The International Journal of Biostatistics

Volume 2, Issue 1

2006

Article 5

The Two Sample Problem for Multiple Categorical Variables

A. G. DiRienzo, *Harvard University*

Recommended Citation:

DiRienzo, A. G. (2006) "The Two Sample Problem for Multiple Categorical Variables," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 5.

DOI: 10.2202/1557-4679.1019

The Two Sample Problem for Multiple Categorical Variables

A. G. DiRienzo

Abstract

Comparing two large multivariate distributions is potentially complicated at least for the following reasons. First, some variable/level combinations may have a redundant difference in prevalence between groups in the sense that the difference can be completely explained in terms of lower-order combinations. Second, the total number of variable/level combinations to compare between groups is very large, and likely computationally prohibitive. In this paper, for both the paired and independent sample case, an approximate comparison method is proposed, along with a computationally efficient algorithm, that estimates the set of variable/level combinations that have a non-redundant different prevalence between two populations. The probability that the estimate contains one or more false or redundant differences is asymptotically bounded above by any pre-specified level for arbitrary data-generating distributions. The method is shown to perform well for finite samples in a simulation study, and is used to investigate HIV-1 genotype evolution in a recent AIDS clinical trial.

KEYWORDS: conditional independence, family-wise error rate, genomics, genotype sequence, HIV-1 genotype evolution, multiple hypothesis testing, simultaneous inference

Author Notes: This work was supported by grant R01 AI058836 from the U.S. National Institutes of Health. The author thanks Marcello Pagano, Victor DeGruttola and John Mellors for fruitful discussions, and two reviewers whose comments improved the clarity and presentation of the article.

1 Introduction

The development of new technologies, for example, genotype sequencing and gene expression arrays, has sparked an increased interest in the field of multiple hypothesis testing. Such technologies typically give rise to datasets with many more variables than data points; furthermore, the variables potentially have a complex and unknown high-dimensional dependency structure. For example, AIDS Clinical Trials Group study 398 (ACTG 398) randomized treatment experienced subjects with loss of virological suppression to one of four salvage regimens. The primary endpoint was the time from randomization to virologic failure. An HIV-1 genotype sequence from the protease (PR) and reverse-transcriptase (RT) regions was obtained for each subject at baseline and at the time of virologic failure for each subject experiencing this event. The variable at each position (codon) in the genotype sequence is an amino acid, taking one of 20 possible unordered values. Drug-resistant virus is known to manifest with potentially complex mutational patterns in the HIV-1 genotype sequence. Because sequencing HIV-1 has dropped to a cost that allows its use in routine subject management, identifying those patterns of mutations evident of resistance is potentially useful for subject-specific drug selection. Therefore, an important question in ACTG 398 is what specific genotype changes are associated with the risk of virologic failure. That is, for those subjects observed to experience virologic failure, what components of the joint distribution of HIV-1 genotype sequence differ between the baseline and post-baseline populations. Note that post-baseline genotypes are not available for those subjects not observed to experience virologic failure, since there is not enough virus for amplification. For such cases virologists typically assume no changes in genotype sequence.

The problem of comparing two high-dimensional multivariate distributions is complicated at least for the following reasons. First, there are possibly redundant differences between groups. For example, suppose the event A has different prevalence between groups and the event B is independent of A within each group but has equal prevalence between groups. Then, the event $(A \cap B)$ has different prevalence between groups but is redundant because it is solely driven by the event A . Second, the number of variable/level combinations to compare between groups is very large. Suppose there are R variables of interest. A full comparison between groups gives consideration to all marginal (first-order) variable/level pairs and all two-way thru R -way combinations of variable/level pairs, which in general will not be computationally feasible.

There has been much recent work on the topic of controlling error rates in simultaneous multiple hypothesis testing. In particular, van der Laan, Du-

doit and Pollard (2004a) propose stepdown bootstrap-based methods for testing multiple hypotheses that correspond to general submodels for the data-generating distribution. For asymptotically linear summary statistics, these methods provide asymptotic control the family-wise error rate (FWER) for arbitrary data-generating distributions. Note that the general hypothesis testing framework of van der Laan et al. (2004a) was initially proposed by Pollard and van der Laan (2003) and more rigorously developed in Dudoit, van der Laan and Pollard (2004), both for a single-step testing procedure. Romano and Wolf (2005) provide general conditions for stepdown methods to provide either finite or asymptotic control of the FWER under arbitrary data-generating distributions. Basically, the required conditions in Romano and Wolf (2005) are (i) monotonicity of the estimated critical values thru the steps, and (ii) each test is at or below target level α , either asymptotically (for asymptotic control) or in finite samples (for finite-sample control). However, the problem of comparing the joint distribution of R random variables has typically been attacked by restricting attention to only the R univariate distributions (Romano and Wolf, 2005, Example 5; Troendle (1995)). An exception is the recent work by Cox and Reid (2004), who consider approximating the full R -dimensional likelihood with all univariate and bivariate density functions. In that article, the authors outline the asymptotic properties of the resulting maximum (approximate) likelihood estimators and provide some examples, e.g. one-way random effects models.

