The International Journal of Biostatistics

Volume 4, *Issue* 1 2008 *Article* 18

Testing for Associations with Missing High-Dimensional Categorical Covariates

Jennifer Schumi, *Statistics Collaborative, Inc.*
A. Gregory DiRienzo, *Harvard University*
Victor DeGruttola, *Harvard University*

Recommended Citation:
Schumi, Jennifer; DiRienzo, A. Gregory; and DeGruttola, Victor (2008) "Testing for Associations with Missing High-Dimensional Categorical Covariates," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 18.
DOI: 10.2202/1557-4679.1102

Testing for Associations with Missing High-Dimensional Categorical Covariates

Jennifer Schumi, A. Gregory DiRienzo, and Victor DeGruttola

Abstract

Understanding how long-term clinical outcomes relate to short-term response to therapy is an important topic of research with a variety of applications. In HIV, early measures of viral RNA levels are known to be a strong prognostic indicator of future viral load response. However, mutations observed in the high-dimensional viral genotype at an early time point may change this prognosis. Unfortunately, some subjects may not have a viral genetic sequence measured at the early time point, and the sequence may be missing for reasons related to the outcome. Complete-case analyses of missing data are generally biased when the assumption that data are missing completely at random is not met, and methods incorporating multiple imputation may not be well-suited for the analysis of high-dimensional data. We propose a semiparametric multiple testing approach to the problem of identifying associations between potentially missing high-dimensional covariates and response. Following the recent exposition by Tsiatis, unbiased nonparametric summary statistics are constructed by inversely weighting the complete cases according to the conditional probability of being observed, given data that is observed for each subject. Resulting summary statistics will be unbiased under the assumption of missing at random. We illustrate our approach through an application to data from a recent AIDS clinical trial, and demonstrate finite sample properties with simulations.

KEYWORDS: family-wise error rate, genotype, inverse probability weighting, missing at random, multiple testing, non-parametric bootstrap, simultaneous inference

Author Notes: This work was partially supported by the National Institute of Allergy and Infectious Diseases grants AI051164 and AI058836, and the National Institutes of Health grant AI007358. We thank the reviewers for helpful comments that improved the presentation of the paper.

1 Introduction

For many diseases, clinicians are interested in using short-term responses to therapy to predict longer-term outcomes. One such outcome is plasma HIV-1 RNA; it is well known [6] that values measured soon after initiation of treatment provide information for predicting longer-term response. Unless there are concerns about non-adherence to the drug regimen or development of antiretroviral resistance, HIV-infected individuals receiving potent antiretroviral treatment who have good initial responses to therapy tend to maintain viral load below levels of detection in the longer term. The development of viral genetic mutations associated with drug resistance at any time during therapy, however, may result in viral rebound, even among patients with a good initial response to therapy who are being maintained on the same regimen.

We can formulate this question statistically by testing whether viral genotype provides additional information about the expected viral load in the future, beyond that provided by the viral load measured at an earlier time point. To illustrate, we obtained data from the AIDS Clinical Trials Group (ACTG) study 398, a Phase II randomized trial of four salvage regimens for treatment-experienced subjects who had experienced a loss of virologic control [5]. In this study, HIV-1 RNA values were measured 8 and 24 weeks post-baseline, and HIV-1 genotyping was performed at week 8. For simplicity, we consider the presence or absence of a mutation at each codon as a binary variable, although our methods would allow consideration of different types of mutations at each codon as well. To do so, we would simply use an indicator function for each type of mutation at each codon.

The viral genotype is high dimensional and sparsely measured, and the entire sequence may be missing for various reasons. For example, in ACTG 398, some week 8 plasma samples were not genotyped; others did not have sufficient virus for amplification (in this study, the threshold was approximately 1000 HIV-1 RNA copies/mL). We propose a semiparametric approach for evaluating whether this high dimensional, potentially missing, genetic data provides further information about long-term response beyond that explained by early measures of viral load .

We construct a collection of multiple null hypotheses of the general form:

$$H^0(\boldsymbol{X}, \boldsymbol{Z}) : E(Y \mid \boldsymbol{X}, \boldsymbol{Z}) = E(Y \mid \boldsymbol{Z}), \tag{1}$$

where Y is the long-term response, for example, viral load at week 24, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the high-dimensional and low-dimensional short term responses, respectively. In this paper, $\boldsymbol{X}$ corresponds to the codons in the viral genotype, and $\boldsymbol{Z}$ to categorical viral load at week 8. The first conditional mean, $E(Y \mid$

$\boldsymbol{Z}$), thus represents the expected week 24 viral load for some level of the week 8 viral load, and can be estimated directly from the available data. The second conditional mean, $E(Y \mid \boldsymbol{X}, \boldsymbol{Z})$, corresponds to the expected week 24 viral load for a given level of the week 8 viral load and genetic sequence $\boldsymbol{X}$. Estimating this second term requires incorporation of the missing high-dimensional data via appropriately chosen weights, as discussed below.

For illustration, we apply the nonparametric bootstrap approach of van der Laan, Dudoit, and Pollard [15] to simultaneously test these null hypotheses with asymptotic control of the family-wise error rate (FWER). Other error rates may be used, for example the generalized family-wise error rate [16] or the tail proportion of false positives [14].

A complete case analysis (in this example, only those subjects with complete viral genotypes available at week 8) depends upon the strong assumption of missing-completely-at-random (MCAR) for the summary statistics to be unbiased. Briefly, the condition MCAR means that the probability that a subject's viral genotype is missing at week 8 does not depend on that subject's observed (here viral load at weeks 8 and 24) and unobserved data. Since one reason for the viral genotype to be missing in ACTG 398 was week 8 viral load below a certain threshold, this assumption is not likely to hold for our example.

