0.05, 0.15)	0.038	(-0.02, 0.13)	(0, 0.14)
ViroSeq	0.072	0.065	(-0.05, 0.20)	0.048	(-0.02, 0.17)	(0, 0.19)
ANRS	0.078	0.051	(-0.02, 0.18)	0.038	(0, 0.15)	(0.02, 0.17)

Table 3.9 – Comparison of influence curve (IC)-derived inference with nonparametric bootstrap-derived inference. DR-IPCW estimator, comprehensive weighting.

rGSS < 1						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	-0.232	0.095	(-0.42, -0.05)	0.099	(-0.43, -0.04)	(-0.38, 0.01)
Rega	-0.250	0.098	(-0.44, -0.06)	0.098	(-0.44, -0.06)	(-0.36, 0.01)
ViroSeq	-0.166	0.124	(-0.41, 0.08)	0.094	(-0.35, 0.02)	(-0.36, 0.01)
ANRS	-0.272	0.072	(-0.41, -0.13)	0.103	(-0.47, -0.07)	(-0.43, -0.02)
1 ≤ rGSS < 2						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.006	0.030	(-0.05, 0.06)	0.028	(-0.05, 0.06)	(-0.06, 0.05)
Rega	-0.180	0.056	(-0.29, -0.07)	0.065	(-0.31, -0.05)	(-0.33, -0.07)
ViroSeq	-0.121	0.045	(-0.21, -0.03)	0.050	(-0.22, -0.02)	(-0.23, -0.03)
ANRS	-0.076	0.048	(-0.17, 0.02)	0.053	(-0.18, 0.03)	(-0.19, 0.02)
2 ≤ rGSS < 3						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.067	0.024	(0.02, 0.11)	0.026	(0.02, 0.12)	(0.02, 0.13)
Rega	0.039	0.021	(0.00, 0.08)	0.021	(0.00, 0.08)	(0.00, 0.08)
ViroSeq	0.062	0.022	(0.02, 0.11)	0.024	(0.02, 0.11)	(0.02, 0.11)
ANRS	0.007	0.028	(-0.05, 0.06)	0.030	(-0.05, 0.06)	(-0.04, 0.08)
rGSS ≥ 3						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.078	0.038	(0.00, 0.15)	0.051	(-0.02, 0.18)	(-0.02, 0.18)
Rega	0.091	0.027	(0.04, 0.14)	0.031	(0.03, 0.15)	(0.03, 0.15)
ViroSeq	0.116	0.027	(0.06, 0.17)	0.033	(0.05, 0.18)	(0.04, 0.17)
ANRS	0.111	0.023	(0.07, 0.16)	0.028	(0.06, 0.17)	(0.06, 0.16)

Table 3.10 – Comparison of influence curve (IC)-derived inference with nonparametric bootstrap-derived inference. TMLE, comprehensive weighting.

rGSS < 1						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	-0.239	0.091	(-0.42, -0.06)	0.098	(-0.43, -0.05)	(-0.41, -0.02)
Rega	-0.239	0.101	(-0.44, -0.04)	0.117	(-0.47, -0.01)	(-0.49, -0.01)
ViroSeq	-0.187	0.110	(-0.40, 0.03)	0.106	(-0.39, 0.02)	(-0.46, -0.02)
ANRS	-0.233	0.064	(-0.36, -0.11)	0.092	(-0.41, -0.05)	(-0.45, -0.09)
1 ≤ rGSS < 2						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.006	0.030	(-0.05, 0.07)	0.026	(-0.04, 0.06)	(-0.05, 0.05)
Rega	-0.177	0.057	(-0.29, -0.06)	0.060	(-0.29, -0.06)	(-0.32, -0.08)
ViroSeq	-0.118	0.047	(-0.21, -0.03)	0.044	(-0.2, -0.03)	(-0.21, -0.04)
ANRS	-0.075	0.048	(-0.17, 0.02)	0.049	(-0.17, 0.02)	(-0.17, 0.01)
2 ≤ rGSS < 3						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.064	0.023	(0.02, 0.11)	0.026	(0.01, 0.12)	(0.02, 0.12)
Rega	0.039	0.021	(0.00, 0.08)	0.021	(0, 0.08)	(0, 0.08)
ViroSeq	0.062	0.022	(0.02, 0.10)	0.023	(0.02, 0.11)	(0.02, 0.11)
ANRS	0.008	0.028	(-0.05, 0.06)	0.028	(-0.05, 0.06)	(-0.04, 0.07)
rGSS ≥ 3						
		Influence Curve		Bootstrap		
	Estimate	SE	95% CI	SE	95% CI (Wald-type)	95% CI (quantiles)
HIVdb	0.103	0.035	(0.03, 0.17)	0.054	(0, 0.21)	(-0.02, 0.19)
Rega	0.090	0.026	(0.04, 0.14)	0.033	(0.02, 0.16)	(0.02, 0.16)
ViroSeq	0.118	0.027	(0.07, 0.17)	0.038	(0.04, 0.19)	(0.04, 0.18)
ANRS	0.112	0.023	(0.07, 0.16)	0.031	(0.05, 0.17)	(0.05, 0.17)

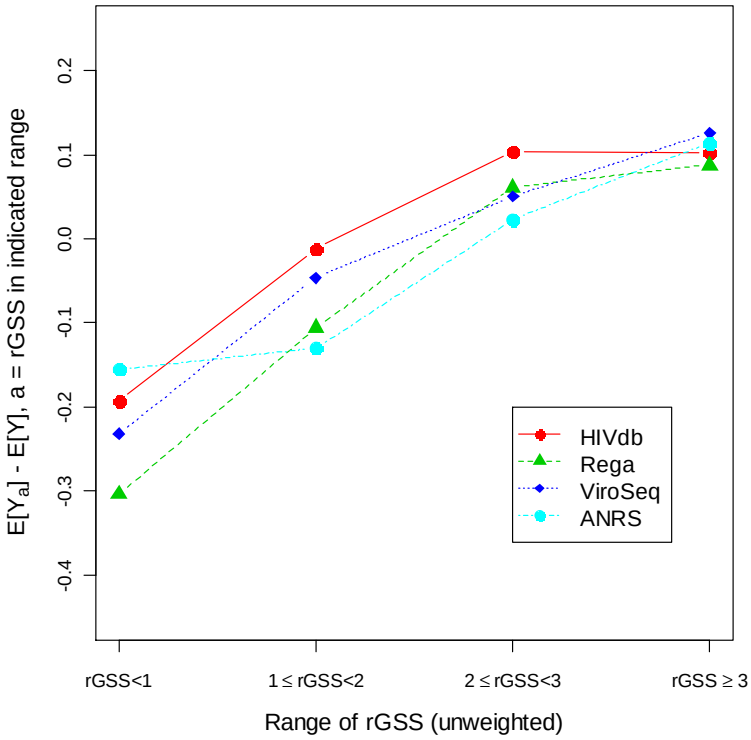

Figure 3.2 TMLE estimates of the difference in the probability of virologic suppression for salvage regimens with unweighted rGSS in the indicated range versus observed values. A larger rGSS should indicate a more effective treatment regimen.

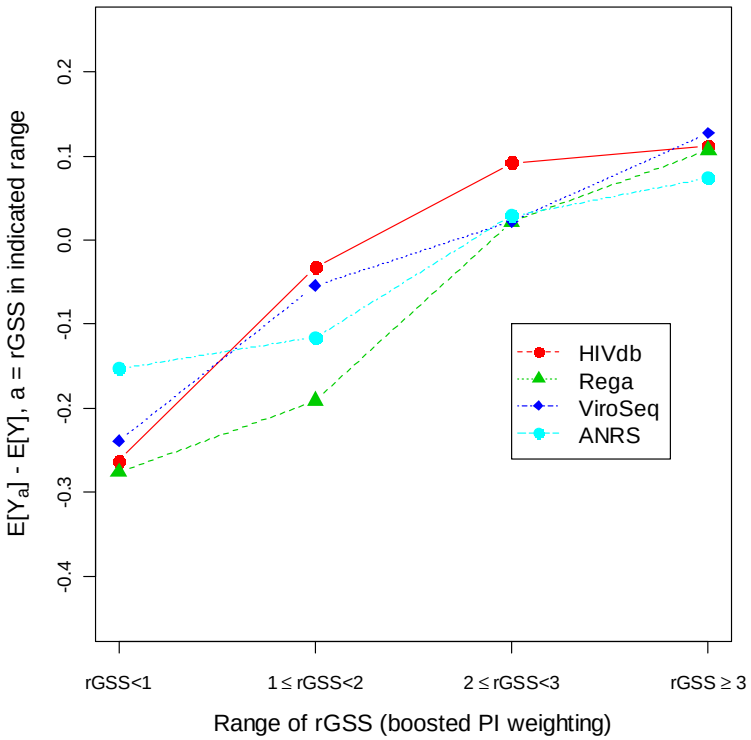

Figure 3.1 TMLE estimates of the difference in the probability of virologic suppression for salvage regimens with boosted PI-weighted rGSS in the indicated range versus observed values.

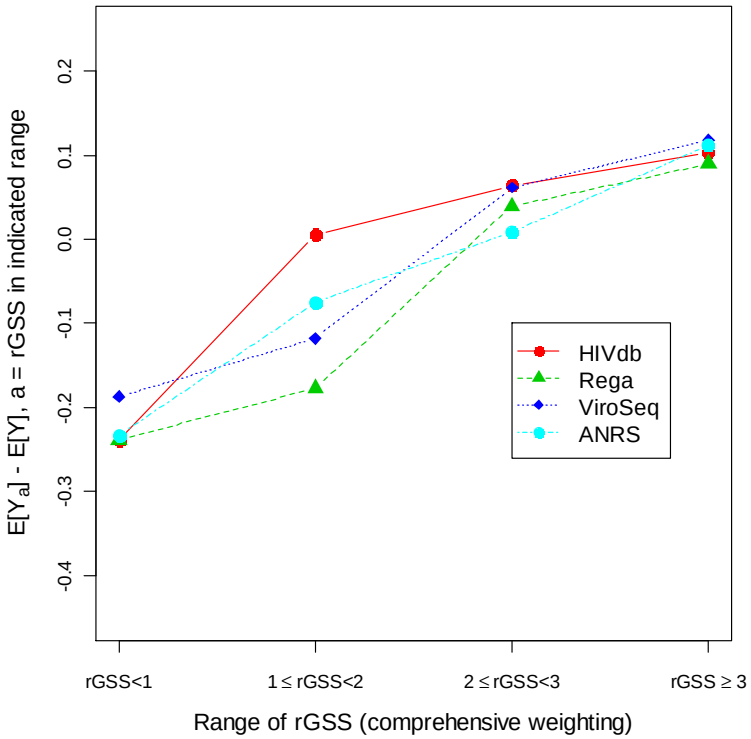

Figure 3.3 TMLE estimates of the difference in the probability of virologic suppression for salvage regimens with comprehensively weighted rGSS in the indicated range versus observed values.

Table 3.11 – Estimates of bias induced by ETA violations for ten target variables.

	IPCW			DR-IPCW			TMLE		
	Estimate	SE	$Bias_{ETA}$	Estimate	SE	$Bias_{ETA}$	Estimate	SE	$Bias_{ETA}$
Boosted PI weighting, A=0 when rGSS < 1									
HIVdb	-0.400	0.143	-0.054	-0.295	0.083	-0.006	-0.262	0.062	0.016
ViroSeq	-0.405	0.143	-0.051	-0.266	0.076	0.003	-0.239	0.061	0.010
ANRS	-0.314	0.179	-0.070	-0.193	0.129	-0.016	-0.153	0.089	-0.011
Comprehensive weighting, A=0 when rGSS < 1									
HIVdb	-0.286	0.150	-0.026	-0.232	0.095	0.003	-0.239	0.091	0.011
ViroSeq	-0.256	0.188	-0.041	-0.166	0.124	-0.007	-0.187	0.110	-0.003
Unweighted, A=0 when rGSS < 1									
HIVdb	-0.360	0.138	-0.087	-0.224	0.074	-0.009	-0.192	0.058	0.002
Rega	-0.432	0.111	-0.036	-0.275	0.060	0.001	-0.304	0.066	0.002
ViroSeq	-0.405	0.143	-0.047	-0.262	0.076	-0.002	-0.232	0.060	0.006
ANRS	-0.314	0.179	-0.061	-0.195	0.128	-0.008	-0.155	0.088	<0.001
Unweighted, A=0 when 2 ≤ rGSS < 3									
ANRS	0.013	0.044	0.001	0.021	0.023	0.002	0.023	0.023	0.002

3.3.2 Prediction

Figures 3.4, 3.5, and 3.6 show cross-validated estimated ROC curves for the full super learner prediction models incorporating the unweighted, boosted PI-weighted, and comprehensively weighted rGSS, respectively. ROC curves for each of the four genotypic resistance test interpretation algorithms are shown separately, and compared with the ROC curve for the super learner prediction model including all other explanatory variables but excluding any rGSS variable. The ROC curves for the different genotypic resistance test interpretation algorithms appear nearly indistinguishable, and their performance in terms of the area under the ROC curve (AUC) is also very similar, ranging from 0.77 to 0.80. The best performer in terms of AUC is HIVdb, with AUC = 0.80 for the boosted PI and comprehensive weighting schemes, and AUC = 0.79 for the unweighted rGSS. The worst performer is ANRS, with AUC = 0.77 for the unweighted rGSS and AUC = 0.78 for the other two weighting schemes. The prediction models including an rGSS variable appear to perform very slightly better than the model containing no rGSS variable (AUC = 0.76).

Figure 3.7 compares weighting schemes, with ROC curves averaged across genotypic resistance test interpretation algorithms. The choice of rGSS weighting scheme does not appear to make much of a difference in prediction model performance; in terms of AUC, the model allowing the super learner to weight the drug class-specific GSS totals ("dynamic weighting") performs better than the other weighting schemes (AUC = 0.79), but since the worst performer (unweighted) has an AUC = 0.77, the improvement is barely detectible. The curves appear almost interchangeable.

Figure 3.8 compares weighting schemes again, this time across logistic regression models with the rGSS as the sole predictor. The ROC curves for the full prediction model including the comprehensively weighted rGSS and the prediction model with no rGSS are also shown for comparison. The ROC curves for the weighting schemes are noticeably different in this comparison, with the unweighted rGSS performing worst (AUC = 0.66), and the comprehensive and dynamic weighting performing best with AUC = 0.73 and AUC = 0.74, respectively. Interestingly, the performance of the rGSS-only models for the comprehensive and dynamic weighting schemes appears only marginally different from the model with no rGSS, which includes 31 explanatory variables (AUC = 0.76).

The simplistic "scaled rGSS" approaches, in which the rGSS values were scaled to fall between 0 and 1 and then treated as predicted probabilities, were also compared to the more sophisticated modeling approaches (ROC curves not shown). The AUC for the unweighted rGSS-only logistic regression model was no better than the AUC for the scaled unweighted rGSS (AUC = 0.67). The scaled comprehensively weighted rGSS performed better in terms of AUC (AUC = 0.71) than both of the rGSS-only logistic regression models for the unweighted and boosted PI weighting schemes

The ROC curves in figure 3.9 compare the best performers of each model type. The full prediction models with the dynamically and comprehensively weighted rGSS perform best, the scaled comprehensively weighted rGSS performs worst, and the prediction models with no rGSS and with the comprehensively weighted rGSS as the sole predictor fall in between. Overall, the AUC values range between 0.71 and 0.79.

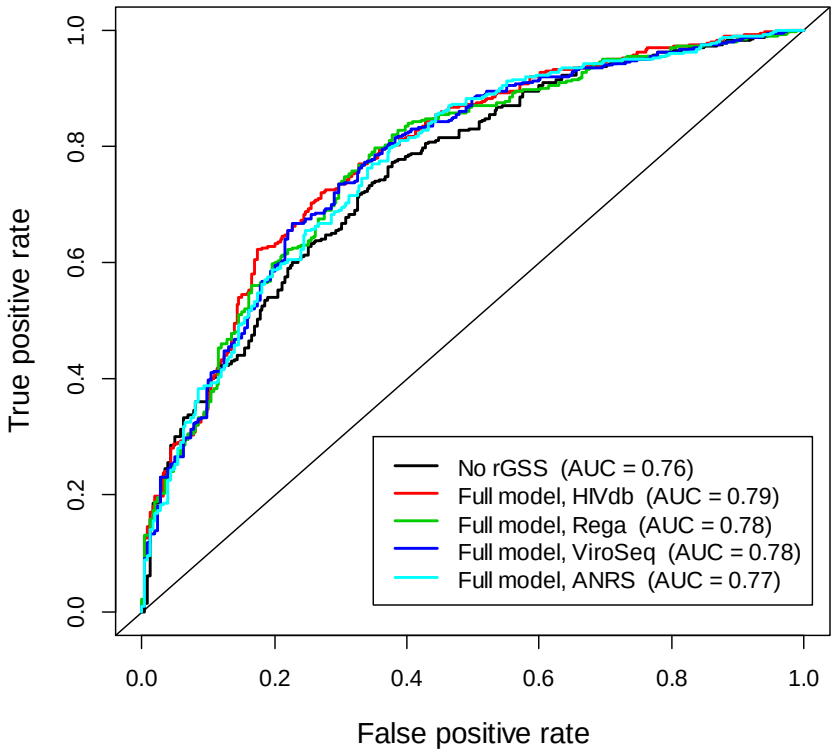

Figure 3.4 ROC curves for full prediction models including the unweighted rGSS across different genotypic resistance test interpretation algorithms.

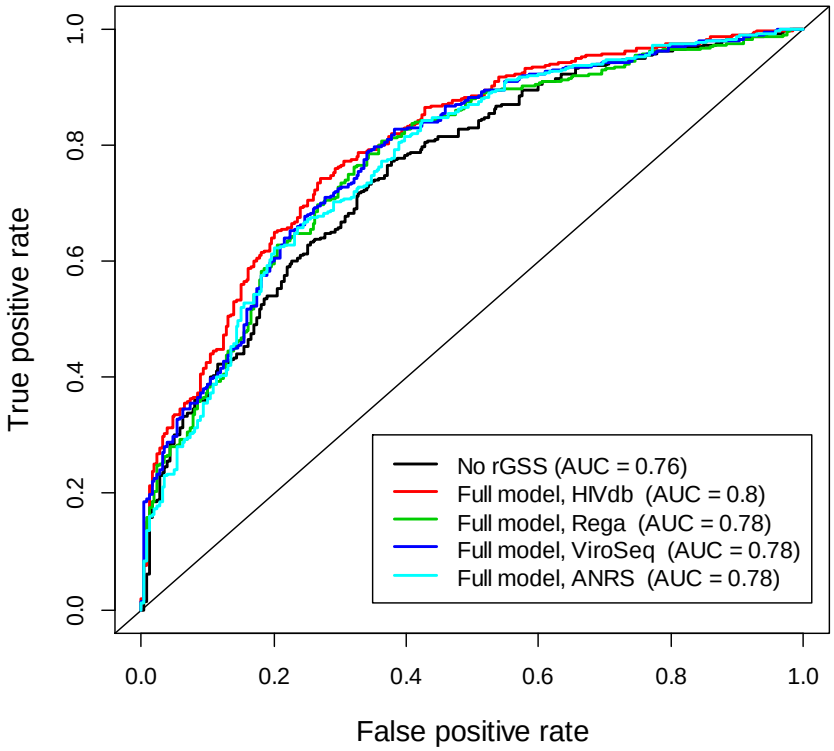

Figure 3.5 ROC curves for full prediction models including the boosted PI-weighted rGSS across different genotypic resistance test interpretation algorithms.