Proposed in this paper is a computationally efficient nonparametric method for approximating a full parsimonious comparison of two large-scale joint distributions. More specifically, the approximation limits the number of variable/level pairs in a combination at some value $K < R$, where K is chosen by the user. If the largest order of dependency among variables is of order K or less, then limiting the number of variable/level pairs in a combination at K actually results in a full (not approximate) parsimonious comparison of the distributions. This is because differences between groups with respect to prevalence of combinations of more than K variable/level pairs would be redundant in the sense that the difference is completely explained by differences in prevalence of combinations of order K or less. Since in practice, one never knows the largest order of dependency among variables, limiting the number of variable/level pairs in a combination at K in general results in an approximate comparison. Further approximations are also presented that consider only those combinations for which at least a pre-specified number of variable/level pairs have a significant marginal (one-way) group effect. The computer algorithm proposed to execute the method converts a potentially computer-memory intensive method to a computer-processor intensive

method — the latter resource being much cheaper relative to the former.

For both the paired and independent sample case, the approximate parsimonious comparison method necessarily tests two types of multiple null hypotheses: equal prevalence and conditional independence. For each combination of variable/level pairs under consideration, one first tests if the prevalence of the combination is the same in the two groups. If the null hypothesis of equal prevalence is not rejected, the combination is excluded from further analysis. On the other hand, if the combination is declared to occur more frequently in one group, then a parsimonious comparison method requires testing if the prevalence of the combination can be written as the same product of prevalences of lower-order combinations in each group. That is, it is required to test for conditional independence of events defining the combination given group membership. If the hypothesis of conditional independence is not rejected in both groups, then even though the combination was declared to have different prevalence between groups, the difference is considered redundant and the combination is discarded. This is because the difference in prevalence is completely driven by lower-order combinations. For a given combination, the number of conditional independence null hypotheses to test in each group depends on the number of variable/level pairs in the combination. For example, a combination of two variable/level pairs has one corresponding conditional independence null hypothesis per group, whereas a combination of three variable/level pairs has three corresponding conditional independence null hypotheses per group.

The goal of the proposed method is to estimate the set of variable/level combinations that have a non-redundant different prevalence between the two populations. For each combination in the estimated set, it is shown that the probability that the combination either has equal prevalence between groups, or corresponds to a redundant group difference, is bounded above asymptotically by any pre-specified level, for any data-generating mechanism.

The paper is organized as follows. The observed data is described in Section 2. Section 3 outlines the issues of redundancy and computation when comparing the full distributions and proposes an approximation to the full parsimonious comparison. The null hypotheses of interest and associated summary statistics are provided in Section 4. Section 5 presents the general bootstrap inference procedure and Section 6 describes the proposed estimation method. A simulation study is presented in Section 7 and shows that the method performs well in finite samples. An analysis of ACTG 398 is provided in Section 8. Some concluding remarks are given in Section 9.

2 Data

Let $\mathbf{Y} = (Y_1, \dots, Y_R)$ denote an R -vector of categorical variables, where R is considered large. Suppose N^0 independent realizations of $\mathbf{Y}$ are generated from the unknown probability distribution P^0 , denoted $\mathbf{Y}_n^0 = (Y_{n,1}^0, \dots, Y_{n,R}^0)$, $n = 1, \dots, N^0$. For the independent sample case, suppose N^1 independent realizations of $\mathbf{Y}$ are generated from the unknown probability distribution P^1 , denoted $\mathbf{Y}_n^1 = (Y_{n,1}^1, \dots, Y_{n,R}^1)$, $n = 1, \dots, N^1$. For the paired sample case, each of the $N^1 = N^0$ independent realizations $\mathbf{Y}_n^1$ are from the unknown conditional probability distribution of $\mathbf{Y}$ in group 1 given $\mathbf{Y}^0 = \mathbf{Y}_n^0$, denoted $P^{1|\mathbf{Y}_n^0}$, $n = 1, \dots, N^1$. Finally, denote by W a binary variable indicating membership to group 0 ($W = 0$) or group 1 ($W = 1$).

The complete data-generating mechanism is denoted by the collection of distributions $P = (P^0, P^1)$ (independent sample case) or $P = \{P^0, (P^{1|\mathbf{y}})_{\mathbf{y}}\}$ (paired sample case), where $\mathbf{y}$ is a potential value for $\mathbf{Y}$. The collection P is considered to belong to a family of probability distributions, say Ω ; the model Ω may be nonparametric, parametric, or semiparametric. Only the nonparametric model is considered in this paper.