Multiple imputation techniques require that one posit a model for the missing data given the data observed for each subject. Such model specification may be difficult with high-dimensional variables. Efron [4] discussed bootstrap alternatives to multiple imputation. The non-parametric bootstrap does not depend on the missing data mechanism, and is unbiased under the missing at random (MAR) assumption. The MAR condition holds when the probability that an observation is missing given the observed and unobserved data depends only on the observed data. The full mechanism bootstrap requires either that the data are MCAR or that there is a model for the conditional distribution of the missing data given the completely observed data. This presents similar problems as for multiple imputation techniques.

Our approach attempts to calculate unbiased summary statistics using the inverse-probability-weighting methods proposed by Robins and colleagues ([7], [8]), recently illustrated in Tsiatis [13], which relax the MCAR assumption to the assumption that the data are MAR. Robins and colleagues showed that estimators incorporating such weights are asymptotically normal and unbiased when three conditions were met: (a) the data are MAR following Rubin [9]; (b) the probability that an observation is completely observed is bounded away from 0; and (c) the missingness probabilities are either known (if observations are missing by design) or can be consistently estimated. We then use these weighted summary statistics to test the sequence of null hypotheses as defined above using existing methods for simultaneous hypothesis testing.

Section 2 provides further motivation and discussion of the ACTG 398 data. We present our model and notation in Section 3 and discuss hypothesis testing with known weights in Section 4. Inference when the weights are estimated is discussed in Section 5. Simulation studies investigating empirical power and error rate control are presented in Section 6, and the results of our application to the ACTG 398 data in Section 7. We conclude with a discussion.

2 ACTG 398 Data

ACTG 398 was a Phase II randomized trial comparing use of amprenavir as the sole protease inhibitor in a drug regimen or in combination with three other protease inhibitors for treatment-experienced patients with a loss of virologic control [5]. Subjects were categorized on the basis of prior exposure to non-nucleoside reverse transcriptase inhibitors (NNRTIs) and randomized to one of four treatment arms. 481 subjects were randomized, 211 (44%) of whom were NNRTI experienced. HIV-1 RNA was measured at baseline, week 8, and week 24; viral genotype at baseline and week 8.

We restrict our analysis to those 368 subjects with known values of baseline HIV-1 genotype and RNA measured at baseline, week 8, and week 24, 39% of whom also had week 8 genotype data. As the lower limit of detection of the HIV-1 RNA assay used in this study was 200 copies/mL, we discretized the outcome, week 24 HIV-1 RNA, into the groups 1-200, 201-400, etc. to avoid the problem of left censoring below the limit of the assay. We used the $\log_{10}$ of the upper limit of each interval for our analysis. At week 8, these subjects had a mean HIV-1 RNA of 3.4 logs (median = 2.8). 116 subjects (32%) had RNA levels at the limit of detection (2.3 logs) at Week 8. For our analysis, we categorized week 8 RNA at a clinically relevant threshold value ($<$ 4.0 logs (N = 257) vs $\geq$ 4.0 logs (N = 111)).

The study virologist identified a subset of 45 codons of interest, the 19 protease positions

$$\mathbf{10}, \mathbf{20}, \mathbf{24}, \mathbf{30}, \mathbf{32}, \mathbf{33}, \mathbf{36}, \mathbf{46}, \mathbf{47}, \mathbf{48}, \mathbf{53}, \mathbf{54}, \mathbf{71}, \mathbf{73}, \mathbf{77}, \mathbf{82}, \mathbf{84}, \mathbf{88}, \mathbf{90}$$

and the 26 reverse-transcriptase (RT) positions

$$\mathbf{41}, \mathbf{44}, \mathbf{62}, \mathbf{65}, \mathbf{67}, \mathbf{69}, \mathbf{70}, \mathbf{74}, \mathbf{75}, \mathbf{77}, \mathbf{100}, \mathbf{103}, \mathbf{106}, \mathbf{108}, \mathbf{116}, \mathbf{118}, \mathbf{151},$$
$$\mathbf{181}, \mathbf{184}, \mathbf{188}, \mathbf{190}, \mathbf{210}, \mathbf{215}, \mathbf{219}, \mathbf{225}, \mathbf{230}.$$

We treated each of the codons in our analysis as a binary variable. Tables 1 and 2 show the number of subjects with observed genotype data in the two week 8 RNA categories.

Table 1: *ACTG 398 Subject accounting: Protease*

	$< 4.0 \log_{10}$		$\geq 4.0 \log_{10}$	
Codon	Mut	WT	Mut	WT
PR10	44	13	59	28
PR20	12	45	25	62
PR24	7	50	6	81
PR30	**3**	**54**	**2**	**85**
PR32	5	52	10	77
PR33	**4**	**53**	**6**	**81**
PR36	16	41	35	52
PR46	34	23	33	54
PR47	**2**	**55**	**2**	**85**
PR48	5	52	9	78
PR53	9	48	7	80
PR54	26	31	36	51
PR71	40	17	50	37
PR73	17	40	13	74
PR77	25	32	35	52
PR82	27	30	36	51
PR84	14	43	27	60
PR88	5	52	2	85
PR90	35	22	48	39

The total number of observed week 8 genotype samples was 57 (22% of subjects) for those with week 8 RNA below 4.0 logs and 87 (78% of subjects) for those with week 8 RNA above 4.0 logs. The bolded rows indicate those codons with fewer than five observed genotype samples in any of the four strata.