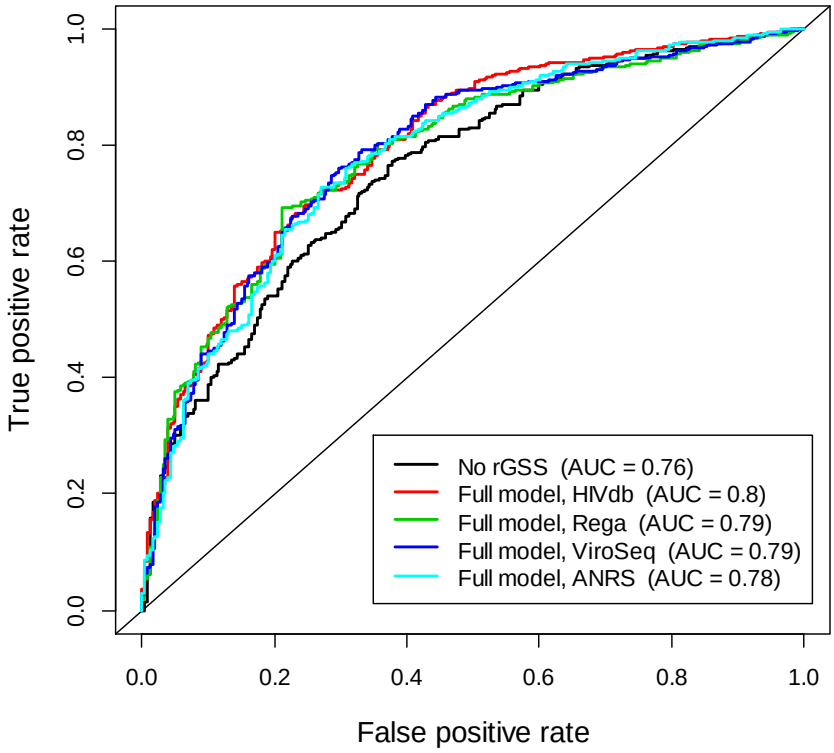

Figure 3.6 ROC curves for full prediction models including the comprehensively weighted rGSS across different genotypic resistance test interpretation algorithms.

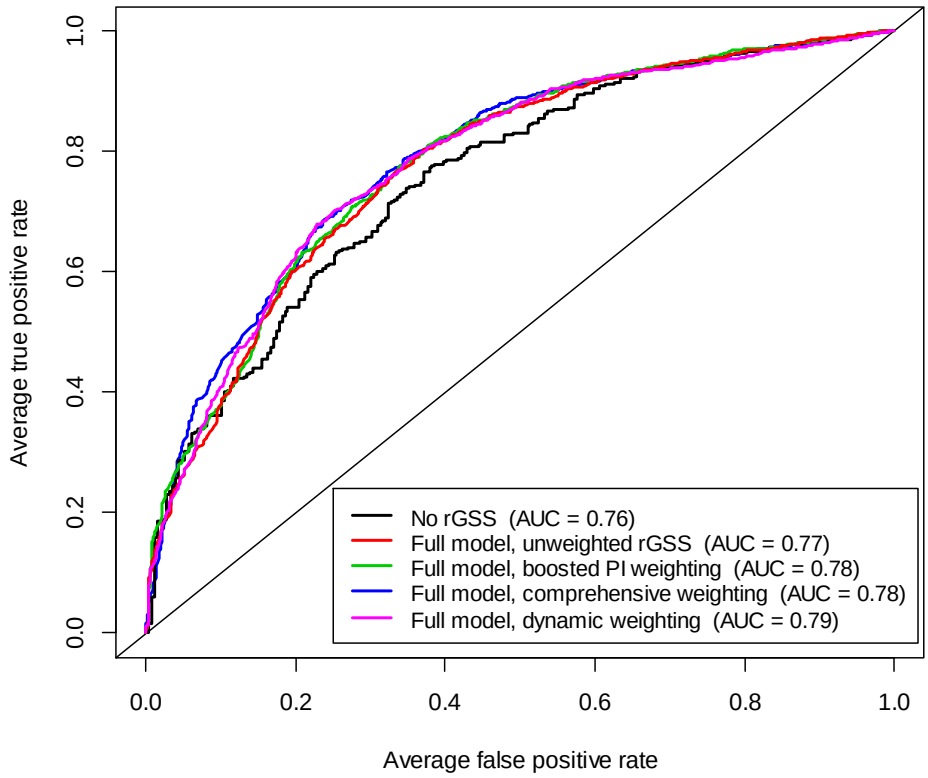

Figure 3.7 ROC curves for each rGSS weighting scheme, averaged across the four genotypic resistance test interpretation algorithms.

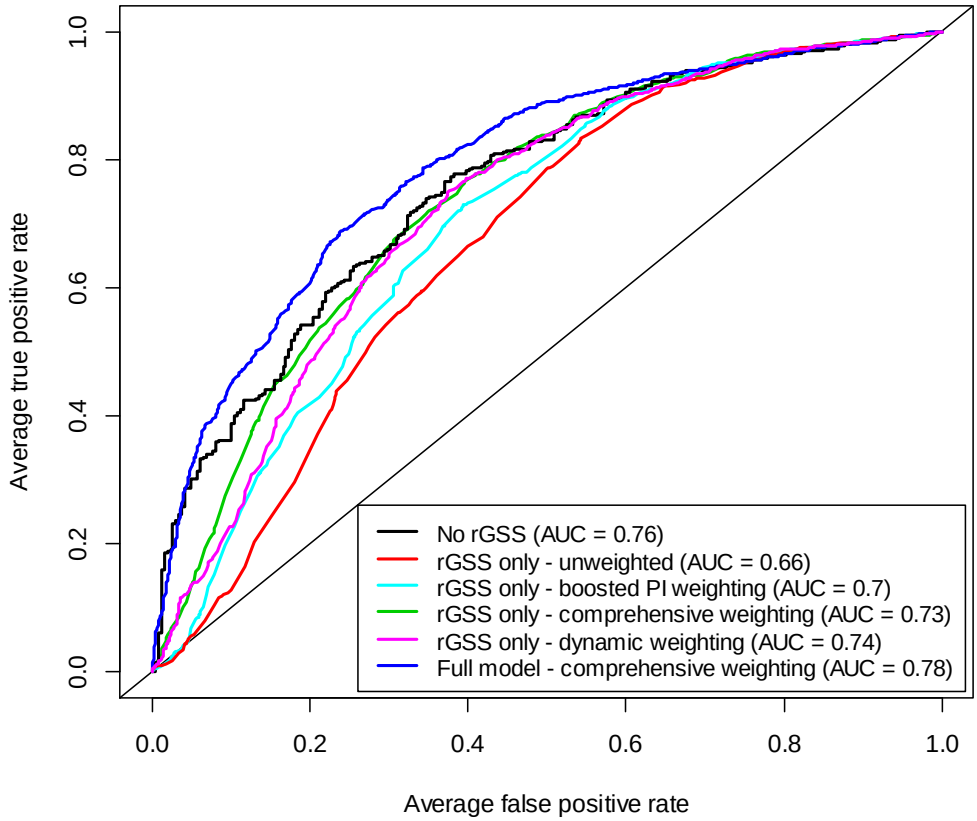

Figure 3.8 ROC curves comparing prediction models using rGSS as the sole predictor, separated by rGSS weighting scheme and averaged across genotypic resistance test interpretation algorithms.

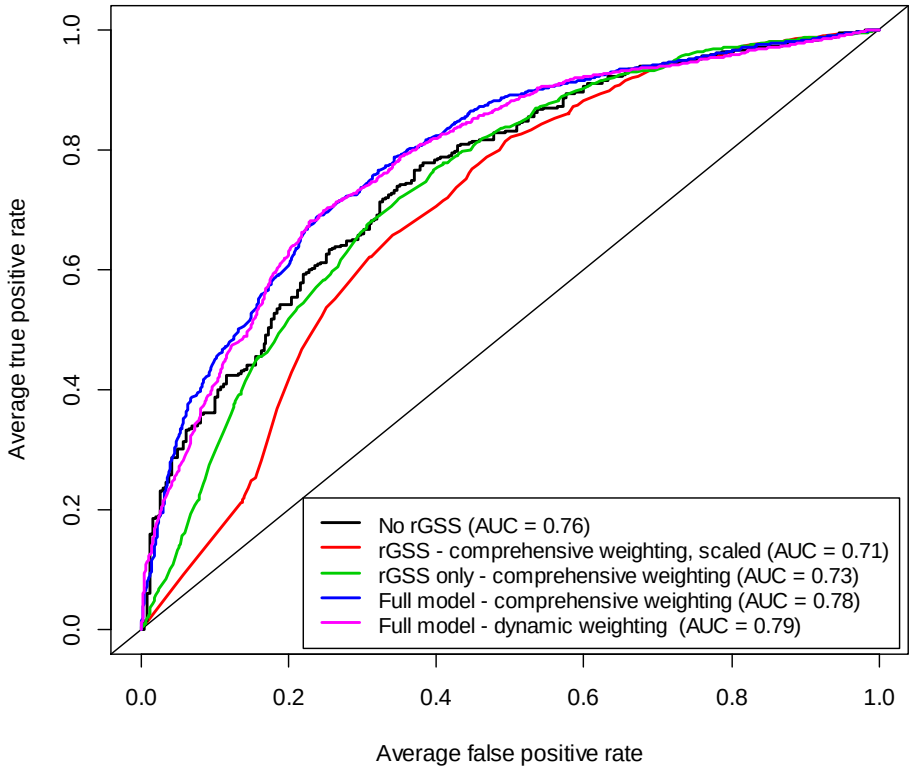

Figure 3.9 ROC curves for the best performers of each model type.

3.4 Discussion

In the context of this analysis, there did not appear to be much difference between the different genotypic resistance test interpretation algorithms, either in terms of variable importance or prediction. Different algorithms did yield differing parameter estimates for the different dichotomizations, but mostly these differences were small in comparison with the estimated standard errors. The rGSS for all algorithms was found to be associated with the virologic outcome, even after adjusting for the many other explanatory variables. The rGSS also moderately improved predictive power when added to a prediction model including the other explanatory variables.

There appears to be a fair amount of shared information between the rGSS and the other explanatory variables, manifesting itself in both ETA violations and in the predictive value of the rGSS when used as the sole predictor of the virologic outcome. While the addition of the rGSS to a prediction model including the other explanatory variables resulted in a moderate gain in predictive power, the rGSS alone performed almost as well as the prediction model including all 31 other explanatory variables but no rGSS. While a full prediction model would be preferable, this does seem to imply that the rGSS alone could still be useful in identifying patients at high risk for virologic failure after a treatment change. The rGSS weighting scheme only seemed to matter when the rGSS was used alone in predicting virologic suppression – in that case, the comprehensive weighting scheme was the best choice for prediction in this dataset. There did not seem to be an appreciable gain in predictive value in allowing data-adaptive weighting of the drug class-specific GSS.

Practical violations of the ETA or positivity assumption were extreme in this analysis; out of the 48 estimates calculated, 16 had minimum values for $g_n(0|W)$ less than 0.005, 31 had minimum values less than 0.025, and 39 had minimum values less than 0.05. The cost of these violations in positivity seems to have been increased estimator variance rather than bias. This is suggested by both the large estimated standard errors and by the results of the parametric bootstrap diagnostic, which did not raise any red flags in terms of estimator bias due to ETA violations. Increased bounding on $g_n(0|W)$ would likely have improved variance, but could also have induced more bias, particularly for the IPCW estimator.

In this analysis, the IPCW estimator performed worst in terms of variance and estimated ETA violation-induced bias, with standard error estimates consistently close to twice as large as the next largest standard error for any other estimator. The estimated ETA bias for IPCW, though not large enough to raise a red flag according to the bias diagnostic, was in all instances still many times larger than the estimated ETA bias for either DR-IPCW or TMLE. Both double-robust estimators performed similarly in terms of variance in most cases, but it should be noted that TMLE was always more efficient in the seven cases in which the largest difference in standard error between DR-IPCW and TMLE was observed. Examination of the estimated influence curves in these seven cases revealed that the estimated DR-IPCW influence curve always contained more extreme positive values than its TMLE counterpart. Since these seven cases coincided with the most extreme practical violations of the ETA assumption, it appears that

TMLE may have an advantage over DR-IPCW in terms of efficiency when $g_n(0|W)$ nears zero for some observations.

The G-computation standard errors for the comprehensively weighted rGSS dichotomizations were in many cases smaller than those of the double-robust estimators, most likely due to the fact that G-computation estimates will be least affected by ETA violations. The differences are rarely large, however, and the standard error estimates for DR-IPCW and TMLE were almost always comparable to those of G-computation, and in some cases were smaller. The lack of any appreciable efficiency gain, combined with the fact that G-computation relies completely upon correct specification of the model for $E[Y \mid A=0,W]$, suggests that a double-robust estimator would be a better choice. Standard errors for G-computation also must be estimated using the nonparametric bootstrap, which can become cumbersome when the number of estimated parameters is large.

Both G-computation and TMLE are substitution estimators, which has the benefit that they respect the bounds on the parameter. In this analysis, the true parameter cannot exceed in absolute value the observed rate of virologic suppression (64.7%). IPCW and DR-IPCW are not substitution estimators, and as such theoretically can exceed the bounds on the parameter. This is seen numerous times for the IPCW 95% confidence intervals – in many cases, the lower bound exceeds in absolute value the observed rate of virologic suppression. For the parameter to attain this value, it would have to be possible for the rate of virologic suppression to fall below zero, which is impossible. The DR-IPCW estimates in this analysis do not encounter this problem – however, the theoretical possibility is still there. While noticeable differences in performance between the two double-robust estimators were not observed in this analysis, this theoretical difference could make TMLE preferable – particularly when parameter estimates or confidence interval limits could reasonably be expected to come up against the bounds on the parameter. TMLE also is the only estimator option designed to reduce bias with respect to the desired feature of the data-generating distribution (the parameter of interest).

It is important in any analysis that involves estimates of variable importance in relation to an outcome of interest to choose a parameter definition that has meaning in the real world, and not only in the context of an arbitrarily pre-specified model. The statistical parameter (the parameter that is identifiable under the observed data distribution P_0) should also have subject-matter value, so the results do not depend completely upon the validity of the usually untestable causal assumptions. The estimator used for estimating the parameter of interest (such as IPCW, DR-IPCW, G-computation, or TMLE) does matter, particularly when there are ETA violations. Of the estimators considered in this analysis, DR-IPCW and TMLE appeared to be the best choices, with TMLE having a slight theoretical edge. Other estimators are available, such as collaborative TMLE, which can have smaller bias and variance when ETA violations are present than any of the estimators used here, due to its adaptive process for selecting the best covariates to include in $g_n(A|W)$ (van der Laan & Gruber, 2009). Good theoretical properties are important, because while many estimators may perform equally under ideal conditions (correct model specification, no ETA violations, etc.), such conditions are rarely in evidence in real applications – therefore, it is worthwhile to consider the behavior of the estimator of choice when conditions are not ideal. When ETA violations are present, this could involve employing the parametric bootstrap to

estimate the degree of ETA violation-induced estimator bias, and possibly using this diagnostic to select a bound on $g_n(A|W)$ that provides an acceptable tradeoff between variance and bias.

Use of asymptotic results to estimate standard errors (i.e. using the estimated influence curve) is convenient, particularly when a large number of estimates must be calculated. The nonparametric bootstrap is also available, however, when sample sizes are small or when there are other concerns regarding the behavior of an estimator under the conditions of the application of interest. In this analysis, the influence curve-based standard errors and inference for DR-IPCW and TMLE were found to be comparable to the bootstrap-derived standard errors and inference. Influence curve-based standard errors for IPCW were found to be conservative with respect to the boostrap-derived standard errors. Overall, the comparison of bootstrap-based with influence curve-based results supported the use of asymptotic standard error calculations in this analysis.

Estimation procedures and model selection, whether in calculating variable importance estimates or forming prediction models, should respect what is known about the form of the data-generating distribution, which is usually nothing. Machine learning techniques and techniques that utilize multiple candidate models (such as super learning) for model selection are therefore particularly valuable – they can cast a wide net and do not require any manual intervention from the researcher, other than in choosing the initial set of candidate models. This makes it possible to employ inference in variable importance estimation that includes model choice as part of its estimate of variability – this is impossible when the choice of model is determined via manual researcher intervention. In the prediction context, it can be of value to assess models ranging in complexity, because an increase in model complexity may not necessarily result in a commensurate increase in predictive value.

Chapter 4

An Assessment of Factors Contributing to Hospital Readmission Risk and Evaluation of a Telemanagement Intervention for Heart Failure Patients

4.1 Background

Prevention of unnecessary readmissions to the hospital has been identified as an area of opportunity to improve quality and reduce costs of healthcare delivery, particularly in the hospital setting (Averill, McCullough, Hughes, Goldfield, Vertrees, & Fuller, 2009; Jencks, Williams, & Coleman, 2009). In 2008, the Medicare Payment Advisory Commission (MedPAC) recommended to Congress that high readmission rates for select conditions be used as a basis for reduced Medicare payments to hospitals; the Affordable Care Act, signed into law in March 2010, called for the establishment of programs for hospitals with high severity-adjusted readmission rates to reduce these rates through quality improvement (Agency for Healthcare Research and Quality; Medicare Payment Advisory Commission, 2008). In October 2012, Medicare will begin to penalize hospitals by reducing fee-for-service payments if their readmission rates for heart failure, heart attack, or pneumonia are higher than expected (Andrews, 2011). Despite the increased urgency to improve quality and to implement processes that will positively impact rates of readmission, however, hospital readmission rates have so far proven difficult to impact, and the best approach by which a desired rate reduction can be achieved has yet to be identified (Rau, 2012).