3 Comparing Joint Distributions

Note that $P^1 = \int P^{1|\mathbf{Y}^0=\mathbf{y}} dP^0(\mathbf{y})$. For both the paired and independent sample case, the main goal of this paper is to estimate the parsimonious set of differences between P^0 and P^1 .

Let $y_{r,1}, \dots, y_{r,L_r}$ denote the possible values of variable Y_r , $r = 1, \dots, R$, with $L_r \geq 2$. Define $J_{r;\ell} = I(Y_r = y_{r,\ell})$, $\ell = 1, \dots, L_r$, $r = 1, \dots, R$, $I(\cdot)$ the indicator function. More generally, the variable indicating the k th-order variable/level combination, $R \geq k \geq 1$, is denoted by $J_{r_1;\ell_1, \dots, r_k;\ell_k} = J_{r_1;\ell_1} \times \dots \times J_{r_k;\ell_k}$.

A crude method for comparing P^0 and P^1 is one that attempts to identify all those R th-order variable/level combinations for which $P(J_{r_1;\ell_1, \dots, r_R;\ell_R} = 1|W = 0)$ does not equal $P(J_{r_1;\ell_1, \dots, r_R;\ell_R} = 1|W = 1)$.

However, if $P(J_{r_1;\ell_1, \dots, r_R;\ell_R} = 1|W)$ can be written as the same product of prevalences of combinations of order $R - 1$ or less for $W = 0, 1$, then the statement $P(J_{r_1;\ell_1, \dots, r_R;\ell_R} = 1|W = 0) \neq P(J_{r_1;\ell_1, \dots, r_R;\ell_R} = 1|W = 1)$ is redundant in the sense that the difference is completely explained in terms of prevalences of lower-order combinations. For example, suppose $P(J_{r_1;\ell_1, r_2;\ell_2} = 1|W = 0) \neq P(J_{r_1;\ell_1, r_2;\ell_2} = 1|W = 1)$, but that

$$P(J_{r_1;\ell_1, r_2;\ell_2} = 1|W) = P(J_{r_1;\ell_1} = 1|W)P(J_{r_2;\ell_2} = 1|W)$$

for $W = 0, 1$. Then the difference between groups in $P(J_{r_1;\ell_1,r_2;\ell_2} = 1|W)$ is completely driven by a difference between groups with respect to $P(J_{r_1;\ell_1} = 1|W)$ and/or $P(J_{r_2;\ell_2} = 1|W)$.

For third-order variable/level combinations, for each group W , there are three ways for $P(J_{r_1;\ell_1,r_2;\ell_2,r_3;\ell_3} = 1|W)$ to be written as a product of prevalences of lower-order combinations, namely,

- i) $P(J_{r_1;\ell_1} = 1|W)P(J_{r_2;\ell_2,r_3;\ell_3} = 1|W)$, or
- ii) $P(J_{r_2;\ell_2} = 1|W)P(J_{r_1;\ell_1,r_3;\ell_3} = 1|W)$, or
- iii) $P(J_{r_3;\ell_3} = 1|W)P(J_{r_1;\ell_1,r_2;\ell_2} = 1|W)$.

Case i) states that the event $J_{r_1;\ell_1} = 1$ is conditionally independent of the event $J_{r_2;\ell_2,r_3;\ell_3} = 1$ given W , similarly for case ii) and iii). The seven ways for $P(J_{r_1;\ell_1,r_2;\ell_2,r_3;\ell_3,r_4;\ell_4} = 1|W)$ to be written as a product of prevalences of lower-order combinations are given in the Appendix.

A parsimonious, albeit idealistic, method for comparing P^0 and P^1 is one that first attempts to identify those variable/level pairs whose first-order (marginal) probabilities, $P(J_{r;\ell} = 1|W)$, differ between groups. The method then would attempt to identify those second- thru R th-order variable/level combinations whose prevalence differs between groups and cannot be written as the same product of prevalences of lower-order combinations for both groups $W = 0, 1$. An obvious challenge with developing such an ideal method is the consideration of all possible second- thru R th-order combinations. In any case, it is clear that this ideal method is parsimonious in that it does not identify redundant differences between P^0 and P^1 .

A first step in constructing a method for parsimoniously comparing P^0 and P^1 that is feasibly implemented can be to fix the largest considered combination of variable/level pairs at some value K , with $R \geq K \geq 1$, e.g. $K = 3$ or 4. Note that K is regarded as the maximal order of non-negligible dependency among variables. If in fact it is true that the prevalence of each combination of order $K + 1$ and larger can be written as the same product of prevalences of lower-order combinations for $W = 0, 1$, then a full parsimonious comparison between P^0 and P^1 only requires consideration of all first- thru K th-order combinations. When the prevalence of one or more combinations of order $K + 1$ or larger differs between groups and cannot be written as the same product of prevalences of lower-order combinations for $W = 0, 1$, limiting the largest considered combination at order K approximates a full parsimonious comparison between P^0 and P^1 . Although the value of K is independent of sample size, because of the curse of dimensionality, the practical range for K

certainly increases with sample size. The value of K may be obvious in a particular subject matter of interest, or determined by the computational and/or data resources at hand.