Since estimation of the conditional means may be unstable with a small

Table 2: *ACTG 398 Subject accounting: Reverse transcriptase*

	$< 4.0 \log_{10}$		$\geq 4.0 \log_{10}$	
Codon	Mut	WT	Mut	WT
RT41	36	21	40	47
RT44	**4**	**53**	**14**	**73**
RT62	**2**	**55**	**5**	**82**
RT65	**1**	**56**	**1**	**86**
RT67	32	25	49	38
RT69	15	42	18	69
RT70	22	35	30	57
RT74	9	48	23	64
RT75	**3**	**35**	**7**	**80**
RT77	**0**	**57**	**1**	**86**
RT100	5	52	12	75
RT103	32	25	59	28
RT106	**0**	**57**	**3**	**84**
RT108	10	47	16	71
RT116	**0**	**57**	**6**	**81**
RT118	18	39	15	72
RT151	**1**	**56**	**8**	**79**
RT181	16	41	37	50
RT184	25	32	31	56
RT188	**3**	**54**	**8**	**79**
RT190	8	49	38	49
RT210	19	38	29	58
RT215	45	12	58	29
RT219	23	34	36	51
RT225	**1**	**56**	**0**	**87**
RT230	**0**	**57**	**1**	**86**

number of complete cases, we conducted the analysis with and without the bolded codons.

3 Model and Notation

Let Y denote a completely observed outcome variable taking at least two levels; we discuss below how to simultaneously incorporate several outcome variables. $\boldsymbol{Z}$ is a J-vector of covariates, where J is not considered large, and $\boldsymbol{X}$ an R-dimensional covariate, where R is considered large. The high-dimensional covariate $\boldsymbol{X}$ is subject to missingness, where δ is the indicator of whether $\boldsymbol{X}$ is observed ($\delta = 1$) or missing ($\delta = 0$). The observed data consists of N independent realizations of $(Y, \boldsymbol{Z}, \delta\boldsymbol{X}, \delta)$ from an unknown probability distribution P, denoted $(Y_n, \boldsymbol{Z}_n, \delta_n\boldsymbol{X}_n, \delta_n)$, $n = 1, \ldots, N$, where, without loss of generality, we take $\boldsymbol{X}$ to equal an R-vector of zeros when $\delta = 0$. We also require the three conditions set out by Robins [7]: (a) the data are MAR, i.e. $P(\delta = 1|\boldsymbol{X}, f(Y, \boldsymbol{Z})) = P(\delta = 1|f(Y, \boldsymbol{Z}))$, for some general function $f(.)$ of the observed data, e.g. $f(Y, \boldsymbol{Z}) = f(\boldsymbol{Z})$, (b) the probabilities $P(\delta = 1|f(Y, \boldsymbol{Z}))$ are bounded away from 0, and (c) there exist consistent estimators for $P(\delta = 1|f(Y, \boldsymbol{Z}))$.

4 Hypothesis Testing with Known Weights

We are interested in simultaneously testing for the effect of the potentially missing covariate $\boldsymbol{X}$ on the conditional mean of Y above and beyond that already explained by $\boldsymbol{Z}$. For simplicity, we consider the case with R binary covariates $\boldsymbol{X}$ and a categorical scalar $\boldsymbol{Z}$. Extension to multiple levels for $\boldsymbol{X}$ and several $\boldsymbol{Z}$ is straightforward and outlined below. Future work will address continuous $\boldsymbol{Z}$. The $K = 2 * R * J$ null hypotheses take the form

$$H^0_{r,x,j} : E(Y \mid Z = z_j) = E(Y \mid X_r = x, Z = z_j), \qquad (2)$$

for $r = 1, \ldots, R, x = 0, 1, j = 1, \ldots, J$. Note that one could instead consider the $R * J$ null hypotheses

$$H^0_{r,j} : E(Y \mid X_r = 0, Z = z_j) = E(Y \mid X_r = 1, Z = z_j),$$

for $r = 1, \ldots, R, j = 1, \ldots, J$, which may result in an increase in power to detect false null hypotheses, as fewer null hypotheses are tested. On the other hand, the variance of the difference between empirical means for the $2 * R * J$ null hypotheses is generally less than that corresponding to the $R * J$ null hypotheses. We ran the analysis of ACTG 398 using both the $R * J$ and $2 * R * J$ approaches; both identified the same codons that are informative for week 24 RNA given that already explained by week 8 RNA.

When the data are completely observed, corresponding test statistics $T_{r,x,j}$ based on the empirical versions of these expectations can be written as

$$T^{obs}_{r,x,j} = \frac{\frac{1}{N}\sum_{n=1}^{N} Y_n I(Z_n = z_j)}{\frac{1}{N}\sum_{n=1}^{N} I(Z_n = z_j)} - \frac{\frac{1}{N}\sum_{n=1}^{N} Y_n I(X_{r,n} = x) I(Z_n = z_j)}{\frac{1}{N}\sum_{n=1}^{N} I(X_{r,n} = x) I(Z_n = z_j)}.$$

When the data are not completely observed, these test statistics cannot be calculated. The complete case version of these test statistics incorporates the missingness indicator δ,

$$T^{cc}_{r,x,j} = \frac{\frac{1}{N}\sum_{n=1}^{N} Y_n I(Z_n = z_j)}{\frac{1}{N}\sum_{n=1}^{N} I(Z_n = z_j)} - \frac{\frac{1}{N}\sum_{n=1}^{N} \delta_n Y_n I(X_{r,n} = x) I(Z_n = z_j)}{\frac{1}{N}\sum_{n=1}^{N} \delta_n I(X_{r,n} = x) I(Z_n = z_j)}, \quad (3)$$

but this version will be biased when the data are not MCAR.