Heart failure patients have long represented a large fraction of Medicare beneficiaries, and have been identified as one of the populations receiving particular focus from the new Medicare regulations (Krumholz, et al., 1997; Ross, et al., 2008; Andrews, 2011). In 2008, a heart failure program was implemented at two hospitals in Alameda County, California, with the aim of increasing the time to readmission, and thereby reducing short-term readmission rates, for patients initially hospitalized for heart failure. The program was community-level, in that the entire heart failure population at both hospitals was simultaneously targeted for intervention without any randomization or separation into intervention and control groups. The intervention consisted of two main components: (1) a hospital-based intervention, during which patients identified as hospitalized primarily for heart failure were visited by specially trained nurses, and provided education and information pertaining to their disease and specifically to self-management of their heart failure symptoms post-hospitalization; (2) a telephone management intervention, during which patients identified at their hospitalization as being high risk for readmission after discharge (and appropriate for a telephone management program) were followed telephonically by specialty nurses, who provided additional support for symptom and medication management, as well as assistance with coordination of outpatient care. The specific focus of this program was readmission for heart failure, because this is the largest subset amongst the all cause readmission diagnoses after an initial heart failure hospitalization; heart failure readmissions are also the most straightforward to target, because the causes of readmission are easier to identify.

The telephone management portion of the heart failure intervention required identification of patients at high risk for readmission. While many studies have assessed the relationship between various clinical and demographic variables with the risk of readmission after a heart failure hospitalization, only a handful have concerned themselves with the development of readmission risk models or tools that could be used to identify high-risk patients (Ross, et al., 2008). Of these, only the risk tool of Philbin and DiSalvo (1999) specifically targeted readmission for heart failure, as opposed to readmission for any reason (Philbin & DiSalvo, 1999; Ross, et al., 2008;

Wang, et al., 2012). The heart failure program employed a modified version of this risk tool, which was deployed as a paper checklist to be filled in manually by evaluating clinicians. Though the modified risk tool was developed using input from clinicians with many years of experience with heart failure patients, it was never tested or validated with sample data before being put to use.

This analysis considers 30-, 90-, and 180-day readmission outcomes after an initial hospitalization for heart failure, based on two years' worth of retrospective administrative data from the two Alameda County, California hospitals at which the heart failure program was implemented. This encompasses data from the first year of the program and from one pre-intervention year. Both all cause (readmission for any reason) and heart failure readmission outcomes are considered. Three main goals are targeted: (1) the evaluation of the heart failure program's impact on readmission; (2) an investigation risk factors for readmission; and (3) a consideration of the readmission risk score's predictive value. The impact of the heart failure program on readmission outcomes was evaluated using causal inference-inspired semi-parametric variable importance measures. Unlike parameters of arbitrary regression models, these measures are not intrinsically linked to any particular model. This means that the parameter of interest can be interpreted in a real-world context, regardless of the method used to estimate the data-generating distribution. Similar variable importance methods were also applied to evaluate the association of various risk factors with the readmission outcomes. Prediction of readmission outcomes using the heart failure program's readmission risk score was compared to prediction of readmission using multiple explanatory variables. Prediction models utilizing super learning, which can incorporate multiple models without requiring manual model selection by the researcher, were also compared to simple main terms logistic regression models in terms of predictive accuracy.

4.2 Methods

4.2.1 Data

This analysis was conducted on a retrospective dataset of administrative data from two hospitals in Alameda County, California. The dataset consisted of all inpatient hospitalizations for heart failure with discharge dates between August 1, 2006 and July 31, 2007 (control group) and August 1, 2008 to July 31, 2009 (intervention group). Consecutive years could not be considered due to partial implementation of the heart failure intervention in November 2007. Full implementation of the heart failure intervention was achieved in both hospitals on July 12, 2008.

Heart failure was defined by the coded primary diagnosis, according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). See Table A4.1 in the Appendix for the full code list.

Heart failure inpatient hospitalizations were excluded from the dataset if their hospital discharge dispositions indicated that the patient expired, left the hospital against medical advice, or was transferred to another facility (hospital, skilled nursing facility, or long term care) after hospitalization. Also excluded were hospitalizations at which dialysis was received (according to

the hospital admission's coded procedures). These hospitalizations were excluded because they were not the intended target population of the heart failure intervention. The final dataset consisted of 788 hospitalizations for 617 unique patients in the control group, and 588 hospitalizations for 471 unique patients in the intervention group. The combined dataset contained 1376 hospitalizations for 1034 unique patients.

The outcome of interest was readmission to the hospital within a specified period of time. Six binary outcomes were considered: (1) readmission for heart failure within 30 days, (2) readmission for any reason within 30 days, (3) readmission for heart failure within 90 days, (4) readmission for any reason within 90 days, (5) readmission for heart failure within 180 days, and (6) readmission for any reason within 180 days. In order to qualify as a readmission, a given subsequent hospital admission was required to be a non-elective acute inpatient hospitalization at either one of the two hospitals considered. Readmission for heart failure was determined according to the primary diagnosis ICD-9-CM code, again according to the list in Table A4.1. Days to readmission was defined as [Readmission admit date] – [Prior hospitalization discharge date].

The list of potential explanatory variables was motivated primarily by the readmission risk prediction score of Philbin and DiSalvo (1999), because a modified version of this score was used by the heart failure program to identify patients at high risk for readmission (Philbin & DiSalvo, 1999). Some additional explanatory variables thought to be possibly related to enrollment in the heart failure program and to the probability of readmission were also considered. The list of baseline covariates includes patient demographic variables (African American race, age at hospitalization), features of the patient's treatment history (cardiac surgery in the past year, inpatient hospitalization in the past year), current disease status (type of heart failure, presence of certain comorbidities, number of diagnoses), and features of the hospitalization (whether the admit or discharge date occurred on the weekend, hospital length of stay, whether patient was in a telemetry unit during the hospitalization, whether the patient was Medicare or Medicaid insured, whether the patient was discharged to home health). The full list of explanatory variables is shown in Table 4.1. The variables shaded in gray are the components of the readmission risk score used by the heart failure program; the score is constructed by simply adding the individual binary variables together.

For prior cardiac surgery and inpatient hospitalization in the past year, only hospitalizations at the two hospitals examined in this study were considered. The list of ICD-9-CM procedure codes that qualified as cardiac surgery can be found in Table A4.1 in the Appendix, as are the lists of ICD-9-CM diagnosis codes that defined the presence of each disease state listed in Table 4.1. For these additional disease states, all coded diagnoses were considered, not only the primary diagnosis. Medicare or Medicaid insurance was determined according to the payer listed for the hospitalization, and a patient was considered to have received telemetry if he or she was documented as having been housed in a known telemetry unit of the hospital at any point during hospitalization. Discharge to home health service was determined according to the hospitalization discharge disposition.

Because of the very small number of patients who had received cardiac surgery in the prior year, this explanatory variable was excluded from individual analysis and included only in the readmission risk score.

Table 4.1 - Explanatory variables. Components of the readmission risk score are shaded in gray.

Variable Category	Variable Description
Demographic	African American race
	Age at hospital admission
Treatment history	Cardiac surgery in past year
	Inpatient hospitalization in past year
Current disease status	Chronic lung disease
	Diabetes mellitus
	Ischemic heart disease
	Renal disease
	Idiopathic cardiomyopathy
	Valvular heart disease
	Number of diagnoses
Hospitalization	Discharged to home health
	Hospital length of stay (days)
	Medicaid
	Medicare
	Telemetry during hospitalization
	Weekend hospital admission
	Weekend hospital discharge
	Hospital

4.2.2 Variable Importance

Data Structure

The heart failure program can be thought of as a community-level intervention on two "communities:" heart failure hospitalizations that occurred between August 1, 2006 and July 31, 2007, and heart failure hospitalizations that occurred between August 1, 2007 and July 31, 2008. The observed data, a collection of hospital admissions, can be thought of as a random sample of i.i.d. observations of the random variable $O = (E, W, A, Y)$. This random variable follows some unknown distribution P_O, which is itself a component of $\mathcal{M}$, a set of possible probability distributions. The individual hospitalizations $O_1, O_2, \ldots, O_n$ can therefore be defined as

$$O_i = (E_i, W_i, A_i, Y_i),\ i \in \{1, 2, ..., n\}.$$

The elements (E, W, A, Y) that comprise O are as follows: E represents the community-level variables, A represents the binary treatment or target variable, W represents the set of possible individual-level confounders, and Y represents the binary outcome variable. In this application, E

is the time period in which the hospitalization occurred (August 1, 2006 - July 31, 2007 or August 1, 2008 - July 31, 2009), *A* is the presence or absence of the heart failure intervention, *W* is the set of other possible individual-level explanatory variables outlined in Table 4.1, and *Y* is one of the six readmission outcomes described in the previous section.

Model and Target Parameters: Heart Failure Intervention

In the context of the observed data, the community-level nature of the heart failure program means that the environmental variable *E* is completely confounded with the treatment variable *A*, and only one outcome can be observed for each hospitalization – if the hospitalization is part of the control group, it is impossible to observe the outcome *Y* that would have occurred if the hospitalization were part of the intervention group, and vice versa. One could, however, conceive of a hypothetical full data structure *X* under which it would be possible to observe any combination of (E,W,A) and the resulting outcome; the observed data can then be considered a missing data structure on this hypothetical full data. Following the notation in van der Laan (2010), a nonparametric structural equation model (NPSEM) for endogenous $X=(E,W,A,Y)$ can be constructed as follows, assuming exogenous $U=(U_E,U_W,U_A,U_Y)\sim P_U$: (Pearl, 2000; van der Laan M. J., 2010)

$$E = f_E(U_E)$$
$$W = f_W(E,U_W)$$
$$A = f_A(E,W,U_A)$$
$$Y = f_Y(E,W,A,U_Y)$$

Let $E\in\{e_0,e_1\}$, where e_0 is the pre-intervention time period, and e_1 is the intervention time period. $A\in\{0,1\}$ is the absence or presence of the heart failure program. Assume also that $\alpha=P(E=e_1)$ is known.

We can now define the counterfactuals Y_0 and Y_1 on this NPSEM, which are the random variables obtained by setting *A=0* and *A=1*, respectively. We can also define the observed data *O* as i.i.d. observations from the post-intervention counterfactual distribution of the intervention $A=0,E=e_0$ and $A=1,E=e_1$. Specifically, for $n=n_0+n_1$, we observe n_0 observations on the counterfactual $(W(e_0),Y(e_0,0))\sim P_{e_0,0}$ and n_1 observations on the counterfactual $(W(e_1),Y(e_1,1))\sim P_{e_1,1}$. In this analysis, $n_1=588$ and $n_0=788$.

Let us define the random variable $B=Bernoulli(\alpha)\in\{(0,e_0),(1,e_1)\}$. We can now re-define the observed data *O* as $O=(B,W(B),Y(B))$. Conditional on $B=(0,e_0)$, *O* follows $P_{e_0,0}$, and conditional on $B=(1,e_1)$, *O* follows $P_{e_1,1}$.

The first parameter of interest is the additive causal effect, which is defined as follows:

$$\psi\left(P_{U,X}\right) = E\left[Y_1\right] - E\left[Y_0\right].$$

$P_{U,X}$ denotes the probability distribution of (U, X). This is the difference in the probability of hospital readmission in an ideal experiment where control and intervention groups could be randomly sampled from the full data, with *W* and *E* approximately evenly distributed across groups, meaning that the parameter would not be affected by environmental confounding.

The statistical parameter, or the analogous parameter of the observed data distribution, is as follows:

$$\psi\left(P_O\right) = E_{W(B)}\left[E\left(Y(B) \mid W(B), B = (1, e_1)\right) - E\left(Y(B) \mid W(B), B = (0, e_0)\right)\right].$$

$E_{W(B)}$ above indicates that the mean of the difference in expected outcomes is also taken over all individual-level $W(B)$. Assumptions are required in order for $\psi\left(P_O\right)$ to be equivalent to $\psi\left(P_{U,X}\right)$. We must assume the NPSEM defined above, and that the marginal distribution of *E* is known. We must also assume that *E* only affects *Y* through *W*, which has been referred to as the exclusion assumption, or no residual environmental confounding. It is also necessary that there be a positive probability of inclusion in either the intervention or control group given the covariate values present in the sample, or:

$$0 < \mathrm{P}\left(B = (1, e_1) \mid W\right) < 1 \text{ a.e.}$$

Finally, we require the strong randomization assumption, which says that P_U is such that (E, W, A) is independent of $Y(e, w, a)$ for all $e, w, a.$ (van der Laan M. J., 2010).

The second parameter of interest is the additive causal effect among the treated population, which can be thought of as the mean difference in outcome amongst the intervention group if the intervention group had not received the intervention. Recall that the outcome of interest is hospital readmission after initial hospitalization for heart failure. In the context of the same NPSEM defined above, the causal parameter is defined as follows:

$$\psi^t\left(P_{U,X}\right) = E\left[Y_1 - Y_0 \mid (A = 1, E = e_1)\right].$$

Under the observed data distribution P_O we have the following statistical parameter:

$$\psi^t\left(P_O\right) = E_{W(B)}\left[\left\{E\left(Y(B) \mid W(B), B = (1, e_1)\right) - E\left(Y(B) \mid W(B), B = (0, e_0)\right)\right\} \mid B = (1, e_1)\right].$$

The causal parameter $\psi^t\left(P_{U,X}\right)$ and the statistical parameter $\psi^t\left(P_O\right)$ are equivalent under the same assumptions described for the additive treatment effect.

Model and Target Parameters: Explanatory Variables

The additive effect was also of interest for the explanatory variables (these other target variables will be referred to as A^*). Most were already binary, but the few that were not were dichotomized as follows:

- Age at hospitalization: $A^* = 1$ when age > 69
- Hospital length of stay (LOS): $A^* = 1$ when LOS > 4 days
- Number of diagnoses: $A^* = 1$ when number of diagnoses > 12
- Readmission risk score: $A^* = 1$ when score > 5

To avoid uncontrolled differences between the intervention and control groups due to the presence of the heart failure program, each group was considered separately in this portion of the analysis. This resulted in two effect estimates for each target variable A^*.

Let $P_{O^*,e}$ denote the observed data distribution within a particular group – specifically, for the group with environmental variable $E = e$. These observed data $O_e^* = (W, A^*, Y)$ could be considered a missing data structure on a hypothetical full data structure $X^* = (W, A^*, Y)$ which, assuming exogenous $U^* = (U_W, U_{A^*}, U_Y) \sim P_{U^*}$, could have the following NPSEM:

$$W = f_{W^*}(U_W)$$
$$A^* = f_{A^*}(W, U_{A^*})$$
$$Y = f_Y(W, A^*, U_Y).$$

Under this NPSEM, the causal parameter of interest (additive effect) would therefore be

$$\psi\left(P_{U^*,X^*}\right) = E[Y_1] - E[Y_0],$$

and the statistical parameter would be

$$\psi\left(P_{O^*,e}\right) = E_W\left[E\left(Y \mid A^* = 1, W\right) - E\left(Y \mid A^* = 0, W\right)\right].$$

For the equivalence $\psi\left(P_{O^*,e}\right) = \psi\left(P_{U^*,X^*}\right)$ to hold, slightly weaker assumptions are required than in the case where the target variable A is the heart failure intervention; this is due to the fact that the community-level variable E is used in defining the target population, and is therefore no longer an additional variable requiring consideration in the NPSEM. First, as implied by the NPSEM for X^*, we must assume that, given W, A^* is independent of the counterfactual outcome Y_{a^*} for $a^* \in \{0,1\}$. This has been called the "no unmeasured confounding" assumption, and is the analog of the exclusion assumption previously explained (van der Laan & Robins, 2003). Also

required is the positivity or experimental treatment assignment (ETA) assumption (van der Laan & Robins, 2003; Messer, Oakes, & Mason, 2010):

$$\mathrm{P}\left(A^* = a^* \mid W\right) > 0.$$

Finally, we must assume that the observed data O_e^* are a missing data structure on X^* (consistency assumption) (van der Laan & Robins, 2003).

Parameter Estimation

Parameters were estimated using targeted maximum likelihood estimation (TMLE), which combines features of both estimating equation and likelihood approaches (van der Laan & Rubin, 2006; van der Laan & Rose, 2011). First, let us define $Q(A,W)$ as $E\left[Y \mid A,W\right]$ and $g(0 \mid W)$ as the estimated probability that $A = 0$ given W. $Q_n^0(A,W)$ then denotes the initial estimate of $Q(A,W)$ and $g_n(0 \mid W)$ denotes the estimate of $g(0 \mid W)$. A represents the treatment or target variable of interest. The TMLE for the additive effect is then defined as follows:

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}\left[Q_n^1(1,W_i) - Q_n^1(0,W_i)\right], \text{ where}$$

$$\operatorname{logit}\left[Q_n^1(A,W)\right] = \operatorname{logit}\left[Q_n^0(A,W)\right] + \varepsilon_n h(A,W), \text{ and}$$

$$h(A,W) = \frac{I(A=1)}{g_n(1 \mid W)} - \frac{I(A=0)}{g_n(0 \mid W)}.$$

The parameter ε is estimated by maximum likelihood.

The TMLE for the additive effect amongst the treated population is as follows: (Hubbard, Jewell, & van der Laan, 2011)

$$\psi_n^t\left(Q_n^*, g_n^*\right) = \frac{1}{\sum_{i=1}^{n} I\left(A_i = 1\right)}\sum_{i=1}^{n} I\left(A_i = 1\right)\cdot\left[Q_n^*(1,W_i) - Q_n^*(0,W_i)\right].$$

Q_n^* and g_n^* are obtained using an iterative process. Specifically, initial fits Q_n^0 and g_n^0 for Q and g are estimated. At each iteration j, $Q_n^j\left(A,W\right)$ and $g_n^j\left(A \mid W\right)$ are then computed as follows:

$$\operatorname{logit}\left[Q_n^j\left(A,W\right)\right] = \operatorname{logit}\left[Q_n^{j-1}\left(A,W\right)\right] + \varepsilon_{1n}^j \cdot c_1\left(g_n^{j-1}\right)\left(A,W\right) \text{ and}$$

$$\operatorname{logit}\left[g_n^j\left(A \mid W\right)\right] = \operatorname{logit}\left[g_n^{j-1}\left(A \mid W\right)\right] + \varepsilon_{2n}^j \cdot c_2\left(Q_n^{j-1}, g_n^{j-1}\right)\left(W\right).$$

The estimates ε_{1n}^{j} and ε_{2n}^{j} are obtained by maximum likelihood. The definitions of $c_1\left(g_n^{j-1}\right)(A,W)$ and $c_2\left(Q_n^{j-1}, g_n^{j-1}\right)(W)$ are given below:

$$c_1\left(g_n^{j-1}\right)(A,W) = I(A=1) - \frac{I(A=0)g_n^{j-1}(1|W)}{g_n^{j-1}(0|W)}$$

$$c_2\left(Q_n^{j-1}, g_n^{j-1}\right)(W) = Q_n^{j-1}(1,W) - Q_n^{j-1}(0,W) - \psi_n^t\left(Q_n^{j-1}, g_n^{j-1}\right).$$

This process is repeated until ε_{1n}^{j} and ε_{2n}^{j} converge to zero, and Q_n^* and g_n^* are defined as $Q_n^j(A,W)$ and $g_n^j(A|W)$ at the final iteration. Convergence was considered to have been reached when both ε_{1n}^{j} and ε_{2n}^{j} achieved an absolute value less than 10^{-6}.