Now, even with limiting the number of variable/level pairs in each combination at K , the resulting computational demand of a parsimonious comparison method may still be too strong. Note that, for each group W , the total number of first-order variable/level pairs is $\sum_{r=1}^R L_r$; the total number of first- and second-order variable/level combinations is $\sum_{r=1}^R L_r \{(\sum_{s=r+1}^R L_s) + 1\}$; and so-on. Thus the total number of combinations rapidly increases with the combination order. Consider the following motivation to approximating the set of all second- thru K th-order combinations. Suppose some k th-order combination has different prevalence between groups, $K \geq k \geq 2$; it is not necessarily the case that each of the k marginal (first-order) variable/level pairs has different prevalence between groups. In fact, it could be the case that none of the k variable/level pairs has a marginal group effect. Thus, those k th-order combinations for which at least $\max\{(k - K_0), 0\}$ variable/level pairs are marginally associated with group membership, $K \geq K_0 \geq 0$, represent a subset of all possible k th-order combinations and, when $K_0 < k$, constitutes an approximation to this set. For each k th-order combination in this subset, at most K_0 of the k variable/level pairs have no marginal group association. In a sense K_0 can be thought of as a degrees of freedom; $K_0 = K$ corresponds to no approximation, while $K_0 = 0$ considers only those k th-order combinations for which all k variable/level pairs have a marginal association with group membership.

4 Null Hypotheses

The value of K is considered fixed and for ease of exposition suppressed in the notation below. Comparing P^0 and P^1 requires testing two types of multiple null hypotheses: those corresponding to no group effect on the prevalence of variable/level combinations, and those corresponding to conditional independence of events defining combinations given W .

4.1 Equal prevalence

For $K \geq k \geq 1$, define

$$H_{r_1; \ell_1, \dots, r_k; \ell_k}^{(k)} : P(J_{r_1; \ell_1, \dots, r_k; \ell_k} = 1 | W = 0) = P(J_{r_1; \ell_1, \dots, r_k; \ell_k} = 1 | W = 1),$$

with $r_i \in (1, \dots, R)$, $r_i \neq r_j$ ($i \neq j$) and $\ell_i \in (1, \dots, L_{r_i})$. Each alternative hypotheses can be either two-sided or one-sided, and is denoted by $\overline{H}_{r_1; \ell_1, \dots, r_k; \ell_k}^{(k)}$.

To calculate a summary statistic corresponding to each $H_{r_1;\ell_1,\dots,r_k;\ell_k}^{(k)}$, let $J_{r_1;\ell_1,\dots,r_k;\ell_k}^{n,W}$ denote the value of $J_{r_1;\ell_1,\dots,r_k;\ell_k}$ for the n th observation, $n = 1, \dots, N^W$, in group $W = 0, 1$. Define the empirical probability

$$\hat{P}(J_{r_1;\ell_1,\dots,r_k;\ell_k} = 1|W) = \sum_{n=1}^{N^W} J_{r_1;\ell_1,\dots,r_k;\ell_k}^{n,W} / N^W.$$

A summary statistic can be defined as the difference in proportions

$$\hat{S}_{r_1;\ell_1,\dots,r_k;\ell_k} = \hat{P}(J_{r_1;\ell_1,\dots,r_k;\ell_k} = 1|W = 1) - \hat{P}(J_{r_1;\ell_1,\dots,r_k;\ell_k} = 1|W = 0). \quad (1)$$

A more computationally intensive summary statistic arises from calculating Fisher's exact test on the 2×2 table with first row entries $\sum_{n=1}^{N^0} J_{r_1;\ell_1,\dots,r_k;\ell_k}^{n,0}$ and $N^0 - \sum_{n=1}^{N^0} J_{r_1;\ell_1,\dots,r_k;\ell_k}^{n,0}$ and second row entries $\sum_{n=1}^{N^1} J_{r_1;\ell_1,\dots,r_k;\ell_k}^{n,1}$ and $N^1 - \sum_{n=1}^{N^1} J_{r_1;\ell_1,\dots,r_k;\ell_k}^{n,1}$. Let $\theta(r_1; \ell_1, \dots, r_k; \ell_k)$ denote the corresponding exact one- or two-sided p-value. Note that $1 \geq \theta(r_1; \ell_1, \dots, r_k; \ell_k) > 0$. An associated summary statistic is

$$\tilde{S}_{r_1;\ell_1,\dots,r_k;\ell_k} = \Phi^{-1}[\max\{\epsilon, 1 - \theta(r_1; \ell_1, \dots, r_k; \ell_k)\}], \quad (2)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and ϵ is an arbitrary positive number. Although $\tilde{S}_{r_1;\ell_1,\dots,r_k;\ell_k}$ requires more computational time than $\hat{S}_{r_1;\ell_1,\dots,r_k;\ell_k}$, its use may lead to more efficient inference procedures, see section 5.