Consider instead a weighted version,

$$T^{\omega}_{r,x,j} = \frac{\frac{1}{N}\sum_{n=1}^{N} Y_n I(Z_n = z_j)}{\frac{1}{N}\sum_{n=1}^{N} I(Z_n = z_j)} - \frac{\frac{1}{N}\sum_{n=1}^{N} \omega_r(Y_n, x, z_j) Y_n I(X_{r,n} = x) I(Z_n = z_j)}{\omega'_r(x, z_j)\frac{1}{N}\sum_{n=1}^{N} I(X_{r,n} = x) I(Z_n = z_j)} \quad (4)$$

For this statistic to be unbiased, we must choose $\omega_r(Y_n, x, z_j)$ and $\omega'_r(x, z_j)$ so that the second term in $T^{\omega}_{r,x,j}$ converges to the limit of the second term in $T^{obs}_{r,x,j}$, or $E[YI(X_r = x)I(Z = z_j)]/E[I(X_r = x)I(Z = z_j)]$ in the absence of missing data. It is straightforward to show that the second term in $T^{\omega}_{r,x,j}$ converges to

$$\frac{E[YI(X_r = x)I(Z = z_j)\omega_r(Y, x, z_j)P(\delta = 1 \mid Y, X_r, Z)]}{E[I(X_r = x)I(Z = z_j)\omega'_r(x, z_j)P(\delta = 1 \mid X_r, Z)]}.$$

Thus setting $\omega_r(Y_n, x, z_j)$ equal to $1/P(\delta = 1 \mid Y_n, x, z_j)$ and $\omega'_r(x, z_j)$ equal to $1/P(\delta = 1 \mid x, z_j)$ will lead to a weighted test statistic with the appropriate limit in probability. These weights, however, include the potentially missing high dimensional covariate $\boldsymbol{X}$. By making use of the MAR assumption, $P(\delta = 1 \mid Y, X_r, Z) = P(\delta = 1 \mid Y, Z)$ and similarly $P(\delta = 1 \mid X_r, Z) = P(\delta = 1 \mid Z)$. Thus the weights need only depend on completely observed Y and $\boldsymbol{Z}$, and can be written as $\omega(Y, \boldsymbol{Z})$ and $\omega'(\boldsymbol{Z})$.

We want to simultaneously test these K hypotheses with a method that provides asymptotic control of the family-wise error rate (FWER). For clarity, we first describe the method using the generalized single-step common-cut-off bootstrap procedure given in [2]. Efficiency may be gained by using the step-down method in [16], which follows.

4.1 Single step Procedure

Independently sample with replacement N observations from $(Y_n, \boldsymbol{Z}_n, \delta_n \boldsymbol{X}_n, \delta_n)_{n=1}^N$; independently repeat this M times. Corresponding to each of the M bootstrap datasets, calculate the K-vector of test statistics, denoted $(T^\#_{m,1}, \ldots, T^\#_{m,K})$ for $m = 1, \ldots, M$. Next, calculate the mean statistics $\overline{T}^\#_k = M^{-1} \sum_{m=1}^M T^\#_{m,k}$, $k = 1, \ldots, K$, and construct the null realizations

$$W_{m,k} = T^\#_{m,k} - \overline{T}^\#_k, \quad k = 1, \ldots, K, \quad m = 1, \ldots, M.$$

Let $W^\dagger_m = \max\limits_{1 \le k \le K} |W_{m,k}|$, $m = 1, \ldots, M$. The two-sided p-value for H_k adjusted for multiple testing is estimated by

$$\hat{\pi}_k = \frac{1}{M} \sum_{m=1}^M I(W^\dagger_m \ge |T^\circ_k|), k = 1, \ldots, K,$$

where $(T^\circ_k)_{k=1}^K$ are the test statistics calculated from the observed data. Those H_k with $\hat{\pi}_k \le \alpha$ are rejected at target FWER α.

4.2 Step down Procedure

Order the observed absolute test statistics $(|T^\circ_k|)_{k=1}^K$ to obtain $|T^\circ_{O(1)}| \le \cdots \le |T^\circ_{O(K)}|$. Calculate the $M \times K$ matrix $(W_{m,k})$ as above, and follow the algorithm below.

1. Set $s = 0$.
2. Calculate $W^\dagger_m(s) = \max(|W_{m,O(k)}|)_{k=1}^{K-s}$
3. Calculate $\tilde{\pi}_s = \frac{1}{M} \sum_{m=1}^M I\{W^\dagger_m(s) \ge |T^\circ_{O(K-s)}|\}$
4. If $\alpha \ge \tilde{\pi}_s$ then reject $H_{O(K-s)}$, increment s to $s+1$ and return to step 2; otherwise, if $\tilde{\pi}_s > \alpha$ then do not reject $H_{O(K-s)}, \ldots, H_{O(1)}$ and the algorithm stops.

An overall adjusted two-sided p-value estimate for $H_{O(K-s)}$, say $\hat{\pi}_{O(K-s)}$, $s = 0, \ldots, K-1$, is the smallest FWER such that $H_{O(K-s)}$ is rejected by the step down algorithm. This estimator is obtained as $\hat{\pi}_{O(K-s)} = \max(\tilde{\pi}_{s'})_{s'=0}^s$. If the step down algorithm is conducted at FWER $\hat{\pi}_{O(K-s)}$ then $H_{O(K-s)}$ will be rejected.