TMLE has several valuable theoretical properties that make it a good estimation choice. It is a substitution estimator, which means that it respects the bounds on the parameter. TMLE is also double-robust (DR), meaning that ψ_n and ψ_n^t will be consistent if either *g* or *Q* are correctly specified. In this context, consistency means that an estimator ψ_n converges in probability to the true parameter $\psi(P_O)$ as $n \to \infty$; TMLE is also asymptotically efficient when the working model contains the true models. Instead of focusing on the entire distribution P_O, TMLE also attempts to reduce bias in relation to the desired feature of the observed data distribution (the parameter of interest); this is the goal of updating the initial estimate or estimates with *h(A,W)* (additive effect) or with $c_1(g)(A,W)$ and $c_2(Q,g)(W)$ (effect amongst the treated).

Initial Estimation of Q and g

The initial estimate of *Q* was obtained by super learning, as implemented in the R package *SuperLearner*; this implementation uses *V*-fold cross-validation to construct a convex combination of candidate estimators. Super learning is a desirable estimation choice because it (1) respects what is known about the true form of *Q* (nothing), (2) considers multiple models and utilizes data-adaptive methods to increase the possibility of capturing the true *Q*, and (3) avoids manual manipulation of the data in choosing the final model. Super learning also performs as well asymptotically as the so-called "oracle" selector which, in the context of a particular loss function, minimizes risk under the true data-generating distribution (Sinisi S. , Polley, Petersen, Rhee, & van der Laan, 2007; van der Laan, Polley, & Hubbard, 2007).

The library of candidate estimators for the super learner included the following: main terms logistic regression (R function *glm*); logistic regression with the target variable *A* as the sole predictor; generalized additive models (as implemented in the R package *gam*); stepwise logistic regression, with all main terms as the maximum size model (as implemented in the R package step); and polychotomous regression and multiple classification (as implemented in the R package polspline, 5-fold cross-validation) (Kooperberg, Bose, & Stone, 1997). Seven-fold

cross-validation was specified for super learner model choice, and the target variable of interest was always required to be present in the final model selected by each candidate estimator.

For all candidate estimators but *polyclass*, it was possible to force the target variable of interest into the final model. This was not possible within the framework of the *polyclass* function, so the following workaround was constructed. First, *polyclass* was fit on the entire dataset (or, in the case of the super learner, on the training dataset). Second, the predicted probability of *Y =1* was obtained, per the *polyclass* fit. Finally, this fit was used in a logistic regression model containing the target variable of interest *A*:

$$\text{logit}\left[E\left(Y \mid A,W\right)\right] = \gamma_0 + \gamma_1 A + \gamma_2 \text{logit}\left[Z_n\left(A,W\right)\right] \cdot A + \gamma_3 \text{logit}\left[Z_n\left(A,W\right)\right],$$

where $Z_n\left(A,W\right)$ represents the fitted probabilities from *polyclass*.

The generalized additive model used smoothing splines with two target degrees of freedom for covariates with more than four unique values, and linear terms for all other covariates. These are the default specifications of for the *gam* function according to the R function *SuperLearner*.

The so-called "treatment mechanism" *g* (also known as the propensity score) was estimated using forward stepwise logistic regression (R function *step*).Though super learning could also have been applied to estimate *g*, it can be overly aggressive and result in predicted probabilities near zero or one when the number of covariates in *W* is reasonably large and the sample size moderate in comparison. Stepwise logistic regression was therefore determined to be preferable for initial estimation of *g*. Collaborative TMLE can also be applied to minimize this issue, but was not utilized in this analysis (van der Laan & Gruber, 2009). No truncation of predicted probabilities was required.

Inference

Inference was obtained using the estimated influence curve (IC). For *n* large enough, ψ_n and ψ_n^t will approximately follow a normal distribution, with variance equal to $\text{var}\left(IC\right)/\sqrt{n}$. Under the empirical distribution P_n, the estimated influence curves IC_n and IC_n^t for the estimators ψ_n and ψ_n^t, respectively, are given below (van der Laan & Robins, 2003; van der Laan & Rose, 2011; Hubbard, Jewell, & van der Laan, 2011).

$$IC_n(O) = \left(\frac{I(A=1)}{g_n(1\mid W)} - \frac{I(A=0)}{g_n(0\mid W)}\right)\left(Y - Q_n^1\left(A,W\right)\right) + Q_n^1\left(1,W\right) - Q_n^1\left(0,W\right) - \psi_n$$

$$IC_n^t(O) = \left(\frac{I(A=1)}{P(A=1)} - \frac{I(A=0)g_n^*(1|W)}{P(A=1)g_n^*(0|W)} \right) \left(Y - Q_n^*(A,W) \right) + \frac{I(A=1)}{P(A=1)} \left[Q_n^*(1,W) - Q_n^*(0,W) - \psi_n^t \left(Q_n^*, g_n^* \right) \right].$$

For comparison, and to provide the most conservative inference for ψ_n^t, standard errors were also computed using 10-fold cross-validation. At each cross-validation sample split, the entire parameter estimation process for ψ_n^t described above was conducted on the training set. The estimated influence curve was then evaluated at the validation set, with parameter estimates determined by the training set. This was repeated for each cross-validation sample split, and the resulting cross-validated influence curve estimate used to estimate the standard error of ψ_n^t. Though the technique was not applied in this analysis, there can also be advantages to using cross-validation in conjunction with TMLE for calculation of the parameter estimates themselves (Zheng & van der Laan, 2010).

Power Calculation

Because the estimated standard errors were found to be relatively large in relation to the effect estimates, it was of interest to investigate the impact of sample size on the power of the statistical test, within the framework of the estimated effect size. The treatment effect among the treated ψ_n^t for the 30-day heart failure readmission outcome was found to be the largest estimated treatment effect, so this was chosen as the focus of the power calculation. Standard errors for parameter estimates at different sample sizes were estimated using a parametric bootstrap, as described below.

Each parametric bootstrap sample $P^{\#}$ was generated from the estimated observed data distribution $\hat{P}_O$, as defined by the Q_n^0 and g_n^0 previously estimated. Each sample $P^{\#}$ was N i.i.d. observations $O^{\#} = \left(W^{\#}, A^{\#}, Y^{\#} \right)$ of $\hat{P}_O$. For each bootstrap sample $P^{\#}$, $W^{\#}$ was generated first by sampling the rows of W with replacement N times. Next, g_n^0 was applied to $W^{\#}$, and the resulting predicted probabilities used to generate $A^{\#}$ as Bernoulli random variables with probability $p_a = g_n^0 \left(1 | W^{\#} \right)$. Finally, Q_n^0 was applied to $A^{\#}$ and $W^{\#}$ to generate $Y^{\#}$ as Bernoulli random variables with probability $p_y = Q_n^0 \left(1 | A^{\#}, W^{\#} \right)$. The parameter estimate $\psi_N^t \left(P^{\#} \right)$ was then obtained from the bootstrap sample according to the same process employed for the original sample.

B = 500 bootstrap samples were generated for $N \in \{1376, 2500, 3500, 5000\}$. At each N, the power π_N to detect at the 5% level an effect at least as large as the estimated original full data effect $\psi_n^t \left(\hat{P}_O \right)$ was then approximated as

$$\pi_N \approx 1-\Phi\left(1.96-\psi_n^t\left(\hat{P}_O\right)/\sigma_N\right),$$

where Φ represents the standard normal cumulative distribution function and σ_N denotes the sample standard deviation of $\psi_N^t\left(P^{\#}\right)$. This approximation is appropriate because of the asymptotic normality of the estimator, and the reasonable sample size (1376) in the original sample.

4.2.3 Prediction of Hospital Readmission

Five prediction models were constructed for each of the six readmission outcomes: two full prediction models including all explanatory variables in Table 4.1, two prediction models including the risk score components only (shaded in Table 4.1), and one prediction model incorporating the risk score alone. Logistic regression was used to fit the model including the score alone. One full prediction model utilized super learning (10-fold cross-validation), and the other employed main terms logistic regression; the same was true of the two score component models. The super learner library of candidate estimators included the following: main terms logistic regression (*glm*), generalized additive models (*gam*), polychotomous regression and multiple classification (*polyclass*, 5-fold cross-validation), and stepwise logistic regression (*step*). The specifications for the generalized additive models implementation *gam* were the same as described previously. The workaround described previously to fix variables into *polyclass* was not needed here, because no variable was required to be present in the final prediction model.

Because the heart failure intervention was intended to reduce readmissions, only the control group (n = 788) was used in this assessment of predictive performance.

Predicted probabilities of hospital readmission were estimated using 10-fold cross-validation, and receiver operating characteristic (ROC) curves were used to assess the predictive performance of the various models (R package *ROCR*) (Sing, Sander, Beerenwinkel, & Lengauer, 2005). ROC curves plot the true positive rate (correct prediction of hospital readmission in the cases when readmission occurred) against the false positive rate (incorrect prediction of hospital readmission in the cases when readmission did not occur). Rate estimates calculated using V-fold cross-validation will be unbiased for sample size $n\left(1-1/V\right)$, where V is the number of cross-validation folds; in this analysis, $n\left(1-1/V\right)=788\left(1-1/10\right)\approx 709$.

4.3 Results

Table 4.2 shows descriptive statistics for the intervention and control groups. The number of admissions for patients with cardiac surgery in the past year was so small in both groups that the variable had to be excluded from any adjustment set (and from individual analysis).

The variable importance estimates of the additive treatment effect are shown in Table 4.3. In the causal world, these represent the estimated difference in the prevalence of hospital readmission after an initial hospitalization for heart failure if the heart failure intervention had been available

for everyone, and the equivalent prevalence of readmission if the heart failure intervention had been available to no one. The final column in the table shows the estimate as a percent change from the expected mean outcome with no heart failure intervention ($\hat{E}[Y_0]$). The largest estimated percent change is for the 30-day heart failure readmission outcome, at a reduction of 16.8% over the predicted mean outcome with no heart failure intervention (11.8%), with a 95% confidence interval (CI) ranging from a 48.1% reduction to 14.6% increase. All 95% confidence intervals cross zero, however, so the results are not significant at the 5% level (two-tailed).

Table 4.2 - Descriptive statistics. Mean values for each explanatory variable in the intervention and control groups.

	No HF Intervention	HF Intervention
African American race	0.57	0.53
Age at hospital admission	68.5	69.6
Cardiac surgery in past year	0.03	0.01
Chronic lung disease	0.38	0.40
Diabetes mellitus	0.38	0.36
Discharged to Home Health	0.27	0.30
Hospital length of stay (days)	4.0	4.5
Idiopathic cardiomyopathy	0.31	0.35
Inpatient hospitalization in past year	0.55	0.56
Ischemic heart disease	0.38	0.38
Medicaid	0.46	0.47
Medicare	0.59	0.64
Number of diagnoses	9.6	12.6
Renal disease	0.24	0.37
Telemetry during hospitalization	0.87	0.88
Valvular heart disease	0.19	0.26
Weekend hospital admission	0.21	0.21
Weekend hospital discharge	0.24	0.29
Readmission risk score	4.7	4.9
Total hospitalizations	788	588

Table 4.4 shows the estimated treatment effect amongst the treated for the six readmission outcomes. If causal assumptions hold, these estimates can be interpreted as the difference within the intervention group between the mean outcome (prevalence of hospital readmission after an initial hospitalization for heart failure) after implementation of the heart failure intervention and the mean outcome that would have been observed if no intervention had been available. The column $\hat{E}[Y_0 \mid A=1]$ in Table 4.4 is the no-intervention estimated mean outcome within the intervention group, and the final column is the variable importance estimate shown as a percentage of $\hat{E}[Y_0 \mid A=1]$. As in Table 4.3, all confidence intervals include zero, so no estimate is significant at the 5% level (two-tailed). The 30-day heart failure readmission effect estimate once again represents the largest percent change in mean outcome, and is also the largest in

magnitude. This estimate represents a 26.5% reduction in prevalence of 30-day readmission for heart failure (after an initial hospitalization for heart failure) over the predicted mean outcome for the intervention group were there no intervention (12.6%), with a 95% CI ranging from a reduction of 67.3% to an increase of 14.2%.

Table 4.3 - Heart failure intervention, estimated additive treatment effect.

Outcome	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
All Cause Readmission					
30-Day	0.001	0.024	(-0.047, 0.049)	0.208	0.5%
90-Day	0.004	0.028	(-0.052, 0.059)	0.381	0.9%
180-Day	0.021	0.029	(-0.035, 0.078)	0.480	4.5%
Heart Failure Readmission					
30-Day	-0.020	0.019	(-0.057, 0.017)	0.118	-16.8%
90-Day	-0.022	0.023	(-0.067, 0.024)	0.215	-10.0%
180-Day	-0.019	0.026	(-0.069, 0.031)	0.284	-6.7%

Table 4.4 also compares standard errors estimated using the original full data influence curve with the cross-validated influence curve, and the corresponding 95% CIs. In all instances the standard errors are larger using cross-validation. The differences between the two standard errors are smallest for the estimates associated with the 30-day and 90-day heart failure readmission outcomes; the corresponding effect sizes for these estimates are also the largest in magnitude amongst those in Table 4.4.

Table 4.4 - Heart failure intervention, estimated treatment effect amongst the treated.

				Cross-validated			
Outcome	Estimate	SE	95% CI	SE	95% CI	$\hat{E}[Y_0 \mid A=1]$	Estimate as % change
All Cause Readmission							
30-Day	0.000	0.029	(-0.056, 0.056)	0.033	(-0.063, 0.064)	0.213	0.2%
90-Day	-0.004	0.034	(-0.071, 0.062)	0.039	(-0.08, 0.071)	0.395	-1.1%
180-Day	0.009	0.034	(-0.057, 0.074)	0.038	(-0.067, 0.084)	0.494	1.7%
Heart Failure Readmission							
30-Day	-0.034	0.025	(-0.083, 0.016)	0.026	(-0.085, 0.018)	0.126	-26.5%
90-Day	-0.026	0.029	(-0.082, 0.031)	0.030	(-0.084, 0.033)	0.220	-11.7%
180-Day	-0.017	0.031	(-0.078, 0.044)	0.037	(-0.091, 0.056)	0.290	-6.0%

For the 30-day heart failure readmission outcome, the estimated statistical power to detect a treatment effect amongst the treated equivalent in magnitude to the effect estimate in Table 4.4 (0.034) is shown in Table 4.5 for four possible sample sizes *N*. The power estimates assume the estimated observed data distribution $\hat{P}_O$, and a desired two-tailed significance level of 5%. The smallest sample size (1376) is equivalent to the sample size of the original sample, and its estimated standard error from the parametric bootstrap is 0.028, which is very close to the cross-validated standard error (0.026) for the equivalent estimate shown in Table 4.4. The estimated power π_N at the original sample size is very low (0.218), and at N = 5000 (over 3.5 times larger

than the original sample size) is estimated at 0.657, which corresponds to an estimated false negative rate (1-π_N) of 0.343, meaning an estimated 34.3% failure to achieve statistical significance at the 5% level (two-tailed) when the true effect size is ± 0.034.

Table 4.5 - Estimated power at different sample sizes *N* to detect at the 5% level (two-tailed) a treatment effect amongst the treated with absolute value of at least 0.034 for the 30-day heart failure readmission outcome.

N	σ_N	π_N
1376	0.028	0.218
2500	0.021	0.359
3500	0.018	0.476
5000	0.014	0.657

Additive effect estimates for the binary explanatory variables for the 30-day, 90-day, and 180-day readmission outcomes are shown in Tables 4.6, 4.7, and 4.8, respectively. For a given readmission outcome, results are shown only for explanatory variables found to have effect estimates with at least one 95% CI that did not cross zero. Full results for all explanatory variables and all outcomes are listed in Tables A4.3, A4.4, and A4.5 in the Appendix. The causal interpretation of these effect estimates is the difference between the mean outcome were all hospital admissions to have a given characteristic or feature (the binary explanatory variable of interest) versus the mean outcome were no hospital admission to have the same characteristic or feature ($\hat{E}[Y_0]$). As before, the mean outcome is the prevalence of hospital readmission after an initial hospitalization for heart failure.

For the 30-day readmission outcomes (Table 4.6), only one explanatory variable, inpatient hospitalization in the past year, was found to have effect estimates with 95% CIs that excluded zero in all cases; for both heart failure and all cause 30-day readmission outcomes within the intervention and control groups, inpatient hospitalization within the past year was associated with an increase in mean outcome (effect estimates were all positive). Within the control group only, valvular heart disease and age at hospital admission over 69 were associated with a decrease in the mean 30-day all cause readmission outcome; age at hospital admission over 69 was also associated with a decrease in the mean 30-day heart failure readmission outcome (95% CIs excluded zero). African American race and readmission risk score greater than 5 were associated with an increase in the prevalence of 30-day heart failure readmission in the control group, and had effect estimates with 95% CIs excluding zero. Among the effect estimates with 95% CIs excluding zero for the intervention group only, chronic lung disease was associated with an increase in the prevalence of 30-day all cause readmission, and telemetry during hospitalization was associated with an increase in the prevalence of 30-day heart failure readmission; discharge to home health services and weekend hospital discharge were associated with a decrease in the prevalence of 30-day heart failure readmission. The effect estimate largest in magnitude in Table 4.6 (0.130) was associated with inpatient hospitalization in the past year for the 30-day all cause readmission outcome in the control group, and represents a 100.7%

increase in mean outcome over the expected mean outcome with no inpatient hospitalizations in the past year for anyone (12.9%), with a 95% CI ranging from 0.074 to 0.185, or from a 57.3% increase to an 144.1% increase.

Table 4.7 (90-day readmission outcomes) shows more overlap between the explanatory variables found to have effect estimates with 95% CIs excluding zero in the intervention and control groups. African American race, chronic lung disease, inpatient hospitalization in the past year and readmission risk score greater than 5 were found to be associated with an increase in the mean 90-day all cause readmission outcome for both groups, and use of telemetry during hospitalization was associated with an increase in the mean 90-day all cause readmission outcome in the intervention group only. Within both groups, age at hospital admission over 69 was associated with a decrease in the mean 90-day heart failure readmission outcome, and readmission risk score greater than 5 was associated with an increase in the mean 90-day heart failure readmission outcome. African American race, inpatient hospitalization in the past year, and weekend hospital discharge were associated with an increase in the mean 90-day heart failure readmission outcome for the control group only, and chronic lung disease and use of telemetry during hospitalization were associated with an increase in the same mean outcome for the intervention group only. Readmission risk score greater than 5 was the only explanatory variable with effect estimates whose 95% CIs excluded zero for both 90-day readmission outcomes and in both groups. The effect estimate largest in magnitude (0.266) corresponds once again to inpatient hospitalization in the past year for the all cause readmission outcome in the control group, representing a 121.0% increase in the prevalence of all cause 90-day readmission over the expected prevalence were no one to have had an inpatient hospitalization in the past year (22.0%). The 95% CI for this estimate ranged from 0.199 to 0.333, or from a 90.6% increase to a 151.5% increase.