4.2 Conditional independence

Corresponding to each second-order variable/level combination, define the null hypothesis

$$Q_{r_1;\ell_1,r_2;\ell_2}^{W,1} : P(J_{r_1;\ell_1,r_2;\ell_2} = 1|W) = P(J_{r_1;\ell_1} = 1|W)P(J_{r_2;\ell_2} = 1|W),$$

for $r_i \in (1, \dots, R)$, $r_1 \neq r_2$, $\ell_i \in (1, \dots, L_{r_i})$ and $W = 0, 1$.

Note that when $Q_{r_1;\ell_1,r_2;\ell_2}^{W,1}$ is true, the following three statements are also true:

- i) $P(J_{r_1;\ell_1} = 0, J_{r_2;\ell_2} = 1|W) = P(J_{r_1;\ell_1} = 0|W)P(J_{r_2;\ell_2} = 1|W)$,
- ii) $P(J_{r_1;\ell_1} = 1, J_{r_2;\ell_2} = 0|W) = P(J_{r_1;\ell_1} = 1|W)P(J_{r_2;\ell_2} = 0|W)$, and
- iii) $P(J_{r_1;\ell_1} = 0, J_{r_2;\ell_2} = 0|W) = P(J_{r_1;\ell_1} = 0|W)P(J_{r_2;\ell_2} = 0|W)$.

Thus, a two-sided p-value corresponding to $Q_{r_1; \ell_1, r_2; \ell_2}^{W,1}$, denoted $\theta_1(r_1; \ell_1, r_2; \ell_2|W)$, can be obtained by calculating Fisher's exact test on the 2×2 table with first row entries $N^W - \sum_{n=1}^{N^W} J_{r_2; \ell_2}^{n,W} - J_{r_1; \ell_1}^{n,W} + J_{r_1; \ell_1, r_2; \ell_2}^{n,W}$ and $\sum_{n=1}^{N^W} J_{r_2; \ell_2}^{n,W} - J_{r_1; \ell_1, r_2; \ell_2}^{n,W}$, and second row entries $\sum_{n=1}^{N^W} J_{r_1; \ell_1}^{n,W} - J_{r_1; \ell_1, r_2; \ell_2}^{n,W}$ and $\sum_{n=1}^{N^W} J_{r_1; \ell_1, r_2; \ell_2}^{n,W}$. A summary statistic corresponding to $Q_{r_1; \ell_1, r_2; \ell_2}^{W,1}$ is

$$\tilde{S}_{r_1; \ell_1, r_2; \ell_2}^{W,1} = \Phi^{-1}[\max\{\epsilon, 1 - \theta_1(r_1; \ell_1, r_2; \ell_2|W)\}]. \quad (3)$$

The reason for recommending Fisher's exact test here is because considerably less data is used compared to the test for equal prevalence; in addition, one of the row sums in this 2×2 table may be small, corresponding to sparse levels.

Now, corresponding to each third-order variable/level combination are three null hypotheses of interest in each group $W = 0, 1$. These are

$$Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,1} : P(J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3} = 1|W) = P(J_{r_1; \ell_1} = 1|W)P(J_{r_2; \ell_2, r_3; \ell_3} = 1|W),$$

$$Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,2} : P(J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3} = 1|W) = P(J_{r_2; \ell_2} = 1|W)P(J_{r_1; \ell_1, r_3; \ell_3} = 1|W),$$

$$Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,3} : P(J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3} = 1|W) = P(J_{r_3; \ell_3} = 1|W)P(J_{r_1; \ell_1, r_2; \ell_2} = 1|W),$$

for $r_i \in (1, \dots, R)$, $r_i \neq r_j$, ($i \neq j$) and $\ell_i \in (1, \dots, L_{r_i})$.