5 Estimated Weights and Extensions

In practice, the weight functions $\omega(.)$ and $\omega'(.)$ are usually unknown. Standard methods for estimating $P(\delta_n = 1 \mid Y_n, z_j)$ and $P(\delta_n = 1 \mid z_j)$, such as logistic regression models can be used. We focus first on $\omega(\mathrm{y}, \mathbf{z})$ by specifying the following model:

$$P(\delta_n = 1 \mid Y = \mathrm{y}_n, \boldsymbol{Z} = \mathbf{z}_n) = \frac{\exp(\beta_0 + \beta_1 \mathrm{y}_n + \boldsymbol{\beta_2} \mathbf{z}_n + \boldsymbol{\beta_3} \mathrm{y}_n \mathbf{z}_n)}{1 + \exp(\beta_0 + \beta_1 \mathrm{y}_n + \boldsymbol{\beta_2} \mathbf{z}_n + \boldsymbol{\beta_3} \mathrm{y}_n \mathbf{z}_n)} = u(\mathrm{y}_n, \mathbf{z}_n; \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta_2}, \boldsymbol{\beta_3})$.

Tian and colleagues [12] demonstrate that the estimating equation

$$0 = S(\hat{\boldsymbol{\beta}}) = \sum_{n=1}^{N} (\mathrm{y}_n, \mathbf{z}_n)(\delta_n - u(\mathrm{y}_n, \mathbf{z}_n; \boldsymbol{\beta}))$$

can be used to obtain consistent estimates for $\boldsymbol{\beta}$ via an application of the Newton-Raphson algorithm; these estimators will converge to a constant vector even if the working model above is misspecified.

Let $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_3$ denote the estimated regression coefficients and denote

$$\widehat{\omega}(\mathrm{y}_n, \mathbf{z}_n) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{y}_n + \hat{\boldsymbol{\beta}}_2 \mathbf{z}_n + \hat{\boldsymbol{\beta}}_3 \mathrm{y}_n \mathbf{z}_n)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{y}_n + \hat{\boldsymbol{\beta}}_2 \mathbf{z}_n + \hat{\boldsymbol{\beta}}_3 \mathrm{y}_n \mathbf{z}_n)}.$$

An approach for inference is to substitute $\widehat{\omega}(\mathrm{y}, \mathbf{z})$ for $\omega(\mathrm{y}, \mathbf{z})$ (and similarly calculated $\widehat{\omega}(\mathbf{z})$ for $\omega(\mathbf{z})$) in the procedure described in Section 4 above. However, this approach does not account for variability from the estimated weights, which may bias the p-values. According to Efron [3], when using this approach a third-order accurate p-value estimate is required for valid inference. Two approaches can be used to achieve this accuracy, the so-called double bootstrap of Beran [1], and the multiscale bootstrap of Shimodaira [11]. We now describe the double bootstrap for the step down algorithm above.

5.1 Single-step Double Bootstrap

Our estimation approach proceeds as follows:

1. Estimate the weights $\widehat{\omega}(\mathrm{y}, \mathbf{z})$ and $\widehat{\omega}(\mathbf{z})$ using the original data and use them to calculate (T_k°).

2. Independently draw M bootstrap samples of size N from the original data $(Y_n, \boldsymbol{Z}_n, \delta_n \boldsymbol{X}_n, \delta_n)_{n=1}^N$ and calculate the K-vector of test statistics for each sampleusing the estimated weights, $\widehat{\omega}(\mathrm{y}, \mathbf{z})$ and $\widehat{\omega}(\mathbf{z})$, obtaining $\{T^{\#}_{m,1}, \ldots, T^{\#}_{m,K}\}_{m=1}^M$.

3. Use the single-step approach outlined in Section 4.1 to calculate a two-sided p-value, adjusted for multiple testing, for each of the K hypotheses, denoted $\hat{\pi}_k$, $k = 1, \ldots, K$. Denote $\tau_m = \max_{1 \le k \le K} |W_{m,k}|$, $m = 1, \ldots, M$.

4. Re-estimate the weights on each bootstrap sample of data; denote the mth bootstrap sample by $(Y_{n,m}, \boldsymbol{Z}_{n,m}, \delta_{n,m}\boldsymbol{X}_{n,m}, \delta_{n,m})_{n=1}^N$ and the re-estimated weights by $\widehat{\omega}_m(\mathrm{y}, \mathbf{z})$ and $\widehat{\omega}_m(\mathbf{z}), m = 1, \ldots, M$.

5. Draw B additional bootstrap samples of size N from each original bootstrap sample $(Y_{n,m}, \boldsymbol{Z}_{n,m}, \delta_{n,m}\boldsymbol{X}_{n,m}, \delta_{n,m})_{n=1}^N$, $m = 1, \ldots, M$.

6. Using the estimated weights $\widehat{\omega}_m(\mathrm{y}, \mathbf{z})$ and $\widehat{\omega}_m(\mathbf{z})$, calculate the K-vector of test statistics $\{T^{\#}_{m,1,b}, \ldots, T^{\#}_{m,K,b}\}$ for each of the B samples, $b = 1, \ldots, B$, and let $\{\overline{T}^{\#}_{m,1}, \ldots, \overline{T}^{\#}_{m,k}\}$ denote the mean over B of these test statistics for the mth bootstrap sample. Obtain $\{W_{m,1,b}, \ldots, W_{m,K,b}\} = \{T^{\#}_{m,1,b}, \ldots, T^{\#}_{m,K,b}\} - \{\overline{T}^{\#}_{m,1}, \ldots, \overline{T}^{\#}_{m,K}\}, b = 1, \ldots, B$.

7. Let $\tau_{m,b} = \max_{1 \le k \le K} |W_{m,k,b}|$ and calculate the empirical probability

$$\pi(\tau_m) = B^{-1} \sum_{b=1}^{B} I(\tau_{m,b} \ge \tau_m), m = 1, \ldots, M. \qquad (5)$$

The empirical distribution function (EDF) formed by $\{\pi(\tau_m)\}_{m=1}^M$ is then used to adjust the $\hat{\pi}_k$ obtained in step 3 above. The adjusted $\hat{\pi}_k$ is the value of the EDF of $\{\pi(\tau_m)\}_{m=1}^M$ at $1 - \hat{\pi}_k$.