Among the estimates with 95% CIs excluding zero in Table 4.8, inpatient hospitalization in the past year and readmission risk score greater than 5 were associated with an increase in all mean 180-day readmission outcomes in both the intervention and control groups. African American race was associated with an increase in mean 180-day all cause readmission in both groups, and with an increase in mean 180-day heart failure readmission for the control group only. Also in the control group, valvular heart disease was associated with a decrease in mean 180-day all cause readmission, and age at hospital admission over 69 was associated with a decrease in mean 180-day heart failure readmission. In the intervention group, chronic lung disease and use of telemetry during hospitalization were associated with an increase in mean 180-day readmission (both all cause and heart failure); ischemic heart disease, hospital length of stay greater than 4 days, and discharge to home health services were associated with a decrease in mean 180-day readmission, but for the heart failure readmission outcome only. Once again, the effect estimate largest in magnitude in Table 4.8 corresponds to inpatient hospitalization in the past year for the all cause readmission outcome in the control group. This estimate (0.305) represents an increase of 105.0% in the prevalence of 180-day all cause readmission over the expected prevalence (29.0%) were no one to have had an inpatient hospitalization in the past year at the time of initial hospitalization for heart failure (95% CI 0.234 to 0.375, or an 80.7% increase to an 129.2% increase).

Diabetes mellitus, renal disease, idiopathic cardiomyopathy, number of diagnoses > 12, Medicaid payer, Medicare payer, weekend hospital admission, and hospital at which heart failure initial hospitalization occurred were not found to be significantly associated (5% level, two-tailed) with any readmission outcome.

Figures 4.1, 4.2 and 4.3 compare cross-validated estimated ROC curves for models predicting each of the six readmission outcomes in the control group. The largest difference between the ROC curves for the main terms logistic regression models and the equivalent super learner models was observed for the 30-day all cause readmission outcome; the main terms logistic regression score components model performed slightly better than the analogous super learner model, with an AUC of 0.61 versus an AUC of 0.59 for the super learner model. Otherwise, the ROC curves for the main terms logistic regression models were almost identical to the corresponding super learner models, and the AUC values were always within 0.01 of each other.

For all outcomes, the full prediction models performed best, with an area under the ROC curve (AUC) ranging from 0.64 for the super learner model for all cause 30-day readmission to 0.73 for heart failure 90-day and 180-day readmission (both super learner and main terms logistic regression models). The score only models performed worst, with an AUC low of 0.54 for heart failure 30-day readmission and high of 0.6 for heart failure 180-day readmission. The score components models fell in between, with AUC values ranging from a low of 0.59 for the super learner model for all cause 30-day readmission and a high of 0.65 for the main terms logistic regression model for all cause 90-day readmission. The largest difference in predictive power between the full prediction model and the score components model, in terms of AUC, was observed for the 180-day readmission outcomes.

All prediction model types performed worst for the 30-day readmission outcomes. Predictive performance of analogous score components and score only models was similar for heart failure and all cause 30-day readmission outcomes, and the full prediction models predicted heart failure 30-day readmission slightly better than all cause 30-day readmission (AUC = 0.69 versus AUC = 0.64-0.65).

ROC curves and AUC values were fairly similar within model types for analogous 90-day and 180-day readmission outcomes; the largest difference in terms of AUC was observed between the score components predictive models for 90-day and 180-day all cause readmission outcomes, with better prediction of 90-day readmission (AUC difference of 0.04). All cause and heart failure readmission were predicted comparably well within the 90-day and 180-day readmission outcomes, with no AUC difference larger than 0.02.

Table 4.6 - Variable importance estimates for explanatory variables, 30-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals (CIs) that do not cross zero. Only explanatory variables with at least one 95% CI excluding zero are shown.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
All Cause Readmission										
African American	0.072	0.032	(0.009, 0.135)	0.136	52.7%	0.014	0.038	(-0.061, 0.089)	0.202	7.0%
Valvular Heart Disease	-0.107	0.033	(-0.171, -0.043)	0.213	-50.2%	-0.075	0.039	(-0.152, 0.001)	0.229	-32.9%
Chronic Lung Disease	0.026	0.034	(-0.04, 0.092)	0.193	13.6%	0.108	0.039	(0.031, 0.184)	0.168	64.2%
Inpatient hospitalization in past year	0.130	0.029	(0.074, 0.185)	0.129	100.7%	0.092	0.037	(0.02, 0.163)	0.163	56.3%
Age at hospital admission >69	-0.091	0.046	(-0.181, -0.002)	0.214	-42.6%	-0.020	0.062	(-0.143, 0.102)	0.194	-10.5%
Readmission risk score >5	0.095	0.035	(0.025, 0.164)	0.185	51.1%	0.015	0.037	(-0.058, 0.089)	0.214	7.2%
Heart Failure Readmission										
Telemetry during hospitalization	-0.011	0.039	(-0.088, 0.065)	0.119	-9.7%	0.073	0.020	(0.033, 0.112)	0.028	260.1%
Discharged to home health	-0.026	0.021	(-0.067, 0.015)	0.106	-24.5%	-0.087	0.022	(-0.13, -0.044)	0.115	-76.2%
Inpatient hospitalization in past year	0.114	0.022	(0.072, 0.157)	0.047	242.9%	0.057	0.024	(0.01, 0.104)	0.062	92.2%
Weekend hospital discharge	0.049	0.029	(-0.008, 0.106)	0.100	48.5%	-0.058	0.022	(-0.101, -0.015)	0.109	-53.0%
Age at hospital admission >69	-0.090	0.041	(-0.171, -0.01)	0.140	-64.4%	-0.039	0.024	(-0.087, 0.008)	0.093	-42.3%

Table 4.7 - Variable importance estimates for explanatory variables, 90-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals (CIs) that do not cross zero. Only explanatory variables with at least one 95% CI excluding zero are shown.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
All Cause Readmission										
African American	0.114	0.036	(0.044, 0.185)	0.258	44.4%	0.098	0.048	(0.004, 0.192)	0.335	29.3%
Chronic Lung Disease	0.091	0.042	(0.008, 0.174)	0.350	25.9%	0.122	0.047	(0.03, 0.213)	0.362	33.6%
Telemetry during hospitalization	-0.046	0.069	(-0.181, 0.09)	0.400	-11.4%	0.138	0.050	(0.04, 0.237)	0.254	54.3%
Inpatient hospitalization in past year	0.266	0.034	(0.199, 0.333)	0.220	121.0%	0.166	0.043	(0.082, 0.251)	0.295	56.3%
Readmission risk score >5	0.123	0.037	(0.05, 0.197)	0.332	37.1%	0.091	0.044	(0.005, 0.177)	0.356	25.5%
Heart Failure Readmission										
African American	0.104	0.033	(0.041, 0.168)	0.134	77.9%	0.032	0.040	(-0.046, 0.111)	0.174	18.5%
Chronic Lung Disease	0.042	0.036	(-0.028, 0.112)	0.203	20.7%	0.082	0.037	(0.008, 0.155)	0.161	50.6%
Telemetry during hospitalization	-0.054	0.059	(-0.169, 0.061)	0.255	-21.1%	0.086	0.038	(0.012, 0.16)	0.112	76.4%
Inpatient hospitalization in past year	0.211	0.029	(0.154, 0.268)	0.095	221.9%	0.064	0.035	(-0.005, 0.133)	0.165	38.7%
Weekend hospital discharge	0.069	0.035	(0, 0.138)	0.195	35.5%	0.016	0.036	(-0.054, 0.086)	0.187	8.4%
Age at hospital admission >69	-0.123	0.043	(-0.207, -0.039)	0.241	-51.1%	-0.119	0.042	(-0.201, -0.036)	0.222	-53.5%
Readmission risk score >5	0.087	0.032	(0.025, 0.149)	0.188	46.1%	0.103	0.037	(0.031, 0.174)	0.156	66.0%

Table 4.8 - Variable importance estimates for explanatory variables, 180-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals (CIs) that do not cross zero. Only explanatory variables with at least one 95% CI excluding zero are shown.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
All Cause Readmission										
African American	0.108	0.039	(0.032, 0.183)	0.354	30.4%	0.095	0.047	(0.003, 0.188)	0.463	20.6%
Valvular Heart Disease	-0.106	0.051	(-0.206, -0.006)	0.473	-22.4%	-0.007	0.049	(-0.102, 0.088)	0.514	-1.3%
Chronic Lung Disease	0.074	0.043	(-0.009, 0.158)	0.446	16.6%	0.109	0.045	(0.021, 0.197)	0.480	22.7%
Telemetry during hospitalization	-0.015	0.068	(-0.147, 0.118)	0.469	-3.2%	0.208	0.053	(0.105, 0.311)	0.307	67.7%
Inpatient hospitalization in past year	0.305	0.036	(0.234, 0.375)	0.290	105.0%	0.261	0.044	(0.175, 0.348)	0.359	72.7%
Readmission risk score >5	0.088	0.039	(0.012, 0.165)	0.443	20.0%	0.107	0.044	(0.02, 0.193)	0.465	22.9%
Heart Failure Readmission										
African American	0.114	0.036	(0.043, 0.185)	0.187	61.0%	0.047	0.043	(-0.038, 0.131)	0.240	19.4%
Ischemic Heart Disease	0.015	0.032	(-0.048, 0.077)	0.232	6.3%	-0.097	0.041	(-0.178, -0.017)	0.299	-32.5%
Chronic Lung Disease	0.014	0.038	(-0.061, 0.088)	0.275	4.9%	0.107	0.042	(0.025, 0.188)	0.234	45.5%
Telemetry during hospitalization	0.010	0.060	(-0.109, 0.128)	0.264	3.7%	0.146	0.041	(0.066, 0.226)	0.140	104.5%
Discharged to home health	-0.035	0.034	(-0.102, 0.032)	0.263	-13.2%	-0.093	0.040	(-0.171, -0.016)	0.281	-33.2%
Inpatient hospitalization in past year	0.261	0.032	(0.198, 0.323)	0.133	196.2%	0.157	0.039	(0.081, 0.233)	0.189	83.0%
Hospital length of stay (days) >4	0.015	0.036	(-0.057, 0.086)	0.285	5.1%	-0.086	0.037	(-0.158, -0.013)	0.305	-28.0%
Age at hospital admission >69	-0.154	0.067	(-0.286, -0.022)	0.330	-46.7%	-0.094	0.118	(-0.326, 0.138)	0.353	-26.7%
Readmission risk score >5	0.102	0.036	(0.031, 0.172)	0.255	40.0%	0.104	0.040	(0.025, 0.183)	0.234	44.3%

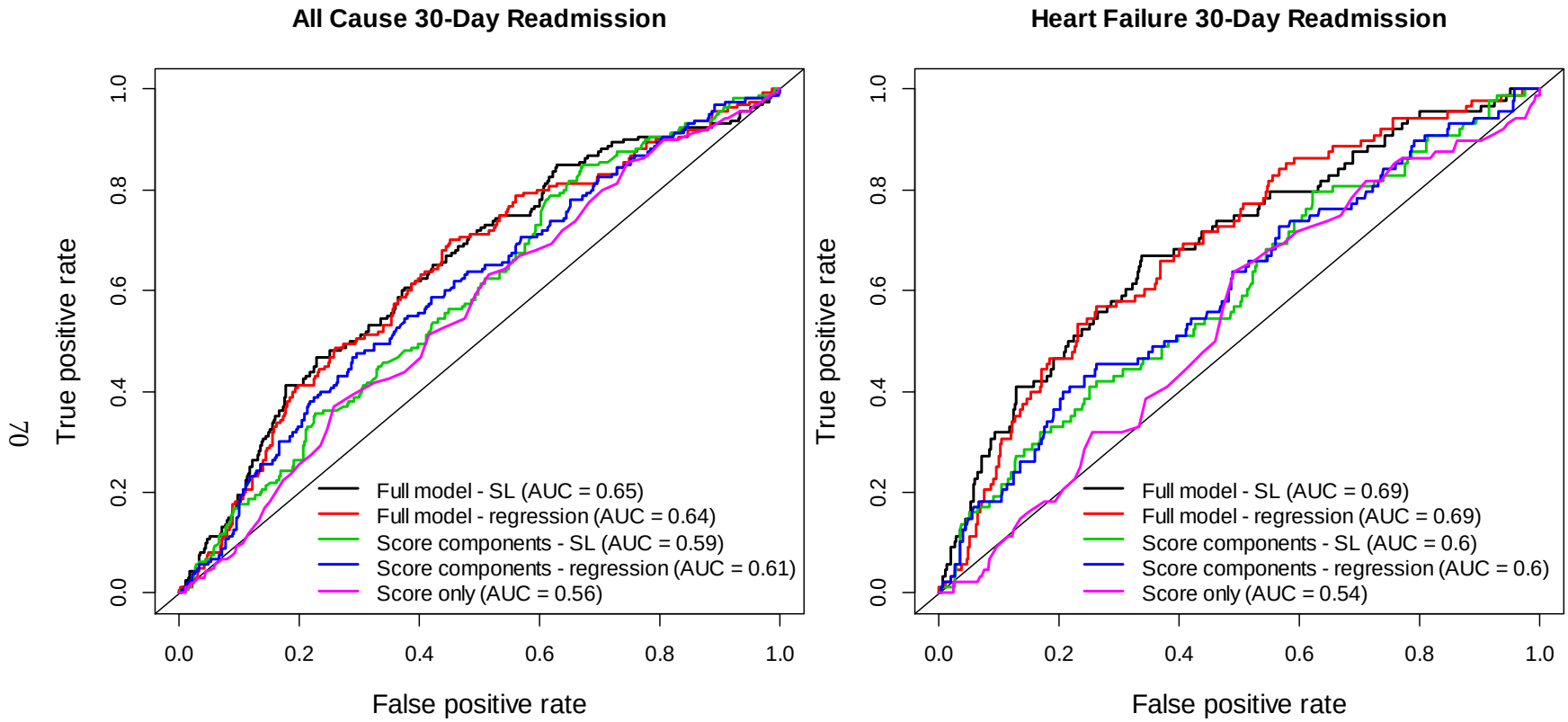

Figure 4.1 ROC curves for prediction of readmission within 30 days after initial hospitalization for heart failure. The left plot shows prediction of readmission for any reason, and the right plot shows prediction of readmission for heart failure. "SL" denotes super learner prediction models.

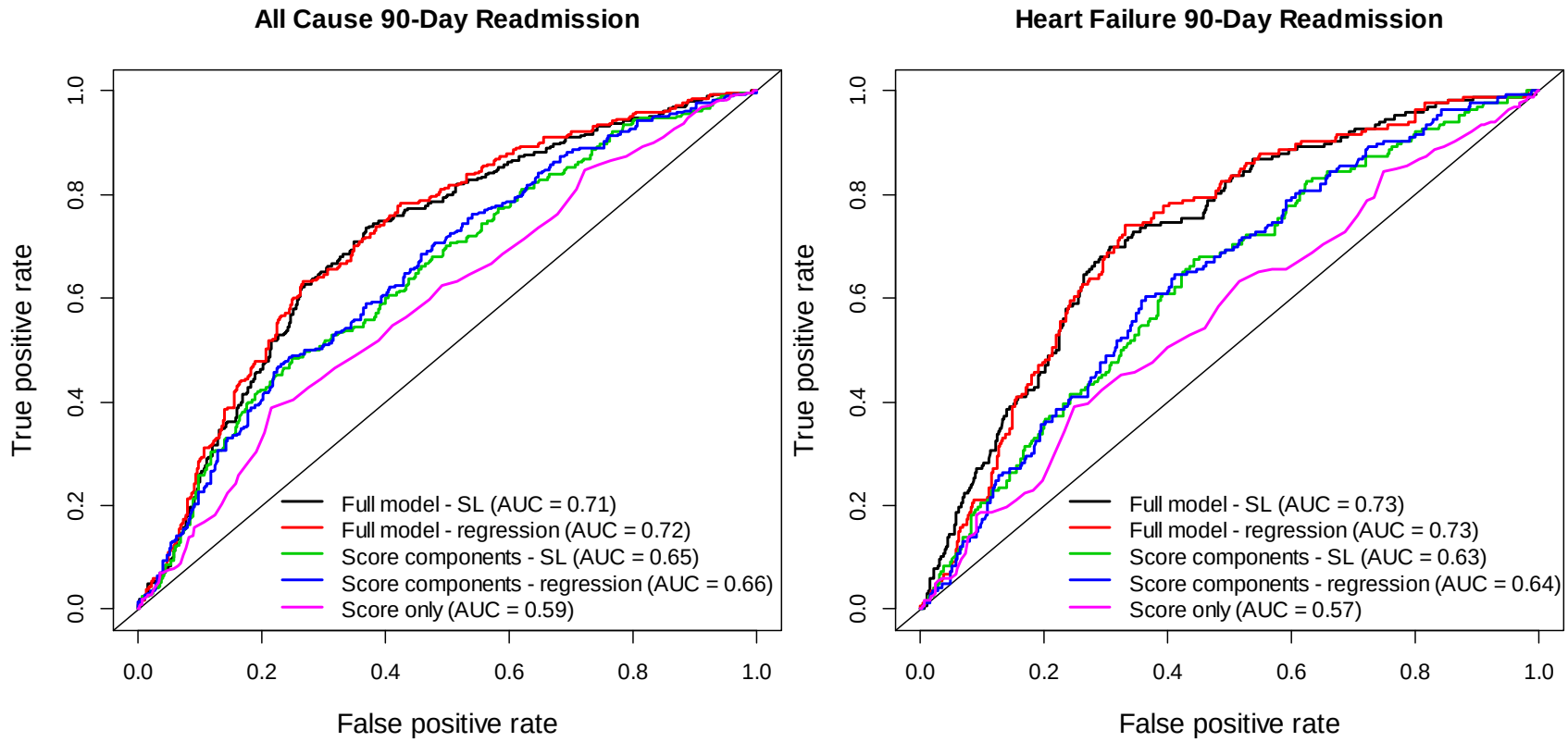

Figure 4.2 ROC curves for prediction of readmission within 90 days after initial hospitalization for heart failure. The left plot shows prediction of readmission for any reason, and the right plot shows prediction of readmission for heart failure. "SL" denotes super learner prediction models.