When $Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,1}$ is true, the following three statements are also true:

$$\text{i)} \quad P(J_{r_1; \ell_1} = 0, J_{r_2; \ell_2, r_3; \ell_3} = 1|W) = P(J_{r_1; \ell_1} = 0|W)P(J_{r_2; \ell_2, r_3; \ell_3} = 1|W),$$

$$\text{ii)} \quad P(J_{r_1; \ell_1} = 1, J_{r_2; \ell_2, r_3; \ell_3} = 0|W) = P(J_{r_1; \ell_1} = 1|W)P(J_{r_2; \ell_2, r_3; \ell_3} = 0|W), \text{ and}$$

$$\text{iii)} \quad P(J_{r_1; \ell_1} = 0, J_{r_2; \ell_2, r_3; \ell_3} = 0|W) = P(J_{r_1; \ell_1} = 0|W)P(J_{r_2; \ell_2, r_3; \ell_3} = 0|W).$$

Similarly for $Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,2}$ and $Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,3}$.

A two-sided p-value on $Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,1}$, denoted $\theta_1(r_1; \ell_1, r_2; \ell_2, r_3; \ell_3|W)$, can be obtained by calculating Fisher's exact test on the 2×2 table with first row entries $N^W - \sum_{n=1}^{N^W} J_{r_2; \ell_2, r_3; \ell_3}^{n,W} - J_{r_1; \ell_1}^{n,W} + J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{n,W}$ and $\sum_{n=1}^{N^W} J_{r_2; \ell_2, r_3; \ell_3}^{n,W} - J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{n,W}$, and second row entries $\sum_{n=1}^{N^W} J_{r_1; \ell_1}^{n,W} - J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{n,W}$ and $\sum_{n=1}^{N^W} J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{n,W}$. Similarly, a two-sided p-value, denoted $\theta_i(r_1; \ell_1, r_2; \ell_2, r_3; \ell_3|W)$, can be calculated for $Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,i}$, $i = 2, 3$. The summary statistic corresponding to $Q_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,i}$ is defined by

$$\tilde{S}_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3}^{W,i} = \Phi^{-1}[\max\{\epsilon, 1 - \theta_i(r_1; \ell_1, r_2; \ell_2, r_3; \ell_3|W)\}], \quad i = 1, 2, 3. \quad (4)$$

A similar set of null hypotheses and summary statistics correspond to fourth-order and larger variable/level combinations.

5 Bootstrap-based Stepdown Inference

Suppose $H_1, \dots, H_B$ is a collection of null hypotheses corresponding to arbitrary parameters of P . Let S_b denote an asymptotically linear summary statistic corresponding to null hypothesis H_b , $b = 1, \dots, B$. When using $(S_b)_{b=1}^B$ as a basis to simultaneously test $(H_b)_{b=1}^B$, van der Laan et al. (2004a) provide nonparametric bootstrap-based stepdown algorithms that asymptotically bound the corresponding FWER above by the target level α for arbitrary data-generating distributions. Presented below is the minP algorithm of van der Laan et al. (2004a) with the following modifications. In the attempt to gain efficiency, unadjusted two-sided p-values are calculated as twice the minimum of left- and right-tailed probabilities of the summary statistic, instead of twice the right-tailed probability of the absolute summary statistic. Also, one minus the unadjusted p-value is used as the test statistic so that larger values represent evidence against the null hypothesis.

Inference differs between the independent and paired sample case only in the way that bootstrap samples are generated. For the independent sample case, each bootstrap sample consists of N^0 independent draws with replacement from $(\mathbf{Y}_n^0)_{n=1}^{N^0}$ and N^1 independent draws with replacement from $(\mathbf{Y}_n^1)_{n=1}^{N^0}$. For the paired sample case, each bootstrap sample consists of $N = N^0 = N^1$ independent draws with replacement of pairs from $(\mathbf{Y}_n^0, \mathbf{Y}_n^1)_{n=1}^N$. In either case, the first step is to generate M independent nonparametric bootstrap samples, with M as large as computationally feasible.

Corresponding to each bootstrap sample, calculate the B -vector of summary statistics and denote it by $(S_{m,1}^*, \dots, S_{m,B}^*)$, for $m = 1, \dots, M$. Denote the vector of summary statistics calculated on the original observed data by $(S_{0,1}, \dots, S_{0,B})$. Calculate $\hat{E}(S_b) = M^{-1} \sum_{m=1}^M S_{m,b}^*$ and form the null realizations $Z_{m,b}^* = \{S_{m,b}^* - \hat{E}(S_b)\}$, $m = 1, \dots, M$, $b = 1, \dots, B$.