5.2 Step-down Double Bootstrap

Application of the double bootstrap to a step-down testing algorithm proceeds analogously. The difference is that for the single-step application, the p-values $(\hat{\pi}_k)_{k=1}^K$ are adjusted; for the step-down application, $\tilde{\pi}_s$ is adjusted at each step s of the algorithm. For the single-step application, $\hat{\pi}_k$ is adjusted using the EDF of $\{\pi(\tau_m)\}_{m=1}^M$. For the step-down application, $\tilde{\pi}_s$ is adjusted in the same way, using the EDF of $\{\pi(\tau_{m,s})\}_{m=1}^M$, where $\pi(\tau_{m,s})$ is defined as $\pi(\tau_m)$

except only using the columns of (T_k°), $(W_{m,k})$ and $(W_{m,k,b})$, that correspond to $H_{O(1)}, \ldots, H_{O(K-s)}$, as opposed to all K columns for $\pi(\tau_m)$.

Adding the second layer of B bootstrap samples to the resampling scheme can greatly increase the computation demand, although methods exist that may reduce the burden [11]. Beran [1] notes that he did not have a good theoretical justification of the choice of values for M and B, but suggested that letting $M = B = 1000$ provided reasonable performance.

5.3 Extensions to Other Outcome Variables

A natural extension to this proposed method would be to simultaneously consider several outcome variables. For example, we could use longitudinal outcomes, such as HIV-1 RNA at weeks 24, 48 and 96 weeks on study, or outcomes of different types, such as CD4 cell count along with HIV-1 RNA. The methods described here can be readily extended to this case. Suppose there are D outcome variables of interest, $Y_1, \ldots, Y_D$, where each Y_d is either continuous or binary. Define the null hypothesis for the dth outcome variable as $H_{r,x,j}^{(d)} : E(Y^{(d)}|Z = z_j) = E(Y^{(d)}|X_r = x, Z = z_j)$, with the corresponding version of the test statistic $T_{r,x,j}^{\omega}$ denoted $\tilde{T}_{r,x,j}^{(d)}$, calculated with $\omega(Y^{(d)}, Z)$ and $\omega'(Z)$. The sequence of null hypotheses to be tested would then be$\{(H_{r,x,j}^{(1)}), \ldots, (H_{r,x,j}^{(D)})\}$, with the corresponding summary statistics $\{(\tilde{T}_{r,x,j}^{(1)}), \ldots, (\tilde{T}_{r,x,j}^{(D)})\}$. The bootstrap samples are calculated by sampling with replacement from the rows of $(Y_{n,1}, \ldots, Y_{n,D}, \boldsymbol{Z}_n, \delta_n \boldsymbol{X}_n, \delta_n)_{n=1}^N$.

The methods described could also extend to the case where Y is an qualitative outcome variable. When Y takes $L > 2$ levels, labeled $\mathrm{y}_1, \ldots, \mathrm{y}_L$, we can generate $L - 1$ dummy variables, $I(Y = \mathrm{y}_\ell)$, $\ell = 2, \ldots, L$, calculate the $L-1$ summary statistics as described above, then average them to arrive at a single summary statistic for Y.

6 Simulation Study

To examine the finite sample properties of the proposed test statistics, we conducted a simulation study in three settings:

(i) the missing data mechanism depends only on the covariates Z, so that the complete-case analysis is unbiased, and the models for $\omega(.)$ and $\omega'(.)$ are properly specified;

(ii) the missing data mechanism depends on both Z and Y, so that the complete-case analysis is biased, and the models for $\omega(.)$ and $\omega'(.)$ are properly specified;

(iii) the missing data mechanism depends on both Z and Y, so that the complete-case analysis is biased, and the models for $\omega(.)$ and $\omega'(.)$ are misspecified.

The covariate vector $\boldsymbol{X}$ consisted of $R = 9$ binary variables whose joint distribution $P(\boldsymbol{X})$ is defined in the Appendix. The distribution of the binary covariate Z was defined as $P(Z = 1|X_9 = 0) = 0.3$, and $P(Z = 1|X_9 = 1) = 0.7$. The continuous outcome variable Y was defined as:

$$Y = 3X_9 + Z - \frac{ZX_9}{2} + \xi,$$

where ξ is distributed normally with mean 0 and standard deviation 1, and is independent of $(\boldsymbol{X}, Z)$. Finally, for setting (i), define

$$\text{logit}\{P(\delta = 1|Y, Z)\} = -2Z$$

and for settings (ii) and (iii) define

$$\text{logit}\{P(\delta = 1|Y, Z)\} = -2Z + \frac{Y}{2} + \frac{3ZY}{4}.$$

We chose the parameters for these logistic regression models to be similar to the corresponding estimates obtained in the analysis of data from ACTG 398. For setting (iii), the working model for logit$\{P(\delta = 1|Y, Z)\}$ was $\beta_0+\beta_1 Z+\beta_2 Y$.

This data structure defined $K = 36$ null hypotheses $(H_{r,\text{x},j})$ of interest. Of these, notice that all of the null hypotheses are true except for $H_{9,0,1}(0.9)$, $H_{9,1,1}, (-2.1)$ $H_{9,0,2}, (1.75)$, $H_{9,1,2}, (-0.75)$, where the number in parentheses denotes the value of $E(Y|Z = \text{z}_j) - E(Y|X_r = \text{x}, Z = \text{z}_j)$ which can be calculated from the expression for Y above.