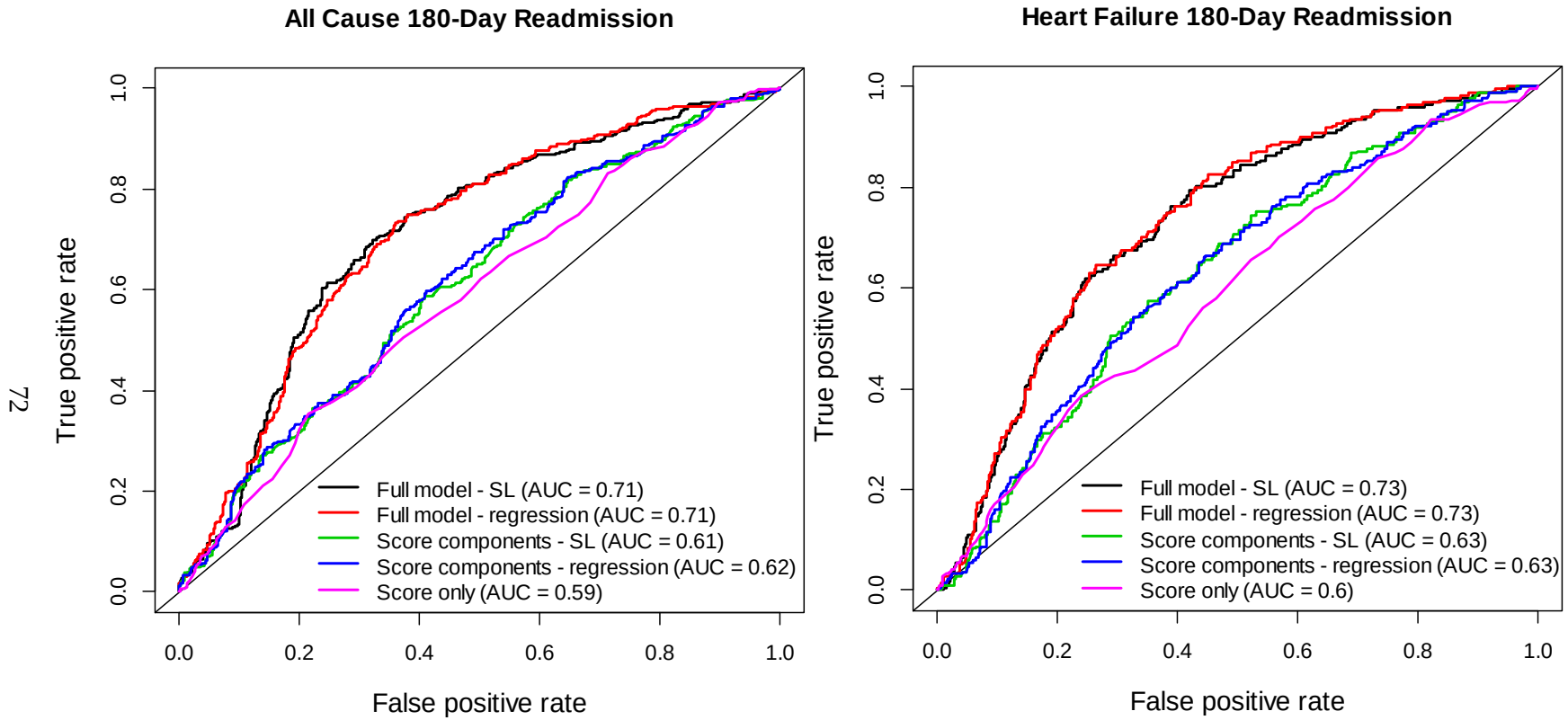

Figure 4.3 ROC curves for prediction of readmission within 180 days after initial hospitalization for heart failure. The left plot shows prediction of readmission for any reason, and the right plot shows prediction of readmission for heart failure. "SL" denotes super learner prediction models.

4.4 Discussion

There was no evidence in this analysis that the heart failure program impacted mean rates of readmission during its first year, in that no treatment effect estimate was significant at the 5% level (two-tailed). There were, however, some interesting differences in the effect estimates for the heart failure versus the all cause readmission outcomes. While both the additive effect and treatment effect amongst the treated estimates were very nearly zero for almost all of the all cause readmission outcomes, the effect estimates for the heart failure readmission outcomes were consistently negative, and in most cases much larger in magnitude than their all cause counterparts. For the estimated treatment effect amongst the treated, differences between effect estimates for the heart failure and all cause outcomes were particularly noticeable. The estimated treatment effect amongst the treated for the 30-day heart failure readmission outcome was the largest in magnitude of any treatment effect estimate, and represented a 26.5% reduction in the mean rate of 30-day heart failure readmission over the expected rate of readmission within the intervention group had there been no heart failure program. Though the estimate's 95% CI included zero, the estimated standard error was large in relation to the small effect size, and the estimated statistical power was very low. This means that while this analysis cannot reject the null hypothesis that the true treatment effect among the treated for the 30-day heart failure readmission outcome was truly zero, there is also not strong evidence to accept the null hypothesis. If possible, further analysis with an increased sample size would be desirable, particularly because one of the heart failure program's stated goals was to reduce readmission rates by 30%, a number very close to the percent change in 30-day heart failure readmission implied by the estimated treatment effect amongst the treated. Based on the estimated effect sizes, the most likely areas of impact are 30-day and 90-day heart failure readmission; it seems less likely that any impact on all cause readmission rates would be discovered.

A major challenge inherent in any attempt to evaluate the impact of the heart failure program is that it was community-level, and the program was completely implemented at both hospitals at the same time. The control group is retrospective, and any treatment effect is completely confounded with the effect of time. The treatment effect is only estimable with assumptions, such as that time only influences the outcome through the measured explanatory variables included in the adjustment set, which may or may not be a reasonable. Any future analysis would benefit from careful consideration of possible individual-level variables that could further express the time differences while retaining the possibility of achieving similar values in both intervention and control groups (van der Laan M. J., 2010).

Out of the 20 explanatory variables considered, 12 were found to be associated with at least one readmission outcome after adjusting for the other explanatory variables. Differences were noted between the sets of explanatory variables found to be associated with heart failure versus all cause readmission, and differences were also observed between those explanatory variables found to be associated with the same readmission outcome in the intervention versus the control group. Inpatient hospitalization in the year prior to hospitalization for heart failure was found to be associated most consistently with readmission after adjusting for the other explanatory variables, and also yielded the largest effect estimates in all but one instance.

There does appear to be an opportunity for increased predictive accuracy of both all cause and heart failure readmission – the addition of just eight variables to the score components model noticeably improved predictive performance for the 90-day and 180-day outcomes. Prediction of 30-day readmission was poor for all models, but could likely be improved with careful selection of additional predictors. By combining a large number of relevant variables available from the electronic health record, Wang et al. (2012) were able to achieve an AUC of 0.82 in prediction of 30-day and 1-year all cause readmission for heart failure patients receiving care from the Veterans Health Administration (Wang, et al., 2012). This analysis does not argue, therefore, for the predictive value of the particular set of explanatory variables or particular models used here; rather, it points out that improved prediction is possible, and should be further investigated if readmission risk assessment is to be used to identify the target population for an intervention.

Philbin and DiSalvo reported an AUC or *c*-statistic of 0.6 for their risk score, and an AUC of 0.62 for the corresponding multiple regression model including the risk score component variables (Philbin & DiSalvo, 1999). For 180-day readmission outcomes, the modified risk score used by the heart failure program performed equivalently in terms of AUC to the original risk score of Philbin and DiSalvo, despite differences in the number of variables included in the score, the study populations being considered, and the definition of the readmission outcome. Performance of the score components models in this study was also comparable to the performance of the multiple regression model of Philbin and DiSalvo. While this is an interesting result, neither the risk score nor the risk components models were strongly predictive of any readmission outcome, and the risk score performed only barely better than chance in predicting 30-day readmission.

There was no observed gain in predictive accuracy by using super learning over simple main terms logistic regression. It should be noted, however, that the super learner library of candidate estimators was quite minimal, and could be easily expanded to include others, possibly with improved results. The advantage of super learning is not that its resultant model will always outperform a simpler model, but that one need not choose between potential candidate models – it is instead possible to include any candidate model that could be predictive of the outcome of interest, and allow the super learner to weight the candidate models for optimal results. Discrete super learning is also possible; instead of returning convex combination of the candidate estimators, discrete super learning will instead use cross-validation to choose the single best estimator among the library of candidate estimators. There is operational value in choosing the simplest prediction method that will achieve the desired results, and more complexity does not guarantee a commensurate increase in predictive accuracy. Given that super learning is available, however, it would be of value to the researcher to compare results from such an approach with simpler methods to ensure that an avenue for improved prediction is not being ignored.

In this analysis, semi-parametric variable importance measures inspired by causal parameters were used to investigate the association of explanatory variables with rates of readmission after hospitalization for heart failure, and to evaluate the impact of a heart failure intervention on those same readmission outcomes. These measures were chosen with the goals of the analysis in mind: to be able to evaluate the questions of interest in a way that would result in effect estimates that could be interpreted in a real-world setting. This ruled out traditional regression approaches, because parameters from such models are only interpretable in the context of the particular

model chosen. Such models are almost always incorrect given the complexity of the systems involved in most public health applications, and it becomes unclear what an estimated parameter of an incorrectly specified regression model truly means in a real-world context. This does not mean that such parameters have no meaning, but they are of limited value if their interpretation cannot be easily defined outside of the estimation approach used. Variable importance measures such as those applied in this analysis address the questions of interest and are defined separately from the estimation methods and models chosen. With assumptions, they can be extended to make causal inferences, but even when causal assumptions are not met, effect estimates remain informative and have interpretations that can be readily understood in the context of the application being considered. These sorts of measures can also utilize the flexibility of machine learning algorithms and methods that utilize multiple candidate models (such as super learning) in the estimation process, maximizing what can be learned from the data and respecting the lack of prior knowledge about the true form of the data-generating distribution.

Chapter 5

Conclusions

This dissertation has been intended as a practical illustration of causal inference-inspired semi-parametric methods and data-adaptive estimation as applied to three real data problems. Because all three analyses handle real data, they face the central issue inherent to all real analyses: lack of knowledge about the true parameter of interest and the true data-generating distribution. In Chapter 2, a different story was told by the semi-parametric variable importance analysis than by the regression analysis: a significant association ($p < 0.05$) with the outcome was found for a particular explanatory variable by the former method but not by the latter. While the semi-parametric variable importance method has theoretical advantages over traditional regression analysis, it remains unknown which method returned the correct result in a general sense: we have no way of knowing whether that particular explanatory variable was in truth associated with the outcome or not. Similarly, in Chapter 3, four estimators were considered, and while these could be compared somewhat in terms of estimated standard errors and parametric bootstrap-estimated ETA bias, it is impossible to determine which estimator came closest to the true parameter value for that particular application.

Simulation is the only way to truly evaluate and compare estimator performance, because only then is the truth actually known. While various simulation studies have been conducted evaluating the behavior of the estimators used in this dissertation, it is unlikely that any closely mimicked the complex data structures analyzed here. There remains, therefore, an opportunity for further evaluation and benchmarking of estimator performance in more complex simulation scenarios. It would also be of interest to investigate in such scenarios whether the use of super learning with a limited library of candidate models to estimate the data-generating distribution resulted in an appreciable gain over the use of simpler estimation methods.

The challenges of real data problems and opportunities for further comparison of estimators and estimation algorithms only highlight the need for flexible variable importance and estimation methods such as those considered in this dissertation. The search for answers to research questions in the real world should always start with a parameter or quantity of interest that has meaning in the real world. When the true form of the data-generating distribution is unknown, this means that the definition of the parameter of interest must be separated from the specifics of the estimation process. Once this separation has been made, there is flexibility to investigate multiple estimation options and make the best choices possible for a given analysis without a resulting change in the definition of the estimated parameter. This is a marked difference from measures of effect based on parameters of pre-specified models, where any change to the model changes the interpretation of a given parameter in the model. If more public health analyses were based on estimation of parameters with straightforward real-world interpretations, one could perhaps hope for more comparability of results across studies of similar applications.

References

Agence National de Recerche sur le SIDA (ANRS). (n.d.). *ANRS genotypic resistance guidelines (version 13)*. Retrieved June 3, 2009, from http://www.hivfrenchresistance.org

Agency for Healthcare Research and Quality. (n.d.). Provisions in the Affordable Care Act that relate to PSOs and reducing unnecessary readmissions. Rockville, Maryland, USA. Retrieved 7 29, 2012, from http://www.pso.ahrq.gov/readmin/psoaca.htm

Ahern, J., Hubbard, A., & Galea, S. (2009). Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. *American Journal of Epidemiology, 169*(9), 1140-1147.

Altmann, A., Sing, T., Vermeiren, H., Winters, B., Van Craenenbroeck, E., Van der Borght, K., et al. (2009). Advantages of predicted phenotypes and statistical learning models in inferring virological response to antiretroviral therapy from HIV genotype. *Antiviral Therapy, 14*(2), 273-283.

Anderson, J. A., Jiang, H., Ding, X., Petch, L., Journigan, T., Fiscus, S. A., et al. (2008). Genotypic susceptibility scores and HIV type 1 RNA responses in treatment-experienced subjects with HIV type 1 infection. *AIDS Research and Human Retroviruses, 24*(5), 685-694.

Andrews, M. (2011, February 22). Health law forces changes to reduce hospitals readmissions. Retrieved from http://www.kaiserhealthnews.org/Features/Insuring-Your-Health/Michelle-Andrews-on-hospital-readmissions.aspx

Assoumou, L., Brun-Vézinet, F., Cozzi-Lepri, A., Kuritzkes, D., Phillips, A., Zolopa, A., et al. (2008). Initiatives for developing and comparing genotype interpretation systems: external validation of existing systems for didanosine against virological response. *Journal of Infectious Diseases, 198*(4), 470-480.

Averill, R. F., McCullough, E. C., Hughes, J. S., Goldfield, N. I., Vertrees, J. C., & Fuller, R. L. (2009). Redesigning the Medicare inpatient PPS to reduce payments to hospitals with high readmission rates. *Health Care Financing Review, 30*(4), 1-15.

Bembom, O., Fessel, J., Shafer, R. W., & van der Laan, M. J. (2008). Data-adaptive selection of the adjustment set in variable importance estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, Working Paper 231.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees.* Chapman & Hall.

Cabrera, C., Cozzi-Lepri, A., Phillips, A. N., Loveday, C., Kirk, O., Ait-Khaled, M., et al. (2004). Baseline resistance and virological outcome in patients with virological failure who start a regimen containing abacavir: EuroSIDA study. *Antiviral Therapy, 9*(5), 787-800.

Clennon, J. A., King, C. H., Muchiri, E. M., Kariuki, H. C., Ouma, J. H., Mungai, P., et al. (2004). Spatial patterns of urinary schistosomiasis infection in a highly endemic area of coastal Kenya. *American Journal of Tropical Medicine and Hygiene, 70*(4), 443-448.

Cole, S. R., & Hernan, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology, 168*(6), 656-664.

R version 2.10.0. (Copyright (C) 2009). The R Foundation for Statistical Computing.

De Luca, A., Cingolani, A., Di Giambenedetto, S., Trotta, M. P., Baldini, F., Rizzo, M. G., et al. (2003). Variable prediction of antiretroviral treatment outcome by different systems for interpreting genotypic human immunodeficiency virus type 1 drug resistance. *Journal of Infectious Diseases, 187*(12), 1934-1943.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Re-sampling Plans* (Vols. CBMS-NSF Regional Conference Series in Applied Mathematics 38). Montpelier: Capital City Press.

Eshleman, S. H., Hackett, J. J., Swanson, P., Cunningham, S. P., Drews, B., Brennan, C., et al. (2004). Performance of the Celera Diagnostics ViroSeq HIV-1 Genotyping System for sequence-based analysis of diverse human immunodeficiency virus type 1 strains. *Journal of Clinical Microbiology, 42*(6), 2711-2717.

Fleischer, N. L., Fernald, L. C., & Hubbard, A. E. (2007). Depressive symptoms in low-income women in rural Mexico. *Epidemiology, 18*(6), 678-685.

Fox, Z. V., Geretti, A. M., Kjaer, J., Dragsted, U. B., Phillips, A. N., Gerstoft, J., et al. (2007). The ability of four genotypic interpretation systems to predict virological response to ritonavir-boosted protease inhibitors. *AIDS, 21*(15), 2033-2042.

Frentz, D., Boucher, C. A., Assel, M., De Luca, A., Fabbiani, M., Incardona, F., et al. (2010). Comparison of HIV-1 genotypic resistance test interpretation systems in predicting virological outcomes over time. *PLoS One, 5*(7), e11505.

Greenland, S., & Drescher, K. (1993). Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics, 49*(3), 865-872.

Helm, M., Walter, H., Ehret, R., Schmit, J. C., Kurowski, M., Knechten, H., et al. (2007). Differences of nine drug resistance interpretation systems in predicting short-term therapy outcomes of treatment-experienced HIV-1 infected patients: a retrospective observational cohort study. *European Journal of Medical Research, 12*(6), 231-242.

Hirsch, M. S., Brun-Vézinet, F., Clotet, B., Conway, B., Kuritzkes, D. R., D'Aquila, R. T., et al. (2003). Antiretroviral drug resistance testing in adults infected with human immunodeficiency virus type 1: 2003 recommendations of an International AIDS Society-USA Panel. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America, 37*(1), 113-128.

Hubbard, A. E., & van der Laan, M. J. (2008). Population intervention models in causal inference. *Biometrika, 95*(1), 35-47.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., et al. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology, 21*(4), 467-74.

Hubbard, A. E., Jewell, N. P., & van der Laan, M. J. (2011). Direct effects and the effect among the treated. In M. J. van der Laan, & S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data* (pp. 133-143). Berlin, Heidelberg, New York: Springer.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1*, pp. 221-223. University of California Press.

Hung, Y. W., & Remais, J. (2008). Quantitative detection of Schistosoma japonicum cercariae in water by real-time PCR. *PLoS Neglected Tropical Diseases, 2*(11), e337.

Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine, 360*(14), 1418-1428.

Katz, N., Chaves, A., & Pellegrino, J. (1972). A simple device for quantitative stool thick-smear technique in Schistosomiasis mansoni. *Revista do Instituto de Medicina Tropical de São Paulo, 14*(6), 397-400.

King, C. H., Dickman, K., & Tisch, D. J. (2005). Reassessment of the cost of chronic helminthic infection: a meta-analysis of disability-related outcomes in endemic Schistosomiasis. *The Lancet, 365*(9470), 1561-9.

Kooperberg, C., Bose, S., & Stone, C. J. (1997). Polychotomous regression. *Journal of the American Statistical Association, 92*(437), 117-127.