For summary statistics taking the form (1), the finite-sample marginal distribution of Z_b^* is not necessarily identical or symmetric for all $b = 1, \dots, B$. In the interest of efficiency, for those H_b with a two-sided alternative hypothesis, calculate

$$F_{m,b}^* = 2 \times \min \begin{cases} M^{-1} \sum_{m'=1}^M I(Z_{m',b}^* \geq Z_{m,b}^*) \\ M^{-1} \sum_{m'=1}^M I(Z_{m',b}^* \leq Z_{m,b}^*) \\ 1/2 \end{cases},$$

and for one-sided alternatives, $F_{m,b}^* = M^{-1} \sum_{m'=1}^M I(Z_{m',b}^* \geq (\leq) Z_{m,b}^*)$, $m = 1, \dots, M$, $b = 1, \dots, B$. Finally, to keep the convention that larger test-statistics represent greater evidence against H_b , calculate $T_{m,b}^* = 1 - F_{m,b}^*$, $m = 1, \dots, M$, $b = 1, \dots, B$. For each summary statistic $S_{0,b}$, corresponding to the observed

data, calculate

$$F_{0,b} = 2 \times \min \begin{cases} M^{-1} \sum_{m=1}^M I(Z_{m,b}^* \geq S_{0,b}) \\ M^{-1} \sum_{m=1}^M I(Z_{m,b}^* \leq S_{0,b}) \\ 1/2 \end{cases}$$

for two-sided alternatives, similarly for one-sided alternatives, and $T_{0,b} = 1 - F_{0,b}$, $b = 1, \dots, B$.

For summary statistics of the form (2), (3) and (4), set $T_{m,b}^* = Z_{m,b}^*$ and $T_{0,b} = S_{0,b}$, $m = 1, \dots, M$, $b = 1, \dots, B$. For either type of summary statistic (1) or (2), the stepdown algorithm below is exactly the same.

Put the $(T_{0,b})_{b=1}^B$ in ascending order to obtain $T_{0,u(1)} \leq \dots \leq T_{0,u(B)}$ and correspondingly reorder the columns of the matrix $(T_{m,b}^*)_{b=1}^B$ and the null hypotheses $(H_b)_{b=1}^B$. The stepdown algorithm is as follows.

1. Set $s = 0$.
2. Calculate $\bar{T}_m^*(s) = \max(T_{m,u(b)}^*)_{b=1}^{B-s}$
3. Calculate $\hat{\pi}_s = M^{-1} \sum_{m=1}^M I\{\bar{T}_m^*(s) \geq T_{0,u(B-s)}\}$
4. If $\alpha \geq \hat{\pi}_s$ then reject $H_{u(B-s)}$, set $s = s + 1$ and return to step 2.; otherwise, if $\hat{\pi}_s > \alpha$ then do not reject $H_{u(B-s)}, \dots, H_{u(1)}$ and stop the algorithm.

An estimator for the multiple testing-adjusted p-value on $H_{u(B-s)}$, $s = 0, \dots, B-1$, is the smallest FWER such that $H_{u(B-s)}$ is rejected by the above algorithm. To approximate this estimator, one could execute the algorithm above for multiple values of the FWER. On the other hand, an upper bound for an estimate of the adjusted p-value on $H_{u(B-s)}$ can easily be obtained from the first step of the algorithm as $\hat{\pi}_{u(B-s)}^* = M^{-1} \sum_{m=1}^M I\{\bar{T}_m^*(0) \geq T_{0,u(B-s)}\}$. If the stepdown algorithm is conducted at FWER $\hat{\pi}_{u(B-s)}^*$ then $H_{u(B-s)}$ will be rejected. Note that the proposed parsimonious comparison method requires simultaneously testing two types of multiple null hypotheses, each at a specific FWER. For each combination in the estimated set of combinations claimed to have a non-redundant group effect, Section 6 describes how to estimate an overall adjusted probability, adjusted for simultaneously testing both types of null hypotheses, that the combination either has equal prevalence between groups or corresponds to a redundant group effect.

Note that, for summary statistics (2), (3) and (4), the distribution of S_b is known exactly when H_b is true. As a result, the null realizations $Z_{m,b}^*$, $m = 1, \dots, M$, $b = 1, \dots, B$, can instead be calculated using the quantile

function-based null distribution of van der Laan and Hubbard (2005). The advantage of using this quantile function-based null distribution is a potential reduction in bias and gain in efficiency compared to the stepdown method above, most notably when the sampling distribution of S_b is skewed. However, using the null distribution of van der Laan and Hubbard (2005) in this setting increases the already large computational burden. For example, for summary statistics corresponding to equal prevalence null hypotheses, Fisher's exact test needs to be calculated instead of the less computationally demanding difference in proportions. For the conditional independence null hypotheses, the added computational burden would primarily arise from having to calculate the empirical distribution of $(S_{m,b}^*)_{m=1}^M$, as opposed to simply the mean of $(S_{m,b}^*)_{m=1}^M$, which is proposed in this paper.