We considered three sample sizes for each of the three settings: $N = 100$, 250 and 500. Separately for each setting and sample size, we generated 1000 independent random samples $(\boldsymbol{X}_n)_{n=1}^N$ from $P(\boldsymbol{X})$. At each of the 1000 simulation iterations, we randomly generated $X_{n,1}$ from $P(X_1)$, then $X_{n,2}$ from $P(X_2|X_{n,1})$, and so on. We then generated the variable Z_n from $P(Z|X_{n,9})$, obtained the random sample $(\xi_n)_{n=1}^N$, and calculated $(Y_n)_{n=1}^N$. With $(Y_n, Z_n)_{n=1}^N$ in hand, we generated each δ_n independently from a Bernoulli distribution with parameter defined by the logistic model above. At each simulation iteration, we estimated the weight functions $\omega(.)$ and $\omega'(.)$ using the observed data and then used in each of the $M = 1000$ independent bootstrap samples. We did

Table 3: *Empirical rejection proportions at $\alpha = 0.05$*

		$H_{9,0,1}$	$H_{9,1,1}$	$H_{9,0,2}$	$H_{9,1,2}$	FWER
				Setting (i)		
$N = 100$	T^{ω} :	0	0.3287	0.0872	0.0180	0.1333
	T^{cc} :	0	0.3390	0.1380	0.0160	0.2480
$N = 250$	T^{ω} :	0.0110	0.4605	0.1612	0	0.0661
	T^{cc} :	0	0.4880	0.2210	0	0.0990
$N = 500$	T^{ω} :	0.0440	0.7500	0.5320	0	0.0240
	T^{cc} :	0.0040	0.8440	0.6000	0.0030	0.0480
				Setting (ii)		
$N = 100$	T^{ω} :	0.0150	0.2490	0.0320	0	0.0240
	T^{cc} :	0	0.8010	0.0730	0.0020	0.1150
$N = 250$	T^{ω} :	0.4380	0.8470	0.1740	0	0.0150
	T^{cc} :	0.0480	1.0000	0.6967	0.2803	0.4925
$N = 500$	T^{ω} :	0.9190	0.9570	0.4540	0	0.0300
	T^{cc} :	0.5940	1.0000	0.9920	0.9890	0.9490
				Setting (iii)		
$N = 100$	T^{ω} :	0.0310	0.0890	0.0470	0	0.0130
	T^{cc} :	0	0.8150	0.0990	0.0030	0.1130
$N = 250$	T^{ω} :	0.5330	0.5780	0.2960	0	0.0350
	T^{cc} :	0.0440	1.0000	0.6840	0.2800	0.4920
$N = 500$	T^{ω} :	0.9520	0.9150	0.7030	0	0.0590
	T^{cc} :	0.5900	1.0000	0.9970	0.9840	0.9430

Note: For $N = 100, 250, 500$, the avg$\sum \delta_n$, taken over the 1000 simulation iterations, were, for Setting (i): 30.79, 77.45, 154.99; Setting (ii): 65.17, 163.33, 326.36; and Setting (iii): 65.50, 162.93, 327.38.

not use the double bootstrap for these simulations. The simulations were executed using Matlab version 7.3 (R2006b) and with 64-bit Intel Xeon 3.2

GHz processors.

Table 3 presents empirical FWERs and rejection percentages for the four false null hypotheses. Notice that with 1000 independent simulation iterations, the normal theory-based 95% two-sided confidence interval for a nominal FWER of 5% extends ±1.35% of the empirical FWER%. For setting (i), statistic (4) had empirical FWER at or below the nominal level for $N = 250$ and $N = 500$, while statistic (3) achieved the nominal FWER only for $N = 500$. For settings (ii) and (iii), statistic (4) produced empirical FWERs at or below the nominal level for all sample sizes. On the other hand, use of (3) resulted in empirical FWERs above the nominal level, with the bias increasing with sample size. Inference using (4) appeared undistorted from this mild misspecification of the model for $\text{logit}\{P(\delta = 1|Y, Z)\}$. The statistics (4) and (3) are only comparable with respect to power for setting (i) with $N = 500$, where (3) showed slightly greater power for $H_{9,1,1}$ and $H_{9,0,2}$.

7 Application to ACTG 398

Recall that in ACTG 398 the outcome variable Y is viral load at week 24, the fully observed low-dimensional short term response $\boldsymbol{Z}$ is viral load at week 8, and $\boldsymbol{X}$ is the viral genotype, consisting of the $R = 45$ binary codons identified by the study virologist. Each variable X_r is defined to be 1 if the amino acid equaled the wildtype amino acid, and 0 otherwise.

We used logistic regression models to estimate the weight functions:

$$\text{logit}\{P(\delta = 1|Y, Z)\} = \beta_0 + \beta_1 Z + \beta_2 Y + \beta_3 ZY$$

and

$$\text{logit}\{P(\delta = 1|Z)\} = \gamma_0 + \gamma_1 Z.$$

The estimated maximum likelihood estimators were $\hat{\beta}_0 = 0.30$, $\hat{\beta}_1 = -4.44$, $\hat{\beta}_2 = 0.20$, $\hat{\beta}_3 = 0.67$ and $\hat{\gamma}_0 = 1.29$, $\hat{\gamma}_1 = -2.54$. Thus, $\hat{P}(\delta = 1|Z = 0) = 0.78$ and $\hat{P}(\delta = 1|Z = 1) = 0.22$, corresponding to the proportion of subjects with missing data at each level of week 8 viral load.

To simultaneously test the $R \times 2 \times J = 180$ null hypotheses ($H_{r,\mathrm{x},j}$), we used the step down procedure outlined in Section 4.2 and used the weights obtained via the logistic models above in each of $M = 1000$ independent bootstrap samples. We did not use the double bootstrap in this application, as the simulation results, which did not use the double bootstrap, showed little distortion of the empirical FWER.