Krumholz, H., Parent, E., Tu, N., Vaccarino, V., Wang, Y., Radford, M., et al. (1997). Readmission after hospitalization for congestive heart failure among Medicare beneficiaries. *Archives of Internal Medicine, 157*(1), 99-104.

Leenstra, T., Coutinho, H. M., Acosta, L. P., Langdon, G. C., Su, L., Olveda, R. M., et al. (2006, November). Schistosoma japonicum Reinfection after Praziquantel Treatment Causes Anemia Associated with Inflammation. *Infection and Immunity, 74*(11), 6398-6407.

Li, Y. S., Sleigh, A. C., Ross, A. G., Williams, G. M., Tanner, M., & McManus, D. P. (2000). Epidemiology of Schistosoma japonicum in China: morbidity and strategies for control in the Dongting Lake region. *International Journal of Parasitology, 30*(3), 273-281.

Li, Y., Sleigh, A. C., Williams, G. M., Ross, A. G., Li, Y., Forsyth, S. J., et al. (2000). Measuring exposure to Schistosoma japonicum in China. III. Activity diaries, snail and human infection, transmission ecology and options for control. *Acta Tropica, 75*(3), 279-289.

Liang, S., Seto, E. Y., Remais, J. V., Zhong, B., Yang, C., Hubbard, A., et al. (2007). Environmental effects on parasitic disease transmission exemplified by schistosomiasis in western China. *Proceedings of the National Academy of Sciences of the United States of America, 104*(17), 7110-7115.

Liu, T. F., & Shafer, R. W. (2006). Web resources for HIV type 1 genotypic-resistance test interpretation. *Clinical Infectious Diseases, 42*(11), 1608-18.

Matthys, B., Tschannen, A. B., Tian-Bi, N. T., Comoe, H., Diabate, S., Traore, M., et al. (2007). Risk factors for Schistosoma mansoni and hookworm in urban farming communities in western Cote d'Ivoire. *Tropical Medicine & International Health, 12*(6), 709-723.

Medicare Payment Advisory Commission. (2008). A path to bundled payment around a rehospitalization. *Report to the Congress: reforming the delivery system*, (pp. 83-103). Washington, DC.

Messer, L. C., Oakes, J. M., & Mason, S. (2010). Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *American Journal of Epidemiology, 171*(6), 664-673.

Messer, L. C., Oakes, J. M., & Mason, S. (2010). Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *American Journal of Epidemiology, 171*(6), 664-673.

Mortimer, K. M., Neugebauer, R., van der Laan, M., & Tager, I. B. (2005). An application of model-fitting procedures for marginal structural models. *American Journal of Epidemiology, 162*(4), 382-388.

Mosteller, R. D. (1987). Simplified calculation of body-surface area. *New England Journal of Medicine, 317*(17), 1098.

Nojima, H., Santos, A., Blas, B., & Kamiya, H. (1980). The emergence of Schistosoma japonicum cercariae from Oncomelania quadrasi. *Journal of Parasitology, 66*(6), 1010-1013.

Ormaasen, V., Sandvik, L., Asjø, B., Holberg-Petersen, M., Gaarder, P. I., & Bruun, J. N. (2004). An algorithm-based genotypic resistance score is associated with clinical outcome in HIV-1-infected adults on antiretroviral therapy. *HIV Medicine, 5*(6), 400-406.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge, U.K.: Cambridge University Press.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & van der Laan, M. J. (2012, February). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research, 21*(1), 31-54.

Philbin, E. F., & DiSalvo, T. G. (1999). Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology, 33*(6), 1560-1566.

Rau, J. (2012, July 19). Hospitals' readmissions rates not budging. Retrieved from http://www.kaiserhealthnews.org/Stories/2012/July/20/hospital-readmissions-rates-still-high.aspx

Remais, J., Hubbard, A., Wu, Z. S., & Spear, R. C. (2007). Weather-driven dynamics of an intermediate host: mechanistic and statistical population modelling of Oncomelania hupensis. *Journal of Applied Ecology, 44*(4), 781-791.

Remais, J., Liang, S., & Spear, R. C. (2008). Coupling hydrologic and infectious disease models to explain regional differences in schistosomiasis transmission in southwestern China. *Environmental Science & Technology, 42*(7), 2643-2649.

Revell, A. D., Wang, D., Boyd, M. A., Emery, S., Pozniak, A. L., De Wolf, F., et al. (2011). The development of an expert system to predict virological response to HIV therapy as part of an online treatment support tool. *AIDS, 25*(15), 1855-1863.

Rhee, S.-Y., Fessel, W. J., Liu, T. F., Marlowe, N. M., Rowland, C. M., Rode, R. A., et al. (2009). Predictive value of HIV-1 genotypic resistance test interpretation algorithms. *Journal of Infectious Diseases, 200*(3), 453-463.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling, 7*, 1393-1512.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran, & D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment and Clinical Trials* (pp. 95-113). New York: Springer.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran, & D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (pp. 95-113). New York: Springer.

Robins, J. M., & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine, 16*, 285-319.

Ross, A. G., Bartley, P. B., Sleigh, A. C., Olds, G. R., Li, Y., Williams, G. M., et al. (2002). Schistosomiasis. *New England Journal of Medicine, 346*, 1212-1220.

Ross, J. S., Mulvey, G. K., Stauffer, B., Patlolla, V., Bernheim, S. M., Keenan, P. S., et al. (2008). Statistical models and patient predictors of readmission for heart failure: a systematic review. *Archives of Internal Medicine, 168*(13), 1371-1386.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics, 6*, 34-58.

Rubin, D. B. (1986). Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers. *Journal of the American Statistical Association, 81*(396), 961-962.

Rudge, J. W., Stothard, J. R., Basanez, M.-G., Mgeni, A. F., Khamis, I. S., Khamis, A. N., et al. (2008). Micro-epidemiology of urinary schistosomiasis in Zanzibar: Local risk factors associated with distribution of infections among schoolchildren and relevance for control. *Acta Tropica, 105*(1), 45-54.

Schmidt, B., Walter, H., Zeitler, N., & Korn, K. (2002). Genotypic drug resistance interpretation systems--the cutting edge of antiretroviral therapy. *AIDS Reviews, 4*(3), 148-156.

Seto, E. Y., Lee, Y. J., Liang, S., & Zhong, B. (2007). Individual and village-level study of water contact patterns and Schistosoma japonicum infection in mountainous rural China. *Tropical Medicine & International Health, 12*(10), 1199-1209.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005, October). ROCR: visualizing classifier performance in R. *Bioinformatics, 21*(20), 3940-3941.

Sinisi, S. E., & 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), 1-38.

Sinisi, S. E., Polley, E. C., Petersen, M., Rhee, S.-Y., & van der Laan, M. J. (2007). Super learning: an application to the prediction of HIV-1 drug resistance. *Statistical Applications in Genetics and Molecular Biology, 6*(1), 1-24.

Sinisi, S., Polley, E. C., Petersen, M. L., Rhee, S.-Y., & van der Laan, M. J. (2007). Super learning: an application to the prediction of HIV-1 drug resistance. *Statistical Applications in Genetics and Molecular Biology, 6*(1), Article 7.

Spear, R. C., Seto, E., Liang, S., Birkner, M., Hubbard, A., Qiu, D., et al. (2004). Factors influencing the transmission of Schistosoma japonicum in the mountains of Sichuan Province of China. *American Journal of Tropical Medicine & Hygiene, 70*(1), 48-56.

Spear, R. C., Zhong, B., Mao, Y., Hubbard, A., Birkner, M., Remais, J., et al. (2004). Spatial and temporal variability in schistosome cercarial density detected by mouse bioassays in village irrigation ditches in Sichuan, China. *American Journal of Tropical Medicine & Hygiene, 71*(5), 554-557.

Stata 10. (n.d.). College Station, TX: StataCorp LP.

Sudat, S. E., Carlton, E. J., Seto, E. Y., Spear, R. C., & Hubbard, A. E. (2010). Using variable importance measures from causal inference to rank risk factors of schistosomiasis infection in a rural setting in China. *Epidemiologic Perspectives & Innovations, 7*(3).

The Office of Endemic Disease Control MoH. (2000). *Handbook of Schistosomiasis Control.* Shanghai: Shanghai Science & Technology Press.

Therneau, T. M., & Atkinson, E. J. (1997). *An Introduction to Recursive Partitioning Using the RPART Routines.* Technical Report 61, Mayo Clinic, Section of Statistics.

van der Laan, M. J. (2010). Estimation of causal effects of community based interventions. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 268.

van der Laan, M. J., & Gruber, S. (2009). Collaborative double robust targeted penalized maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 246.

van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality.* New York: Springer.

van der Laan, M. J., & Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data.* Berlin, Heidelberg, New York: Springer.
van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics, 2*(1), 1-38.
van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics, 2*(1), 1-38.
van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology, 6*(1), Article 25.
Van Laethem, K., De Luca, A., Antinori, A., Cingolani, A., & Vandamme, A. M. (2002). A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antiviral Therapy, 7*(2), 123-129.
Vandamme, A. M., Camacho, R. J., Ceccherini-Silberstein, F., de Luca, A., Palmisano, L., Paraskevis, D., et al. (2011). European recommendations for the clinical use of HIV drug resistance testing: 2011 update. *AIDS Reviews, 13*(2), 77-108.
Wang, L., Porter, B., Maynard, C., Bryson, C., Sun, H., Lowy, E., et al. (2012). Predicting risk of hospitalization or death among patients with heart failure in the veterans health administration. *American Journal of Cardiology,* Epub ahead of print.
Wang, Y., Petersen, M., Bangsberg, D., & van der Laan, M. J. (2006, September). Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. *U.C. Berkeley Division of Biostatistics Working Paper Series,* Working Paper 211.
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48,* 817-830.
WHO. (2006). *Report of the Scientific Working Group Meeting on Schistosomiasis.* Geneva: WHO.
Zazzi, M., Prosperi, M., Vicenti, I., Di Giambenedetto, S., Callegaro, A., Bruzzone, B., et al. (2009). Rules-based HIV-1 genotypic resistance interpretation systems predict 8 week and 24 week virological antiretroviral treatment outcome and benefit from drug potency weighting. *Journal of Antimicrobial Chemotherapy, 64*(3), 616-624.
Zheng, W., & van der Laan, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series,* Paper 273.

Appendix

Table A4.1 - ICD-9-CM codes

Code Category	Codes
Heart Failure	402.01 Malignant, hypertensive heart disease with heart failure 402.11 Benign, hypertensive heart disease with heart failure 402.91 Unspecified, hypertensive heart disease with heart failure 404.01 Malignant hypertensive heart and kidney disease with heart failure 404.03 Malignant hypertensive heart and kidney disease with heart failure and chronic kidney disease 404.11 Benign, hypertensive heart and kidney disease with heart failure 404.13 Benign hypertensive heart and kidney disease with heart failure and chronic kidney disease 404.91 Unspecified, hypertensive heart and kidney disease with heart failure 404.93 Unspecified hypertensive heart and kidney disease with heart failure and chronic kidney disease 428.0 Congestive heart failure, unspecified 428.1 Left heart failure 428.20 Unspecified systolic heart failure 428.21 Acute systolic heart failure 428.22 Chronic systolic heart failure 428.23 Acute on chronic systolic heart failure 428.30 Unspecified diastolic heart failure 428.31 Acute diastolic heart failure 428.32 Chronic diastolic heart failure 428.33 Acute on chronic diastolic heart failure 428.40 Unspecified combined systolic and diastolic heart failure 428.41 Acute combined systolic and diastolic heart failure 428.42 Chronic combined systolic and diastolic heart failure 428.43 Acute on chronic combined systolic and diastolic heart failure 428.9 Heart failure, unspecified
Ischemic Heart Disease	4140 41400 41401 41406 4142 4143 4148 4149
Valvular Heart Disease	3940 3941 3942 3949 3950 3951 3952 3959 3960 3961 3962 3963 3968 3969 3970 3971 3979 4240 4241 4242 4243 42490 42491 42499 7852 7853 V422 V433
Diabetes Mellitus	24900 25000 25001 7902 79021 79022 79029 7915 7916 V4585 V5391 V6546 24901 24910 24911 24920 24921 24930 24931 24940 24941 24950 24951 24960 24961 24970 24971 24980 24981 24990 24991 25002 25003 25010 25011 25012 25013 25020 25021 25022 25023 25030 25031 25032 25033 25040 25041 25042 25043 25050 25051 25052 25053 25060 25061 25062 25063 25070 25071 25072 25073 25080 25081 25082 25083 25090 25091 25092 25093
Renal Disease	5810 5811 5812 5813 58181 58189 5819 5820 5821 5822 5824 58281 58289 5829 5830 5832 5834 5836 5837 58381 58389 5839 587 585 5853 5854 5855 5856 5859 7925
Chronic Lung Disease	4910 4911 4912 49120 49121 49122 4918 4919 4920 4928 494 4940 4941 496 49300 49301 49302 49310 49311 49312 49320 49321 49322 49381 49382 49390 49391 49392 4950 4951 4952 4953 4954 4955 4956 4957 4958 4959 500 501 502 503 504 505 5060 5061 5062 5063 5064 5069 5071 5078 5080 5081 5088 5089
Cardiac Surgery	35.10 OPEN VALVULOPLASTY NOS 35.11 OPN AORTIC VALVULOPLASTY 35.12 OPN MITRAL VALVULOPLASTY 35.13 OPN PULMON VALVULOPLASTY 35.14 OPN TRICUS VALVULOPLASTY 35.20 REPLACE HEART VALVE NOS 35.21 REPLACE AORT VALV-TISSUE 35.22 REPLACE AORTIC VALVE NEC 35.23 REPLACE MITR VALV-TISSUE 35.24 REPLACE MITRAL VALVE NEC 35.25 REPLACE PULM VALV-TISSUE 35.26 REPLACE PULMON VALVE NEC 35.27 REPLACE TRIC VALV-TISSUE 35.28 REPLACE TRICUSP VALV NEC 35.31 PAPILLARY MUSCLE OPS 35.32 CHORDAE TENDINEAE OPS 35.33 ANNULOPLASTY 35.34 INFUNDIBULECTOMY 35.35 TRABECUL CARNEAE CORD OP 35.39 TISS ADJ TO VALV OPS NEC 35.42 CREATE SEPTAL DEFECT 35.50 PROSTH REP HRT SEPTA NOS

Table A4.1 - ICD-9-CM codes

Code Category	Codes
Cardiac Surgery	35.51 PROS REP ATRIAL DEF-OPN
	35.53 PROS REP VENTRIC DEF-OPN
	35.54 PROS REP ENDOCAR CUSHION
	35.60 GRFT REPAIR HRT SEPT NOS
	35.61 GRAFT REPAIR ATRIAL DEF
	35.62 GRAFT REPAIR VENTRIC DEF
	35.63 GRFT REP ENDOCAR CUSHION
	35.70 HEART SEPTA REPAIR NOS
	35.71 ATRIA SEPTA DEF REP NEC
	35.72 VENTR SEPTA DEF REP NEC
	35.73 ENDOCAR CUSHION REP NEC
	35.81 TOT REPAIR TETRAL FALLOT
	35.82 TOTAL REPAIR OF TAPVC
	35.83 TOT REP TRUNCUS ARTERIOS
	35.84 TOT COR TRANSPOS GRT VES
	35.91 INTERAT VEN RETRN TRANSP
	35.92 CONDUIT RT VENT-PUL ART
	35.93 CONDUIT LEFT VENTR-AORTA
	35.94 CONDUIT ARTIUM-PULM ART
	35.98 OTHER HEART SEPTA OPS
	35.99 OTHER HEART VALVE OPS
	36.03 OPEN CORONRY ANGIOPLASTY
	36.10 AORTOCORONARY BYPASS NOS
	36.11 AORTOCOR BYPAS-1 COR ART
	36.12 AORTOCOR BYPAS-2 COR ART
	36.13 AORTOCOR BYPAS-3 COR ART
	36.14 AORTCOR BYPAS-4+ COR ART
	36.15 1 INT MAM-COR ART BYPASS
	36.16 2 INT MAM-COR ART BYPASS
	36.17 ABD-CORON ARTERY BYPASS
	36.19 HRT REVAS BYPS ANAS NEC
	36.31 OPEN CHEST TRANS REVASC
	36.91 CORON VESS ANEURYSM REP
	36.99 HEART VESSEL OP NEC
	37.10 INCISION OF HEART NOS
	37.11 CARDIOTOMY
	37.32 HEART ANEURYSM EXCISION
	37.33 EXC/DEST HRT LESION OPEN
	37.35 PARTIAL VENTRICULECTOMY
	37.36 EXC LEFT ATRIAL APPENDAG
	37.41 IMPL CARDIAC SUPPORT DEV
	37.49 HEART/PERICARD REPR NEC
	37.51 HEART TRANSPLANTATION
	37.52 IMP TOT INT BI HT RP SYS
	37.53 REPL/REP THR UNT TOT HRT
	37.54 REPL/REP OTH TOT HRT SYS
	37.55 REM INT BIVENT HRT SYS
	37.60 IMP BIVN EXT HRT AST SYS
	37.62 INSRT NON-IMPL CIRC DEV
	37.63 REPAIR HEART ASSIST SYS
	37.64 REMVE EXT HRT ASSIST SYS
	37.66 IMPLANTABLE HRT ASSIST
	37.67 IMP CARDIOMYOSTIMUL SYS
Dialysis	39.95 HEMODIALYSIS
	54.98 PERITONEAL DIALYSIS

Table A4.2 - Variable importance estimates for all explanatory variables, 30-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals that do not cross zero.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
All Cause Readmission										
African American	0.072	0.032	(0.009, 0.135)	0.136	52.7%	0.014	0.038	(-0.061, 0.089)	0.202	7.0%
Medicare	0.030	0.041	(-0.049, 0.11)	0.160	18.8%	-0.029	0.066	(-0.158, 0.1)	0.248	-11.7%
Medicaid	0.007	0.035	(-0.062, 0.076)	0.200	3.5%	-0.013	0.033	(-0.077, 0.052)	0.194	-6.7%
Idiopathic Cardiomyopathy	0.069	0.038	(-0.005, 0.143)	0.196	35.1%	0.029	0.041	(-0.052, 0.109)	0.197	14.5%
Ischemic Heart Disease	0.004	0.030	(-0.055, 0.062)	0.182	2.1%	-0.043	0.044	(-0.129, 0.043)	0.219	-19.7%
Valvular Heart Disease	-0.107	0.033	(-0.171, -0.043)	0.213	-50.2%	-0.075	0.039	(-0.152, 0.001)	0.229	-32.9%
Diabetes Mellitus	0.016	0.039	(-0.061, 0.092)	0.213	7.4%	-0.045	0.039	(-0.122, 0.032)	0.238	-18.8%
Renal Disease	0.021	0.039	(-0.056, 0.098)	0.207	10.0%	0.031	0.038	(-0.044, 0.105)	0.186	16.5%
Chronic Lung Disease	0.026	0.034	(-0.04, 0.092)	0.193	13.6%	0.108	0.039	(0.031, 0.184)	0.168	64.2%
Telemetry during hospitalization	-0.082	0.063	(-0.206, 0.041)	0.272	-30.3%	0.079	0.042	(-0.003, 0.161)	0.134	59.1%
Discharged to home health	-0.026	0.029	(-0.082, 0.03)	0.188	-13.7%	-0.059	0.042	(-0.141, 0.023)	0.221	-26.8%
Inpatient hospitalization in past year	0.130	0.029	(0.074, 0.185)	0.129	100.7%	0.092	0.037	(0.02, 0.163)	0.163	56.3%
Facility	0.021	0.031	(-0.04, 0.081)	0.205	10.2%	0.012	0.034	(-0.055, 0.08)	0.208	6.0%
Weekend hospital admission	-0.002	0.037	(-0.075, 0.071)	0.204	-0.9%	0.030	0.044	(-0.056, 0.117)	0.210	14.4%
Weekend hospital discharge	0.010	0.034	(-0.056, 0.077)	0.198	5.3%	-0.001	0.038	(-0.075, 0.073)	0.219	-0.6%
Number of diagnoses >13	-0.007	0.039	(-0.083, 0.068)	0.196	-3.8%	0.055	0.039	(-0.021, 0.131)	0.176	31.1%
Hospital length of stay (days) >4	-0.013	0.034	(-0.079, 0.054)	0.209	-6.1%	-0.020	0.036	(-0.089, 0.05)	0.214	-9.2%
Age at hospital admission >69	-0.091	0.046	(-0.181, -0.002)	0.214	-42.6%	-0.020	0.062	(-0.143, 0.102)	0.194	-10.5%
Readmission risk score >5	0.095	0.035	(0.025, 0.164)	0.185	51.1%	0.015	0.037	(-0.058, 0.089)	0.214	7.2%
Heart Failure Readmission										
African American	0.044	0.028	(-0.011, 0.098)	0.079	55.0%	-0.012	0.028	(-0.066, 0.043)	0.106	-10.9%
Medicare	0.016	0.038	(-0.058, 0.091)	0.105	15.5%	-0.016	0.025	(-0.064, 0.032)	0.082	-19.2%
Medicaid	-0.025	0.027	(-0.077, 0.027)	0.122	-20.1%	-0.018	0.026	(-0.069, 0.033)	0.106	-16.8%
Idiopathic Cardiomyopathy	0.070	0.039	(-0.006, 0.146)	0.101	69.1%	-0.002	0.024	(-0.05, 0.045)	0.090	-2.5%
Ischemic Heart Disease	0.010	0.021	(-0.03, 0.051)	0.092	11.0%	-0.049	0.027	(-0.102, 0.003)	0.104	-47.6%
Valvular Heart Disease	-0.038	0.027	(-0.09, 0.015)	0.111	-33.9%	-0.007	0.028	(-0.061, 0.047)	0.099	-7.0%
Diabetes Mellitus	0.006	0.033	(-0.058, 0.071)	0.115	5.5%	0.010	0.032	(-0.053, 0.073)	0.098	10.7%
Renal Disease	0.036	0.030	(-0.022, 0.094)	0.101	35.9%	0.038	0.031	(-0.023, 0.098)	0.076	49.1%
Chronic Lung Disease	0.036	0.026	(-0.015, 0.088)	0.099	36.8%	0.019	0.024	(-0.028, 0.066)	0.079	23.6%

Table A4.2 - Variable importance estimates for all explanatory variables, 30-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals that do not cross zero.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
Telemetry during hospitalization	-0.011	0.039	(-0.088, 0.065)	0.119	-9.7%	0.073	0.020	(0.033, 0.112)	0.028	260.1%
Discharged to home health	-0.026	0.021	(-0.067, 0.015)	0.106	-24.5%	-0.087	0.022	(-0.13, -0.044)	0.115	-76.2%
Inpatient hospitalization in past year	0.114	0.022	(0.072, 0.157)	0.047	242.9%	0.057	0.024	(0.01, 0.104)	0.062	92.2%
Facility	0.029	0.025	(-0.02, 0.077)	0.106	26.9%	0.015	0.026	(-0.037, 0.067)	0.094	16.2%
Weekend hospital admission	-0.023	0.027	(-0.076, 0.03)	0.116	-19.8%	0.022	0.032	(-0.042, 0.086)	0.093	23.7%
Weekend hospital discharge	0.049	0.029	(-0.008, 0.106)	0.100	48.5%	-0.058	0.022	(-0.101, -0.015)	0.109	-53.0%
Number of diagnoses >13	-0.034	0.024	(-0.081, 0.014)	0.110	-30.7%	-0.017	0.023	(-0.062, 0.028)	0.094	-17.8%
Hospital length of stay (days) >4	-0.004	0.029	(-0.061, 0.052)	0.121	-3.7%	-0.037	0.023	(-0.081, 0.008)	0.101	-36.6%
Age at hospital admission >69	-0.090	0.041	(-0.171, -0.01)	0.140	-64.4%	-0.039	0.024	(-0.087, 0.008)	0.093	-42.3%
Readmission risk score >5	0.041	0.027	(-0.012, 0.094)	0.104	39.3%	0.036	0.026	(-0.014, 0.087)	0.080	45.3%

Table A4.3 - Variable importance estimates for all explanatory variables, 90-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals that do not cross zero.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
All Cause Readmission										
African American	0.114	0.036	(0.044, 0.185)	0.258	44.4%	0.098	0.048	(0.004, 0.192)	0.335	29.3%
Medicare	-0.012	0.046	(-0.103, 0.078)	0.325	-3.8%	-0.103	0.079	(-0.259, 0.053)	0.532	-19.4%
Medicaid	0.024	0.039	(-0.053, 0.101)	0.348	6.8%	0.042	0.045	(-0.048, 0.131)	0.355	11.7%
Idiopathic Cardiomyopathy	0.019	0.041	(-0.062, 0.1)	0.371	5.2%	0.014	0.048	(-0.08, 0.107)	0.367	3.7%
Ischemic Heart Disease	-0.040	0.035	(-0.108, 0.028)	0.362	-11.1%	-0.023	0.052	(-0.126, 0.079)	0.379	-6.2%
Valvular Heart Disease	-0.068	0.050	(-0.166, 0.03)	0.376	-18.1%	0.037	0.049	(-0.058, 0.132)	0.401	9.2%
Diabetes Mellitus	0.058	0.040	(-0.02, 0.136)	0.353	16.5%	-0.027	0.046	(-0.118, 0.063)	0.413	-6.6%
Renal Disease	0.010	0.044	(-0.075, 0.096)	0.363	2.8%	0.026	0.049	(-0.07, 0.121)	0.385	6.7%
Chronic Lung Disease	0.091	0.042	(0.008, 0.174)	0.350	25.9%	0.122	0.047	(0.03, 0.213)	0.362	33.6%
Telemetry during hospitalization	-0.046	0.069	(-0.181, 0.09)	0.400	-11.4%	0.138	0.050	(0.04, 0.237)	0.254	54.3%
Discharged to home health	-0.009	0.038	(-0.083, 0.065)	0.360	-2.5%	-0.052	0.051	(-0.153, 0.049)	0.375	-13.8%
Inpatient hospitalization in past year	0.266	0.034	(0.199, 0.333)	0.220	121.0%	0.166	0.043	(0.082, 0.251)	0.295	56.3%
Facility	0.035	0.033	(-0.029, 0.1)	0.359	9.9%	0.037	0.041	(-0.043, 0.116)	0.376	9.7%
Weekend hospital admission	0.031	0.043	(-0.053, 0.114)	0.359	8.5%	-0.029	0.048	(-0.122, 0.065)	0.394	-7.3%

Table A4.3 - Variable importance estimates for all explanatory variables, 90-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals that do not cross zero.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
Weekend hospital discharge	0.055	0.039	(-0.021, 0.13)	0.355	15.4%	0.031	0.044	(-0.055, 0.117)	0.386	8.1%
Number of diagnoses >13	0.022	0.061	(-0.098, 0.142)	0.364	6.1%	0.037	0.050	(-0.062, 0.135)	0.358	10.2%
Hospital length of stay (days) >4	0.013	0.038	(-0.061, 0.086)	0.373	3.4%	-0.014	0.046	(-0.103, 0.076)	0.406	-3.4%
Age at hospital admission >69	-0.076	0.046	(-0.166, 0.014)	0.341	-22.3%	0.090	0.074	(-0.056, 0.235)	0.369	24.3%
Readmission risk score >5	0.123	0.037	(0.05, 0.197)	0.332	37.1%	0.091	0.044	(0.005, 0.177)	0.356	25.5%
Heart Failure Readmission										
African American	0.104	0.033	(0.041, 0.168)	0.134	77.9%	0.032	0.040	(-0.046, 0.111)	0.174	18.5%
Medicare	0.015	0.040	(-0.062, 0.093)	0.179	8.6%	-0.048	0.060	(-0.166, 0.069)	0.226	-21.4%
Medicaid	0.008	0.031	(-0.054, 0.07)	0.198	4.0%	0.011	0.036	(-0.059, 0.081)	0.188	5.7%
Idiopathic Cardiomyopathy	0.038	0.040	(-0.04, 0.115)	0.203	18.5%	0.007	0.036	(-0.063, 0.077)	0.175	3.8%
Ischemic Heart Disease	0.047	0.029	(-0.009, 0.104)	0.167	28.5%	-0.059	0.040	(-0.136, 0.019)	0.201	-29.2%
Valvular Heart Disease	0.023	0.043	(-0.062, 0.108)	0.202	11.4%	0.077	0.047	(-0.015, 0.168)	0.190	40.4%
Diabetes Mellitus	0.035	0.038	(-0.04, 0.11)	0.195	18.0%	-0.026	0.038	(-0.1, 0.048)	0.209	-12.5%
Renal Disease	0.008	0.036	(-0.062, 0.078)	0.193	4.0%	0.032	0.041	(-0.048, 0.113)	0.183	17.7%
Chronic Lung Disease	0.042	0.036	(-0.028, 0.112)	0.203	20.7%	0.082	0.037	(0.008, 0.155)	0.161	50.6%
Telemetry during hospitalization	-0.054	0.059	(-0.169, 0.061)	0.255	-21.1%	0.086	0.038	(0.012, 0.16)	0.112	76.4%
Discharged to home health	0.004	0.032	(-0.058, 0.067)	0.200	2.2%	-0.063	0.034	(-0.13, 0.003)	0.197	-32.0%
Inpatient hospitalization in past year	0.211	0.029	(0.154, 0.268)	0.095	221.9%	0.064	0.035	(-0.005, 0.133)	0.165	38.7%
Facility	0.043	0.029	(-0.014, 0.101)	0.197	21.9%	0.006	0.033	(-0.059, 0.071)	0.194	3.1%
Weekend hospital admission	0.013	0.035	(-0.056, 0.082)	0.207	6.2%	-0.005	0.039	(-0.082, 0.072)	0.192	-2.4%
Weekend hospital discharge	0.069	0.035	(0, 0.138)	0.195	35.5%	0.016	0.036	(-0.054, 0.086)	0.187	8.4%
Number of diagnoses >13	-0.056	0.033	(-0.12, 0.008)	0.206	-27.3%	-0.041	0.038	(-0.116, 0.034)	0.200	-20.4%
Hospital length of stay (days) >4	0.025	0.032	(-0.037, 0.088)	0.212	12.0%	-0.066	0.034	(-0.131, 0)	0.216	-30.4%
Age at hospital admission >69	-0.123	0.043	(-0.207, -0.039)	0.241	-51.1%	-0.119	0.042	(-0.201, -0.036)	0.222	-53.5%
Readmission risk score >5	0.087	0.032	(0.025, 0.149)	0.188	46.1%	0.103	0.037	(0.031, 0.174)	0.156	66.0%

Table A4.4 - Variable importance estimates for all explanatory variables, 180-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals that do not cross zero.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
All Cause Readmission										
African American	0.108	0.039	(0.032, 0.183)	0.354	30.4%	0.095	0.047	(0.003, 0.188)	0.463	20.6%
Medicare	0.014	0.047	(-0.079, 0.106)	0.391	3.5%	-0.093	0.078	(-0.247, 0.061)	0.619	-15.0%
Medicaid	0.055	0.042	(-0.027, 0.137)	0.416	13.2%	0.030	0.045	(-0.058, 0.118)	0.479	6.3%
Idiopathic Cardiomyopathy	0.006	0.042	(-0.076, 0.089)	0.459	1.4%	-0.016	0.048	(-0.11, 0.079)	0.493	-3.2%
Ischemic Heart Disease	-0.043	0.038	(-0.118, 0.031)	0.448	-9.7%	-0.094	0.051	(-0.193, 0.006)	0.526	-17.8%
Valvular Heart Disease	-0.106	0.051	(-0.206, -0.006)	0.473	-22.4%	-0.007	0.049	(-0.102, 0.088)	0.514	-1.3%
Diabetes Mellitus	0.049	0.040	(-0.029, 0.127)	0.462	10.7%	-0.011	0.045	(-0.099, 0.078)	0.514	-2.1%
Renal Disease	0.017	0.045	(-0.071, 0.106)	0.457	3.8%	0.025	0.047	(-0.067, 0.117)	0.495	5.0%
Chronic Lung Disease	0.074	0.043	(-0.009, 0.158)	0.446	16.6%	0.109	0.045	(0.021, 0.197)	0.480	22.7%
Telemetry during hospitalization	-0.015	0.068	(-0.147, 0.118)	0.469	-3.2%	0.208	0.053	(0.105, 0.311)	0.307	67.7%
Discharged to home health	-0.052	0.040	(-0.129, 0.026)	0.462	-11.2%	-0.065	0.056	(-0.175, 0.045)	0.489	-13.3%
Inpatient hospitalization in past year	0.305	0.036	(0.234, 0.375)	0.290	105.0%	0.261	0.044	(0.175, 0.348)	0.359	72.7%
Facility	0.031	0.035	(-0.038, 0.099)	0.456	6.7%	0.049	0.041	(-0.032, 0.129)	0.486	10.0%
Weekend hospital admission	0.036	0.046	(-0.055, 0.126)	0.453	7.8%	-0.041	0.049	(-0.138, 0.055)	0.509	-8.2%
Weekend hospital discharge	0.038	0.040	(-0.04, 0.117)	0.455	8.4%	0.009	0.045	(-0.079, 0.097)	0.506	1.8%
Number of diagnoses >13	0.041	0.066	(-0.088, 0.169)	0.453	9.0%	0.036	0.054	(-0.069, 0.142)	0.473	7.7%
Hospital length of stay (days) >4	0.008	0.039	(-0.067, 0.084)	0.466	1.8%	-0.008	0.046	(-0.098, 0.082)	0.522	-1.5%
Age at hospital admission >69	-0.103	0.057	(-0.216, 0.01)	0.456	-22.6%	0.043	0.070	(-0.094, 0.18)	0.554	7.7%
Readmission risk score >5	0.088	0.039	(0.012, 0.165)	0.443	20.0%	0.107	0.044	(0.02, 0.193)	0.465	22.9%
Heart Failure Readmission										
African American	0.114	0.036	(0.043, 0.185)	0.187	61.0%	0.047	0.043	(-0.038, 0.131)	0.240	19.4%
Medicare	0.008	0.044	(-0.078, 0.094)	0.244	3.2%	-0.034	0.062	(-0.155, 0.086)	0.275	-12.5%
Medicaid	0.045	0.036	(-0.026, 0.116)	0.245	18.3%	0.026	0.038	(-0.048, 0.099)	0.252	10.1%
Idiopathic Cardiomyopathy	0.050	0.041	(-0.031, 0.131)	0.262	19.0%	0.008	0.041	(-0.073, 0.089)	0.257	2.9%
Ischemic Heart Disease	0.015	0.032	(-0.048, 0.077)	0.232	6.3%	-0.097	0.041	(-0.178, -0.017)	0.299	-32.5%
Valvular Heart Disease	-0.018	0.046	(-0.107, 0.072)	0.270	-6.5%	0.057	0.049	(-0.039, 0.154)	0.260	22.0%
Diabetes Mellitus	0.040	0.040	(-0.039, 0.118)	0.266	14.9%	-0.028	0.041	(-0.108, 0.053)	0.283	-9.7%
Renal Disease	0.012	0.040	(-0.067, 0.09)	0.258	4.5%	0.039	0.046	(-0.05, 0.129)	0.261	15.0%
Chronic Lung Disease	0.014	0.038	(-0.061, 0.088)	0.275	4.9%	0.107	0.042	(0.025, 0.188)	0.234	45.5%

Table A4.4 - Variable importance estimates for all explanatory variables, 180-day readmission outcomes. Estimates highlighted in gray have 95% confidence intervals that do not cross zero.

	No HF Intervention					*HF Intervention*				
	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change	Estimate	SE	95% CI	$\hat{E}[Y_0]$	Estimate as % change
Telemetry during hospitalization	0.010	0.060	(-0.109, 0.128)	0.264	3.7%	0.146	0.041	(0.066, 0.226)	0.140	104.5%
Discharged to home health	-0.035	0.034	(-0.102, 0.032)	0.263	-13.2%	-0.093	0.040	(-0.171, -0.016)	0.281	-33.2%
Inpatient hospitalization in past year	0.261	0.032	(0.198, 0.323)	0.133	196.2%	0.157	0.039	(0.081, 0.233)	0.189	83.0%
Facility	0.034	0.033	(-0.03, 0.098)	0.267	12.6%	0.022	0.036	(-0.049, 0.093)	0.266	8.4%
Weekend hospital admission	0.035	0.039	(-0.041, 0.111)	0.267	13.1%	-0.012	0.042	(-0.094, 0.07)	0.268	-4.3%
Weekend hospital discharge	0.037	0.037	(-0.035, 0.11)	0.265	14.1%	0.004	0.040	(-0.073, 0.082)	0.269	1.6%
Number of diagnoses >13	-0.016	0.057	(-0.127, 0.095)	0.266	-6.1%	-0.009	0.042	(-0.092, 0.074)	0.264	-3.4%
Hospital length of stay (days) >4	0.015	0.036	(-0.057, 0.086)	0.285	5.1%	-0.086	0.037	(-0.158, -0.013)	0.305	-28.0%
Age at hospital admission >69	-0.154	0.067	(-0.286, -0.022)	0.330	-46.7%	-0.094	0.118	(-0.326, 0.138)	0.353	-26.7%
Readmission risk score >5	0.102	0.036	(0.031, 0.172)	0.255	40.0%	0.104	0.040	(0.025, 0.183)	0.234	44.3%