When considering all available codons, we rejected two of the 180 hypotheses at the $\alpha = 0.05$ level, corresponding to RT positions **65** and **151**.

For subjects with week 8 RNA below 4 $\log_{10}$, the mean week 24 RNA was estimated to be 1.78 $\log_{10}$ larger for subjects mutant at **RT65** ($\hat{\pi} = 0.02$), and was estimated to be 1.78 $\log_{10}$ larger for subjects mutant at **RT151** ($\hat{\pi} = 0.02$). These are known resistance mutations to drugs in the backbone regimen used in ACTG 398.

However, we noticed that there was a single subject with week 8 RNA below 4 $\log_{10}$ and mutations at **RT65** and **RT151**. When removing the hypotheses corresponding to the bolded codons from Tables 1 and 2, no hypothesis was rejected at the $\alpha = 0.05$ level; in fact the smallest adjusted p-value was $\hat{\pi} = 0.64$.

8 Discussion

In this paper, we considered the question of whether high-dimensional longitudinal information, e.g. week 8 genotype sequence, can provide more information about a longer-term endpoint than a low-dimensional variable alone. Such methods may be useful for identifying a treatment that is at high risk of failure. Incorporating information from the high-dimensional marker could improve understanding of the causes of future failures by providing insight about the mechanism by which treatments fail. Furthermore, our approach permits the identification of potentially influential outliers, such as the individual with the mutation at **RT65** and **RT151** in ACTG 398. Such exploratory investigations can help in targeting rare, but potentially important, mutations for further research.

Our methods could also be easily adapted to address another question of interest: does the high-dimensional marker enhance the degree to which the low dimensional marker captures the treatment effect on the longer-term endpoint? This relates to the question of whether high dimensional markers can improve the performance of existing surrogate endpoints in clinical trials. For example, week 8 HIV-1 RNA levels may be a good surrogate for longer-term outcomes for some patients; but for others, inclusion of genetic information may be necessary to capture effects of treatment on a longer-term outcome.

The high dimensional nature of the viral genotype lends itself naturally to considering the impact of multiple mutations and interactions between mutations on long-term clinical outcomes. Other work [10] has addressed the potential for interactions among mutations in their effects on viral load; more complex hypotheses regarding the impact of such interactions could be constructed here as well. We were limited by the data available from ACTG 398 in this application, but the method generalizes easily to such settings.

Appendix

Probability distribution of $\boldsymbol{X}$ for simulation study

Denote the probability mass function $p_{X_i|X_j=\mathrm{x}} = P(X_i = 1|X_j = \mathrm{x})$ for the binary variables X_i and X_j, $i \neq j$.

The non-zero components of $P(\mathbf{x})$ are $p_{X_1} = \frac{1}{4}$, $p_{X_2|X_1=\mathrm{x}} = \begin{cases} 1/2; \mathrm{x} = 0 \\ 9/10; \mathrm{x} = 1 \end{cases}$, $p_{X_3} = \frac{1}{2}$, $p_{X_4|X_3=\mathrm{x}} = \begin{cases} 3/10; \mathrm{x} = 0 \\ 7/10; \mathrm{x} = 1 \end{cases}$, $p_{X_5} = \frac{3}{4}$, $p_{X_6|X_5=\mathrm{x}} = \begin{cases} 3/5; \mathrm{x} = 0 \\ 4/5; \mathrm{x} = 1 \end{cases}$, $p_{X_7} = \frac{1}{4}$, $p_{X_8|X_7=\mathrm{x}} = \begin{cases} 1/5; \mathrm{x} = 0 \\ 4/5; \mathrm{x} = 1 \end{cases}$, $p_{X_9} = \frac{1}{2}$.

References

[1] R Beran. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74:457–468, 1987.

[2] S Dudoit, MJ van der Laan, and KS Pollard. Multiple testing. part I, single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(13), 2004.

[3] B Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82:171–185, 1987.

[4] B Efron. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89:463–475, 1994.

[5] SM Hammer et al. Dual vs single protease inhibitor therapy following antiretroviral treatment failure: a randomized trial. *JAMA*, 288:169–180, 2002.

[6] W Huang, V De Gruttola, RM Gulick, M Fischl, D Havlir, SM Hammer, J Mellors, DD Richman, and KE Squires. Pattern of plasma HIV-1 RNA response to antiretroviral therapy. *Journal of Infectious Diseases*, 183(10):1455–1465, 2001.

[7] JM Robins, A Rotnitzky, and LP Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

DOI: 10.2202/1557-4679.1102

[8] JM Robins, A Rotnitzky, and LP Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121, 1995.

[9] DB Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[10] J Schumi and V De Gruttola. Resampling-based analyses of the effects of combinations of HIV genetic mutations on drug susceptibility. *Statistics in Medicine*, DOI: 10.1002/sim.3181, 2008.

[11] H Shimodaira. Approximately unbiased tests of regions using multistep-multiscale bootstrap sampling. *Annals of Statistics*, 32:2616–2641, 2004.

[12] L Tian, T Cai, E Goetghebeur, and LJ Wei. Model evaluation based on the distribution of estimated absolute prediction error. *Biometrika*, 94:297–311, 2007.

[13] AA Tsiatis. *Semiparametric theory and missing data*. Springer, New York, 2006.

[14] MJ van der Laan, MD Birkner, and AE Hubbard. Resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 4(29), 2005.

[15] MJ van der Laan, S Dudoit, and KS Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(15), 2004.

[16] MJ van der Laan, S Dudoit, and KS Pollard. Multiple testing. part II, step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, 3(14), 2004.