6 Comparison Methods

Consider the following method for parsimoniously comparing P^0 and P^1 . Set $K_0 = K$ and form all unique variable/level combinations of order $K, K - 1, \dots, 1$. Simultaneously test the equal prevalence null hypotheses

$$\mathcal{H} = \{(H_{r;\ell}^{(1)}), \dots, (H_{r_1;\ell_1, \dots, r_K;\ell_K}^{(K)})\}$$

at target FWER α using the procedure in Section 5. Now, denote $\hat{\mathcal{G}}_{1,\alpha}$ the set of first-order variable/level pairs corresponding to the rejected $(H_{r;\ell}^{(1)})$ and $\hat{\mathcal{G}}_{2,\alpha}$ the set of second-order and larger variable/level combinations corresponding to the rejected $\{(H_{r_1;\ell_1, r_2;\ell_2}^{(2)}), \dots, (H_{r_1;\ell_1, \dots, r_K;\ell_K}^{(K)})\}$. For a variable/level combination of order k , $k = 2, \dots, K$, let $q(k)$, denote the number of conditional independence null hypotheses to be tested in each group W , e.g. $q(2) = 1, q(3) = 3, q(4) = 7$. Simultaneously test those conditional independence null hypotheses from

$$\mathcal{Q} = \{(Q_{r_1;\ell_1, r_2;\ell_2}^{0,1}), \dots, (Q_{r_1;\ell_1, \dots, r_K;\ell_K}^{0,q(K)}), (Q_{r_1;\ell_1, r_2;\ell_2}^{1,1}), \dots, (Q_{r_1;\ell_1, \dots, r_K;\ell_K}^{1,q(K)})\}$$

that correspond to the combinations in $\hat{\mathcal{G}}_{2,\alpha}$ at target FWER β using the procedure in Section 5. Those combinations $J_{r_1;\ell_1, \dots, r_k;\ell_k}$, $K \geq k \geq 2$, for which $Q_{r_1;\ell_1, \dots, r_k;\ell_k}^{0,i}$ and $Q_{r_1;\ell_1, \dots, r_k;\ell_k}^{1,i}$ are not rejected for one or more i , $i \in \{1, \dots, q(k)\}$, are removed from $\hat{\mathcal{G}}_{2,\alpha}$; let $\hat{\mathcal{G}}'_{2,\alpha,\beta}$ denote this reduced set. The estimate of the set of variable/level combinations corresponding to the false and non-redundant null hypotheses from $\mathcal{H}$ is $\hat{\mathcal{G}}_{0,\alpha,\beta} = (\hat{\mathcal{G}}_{1,\alpha}, \hat{\mathcal{G}}'_{2,\alpha,\beta})$.

The set $\hat{\mathcal{G}}_{0,\alpha,\beta}$ is the estimate of the parsimonious set of components, up to and including order K , distinguishing P^0 and P^1 . The probability that

$\hat{\mathcal{G}}_{0,\alpha,\beta}$ contains one or more combinations that either have the same prevalence between groups, or correspond to redundant differences, is asymptotically bounded above by $1 - \{(1 - \alpha)(1 - \beta)\}$.

The direct testing method above will in general require far too much computer memory to be computationally feasible, even if K_0 were chosen such that $K_0 < K$. A more computationally practical method is one that breaks this large-scale inference problem into multiple smaller-scale problems—requiring extensive processing power instead of memory—then pieces the multiple results together into one final result.

6.1 A computationally efficient method

Using the procedure in Section 5, simultaneously test the first-order equal prevalence null hypotheses $(H_{r;\ell}^{(1)})$ at target FWER α . Denote the set of variable/level pairs corresponding to the rejected $(H_{r;\ell}^{(1)})$ by $\mathcal{C}_\alpha^1$, and the set corresponding those $(H_{r;\ell}^{(1)})$ that were not rejected by $\mathcal{C}_\alpha^0$.

For a given value of K_0 , let $M' = \binom{\#(\mathcal{C}_\alpha^0)}{K_0}$, the total number of ways to select K_0 elements from the set $\mathcal{C}_\alpha^0$, where $\#(\cdot)$ is the cardinality function. Denote each of the M' subsets of K_0 variable/level pairs from $\mathcal{C}_\alpha^0$ by $c_\alpha^0(m)$, $m = 1, \dots, M'$. For each $c_\alpha^0(m)$, $m = 1, \dots, M'$:

1. Form all unique second- thru K th-order variable/level combinations from $\{\mathcal{C}_\alpha^1, c_\alpha^0(m)\}$; denote this set by $\mathcal{G}_0(m)$
2. Remove from $\mathcal{G}_0(m)$ all combinations that either do not contain a variable/level pair from $c_\alpha^0(m)$ or have been previously considered at some iteration $m' < m$; denote the resulting set by $\mathcal{G}_1(m)$
3. Using the procedure in Section 5, simultaneously test at FWER α the equal prevalence null hypotheses from $\{(H_{r_1;\ell_1, r_2;\ell_2}^{(2)}, \dots, (H_{r_1;\ell_1, \dots, r_K;\ell_K}^{(K)})\}$ that correspond to the combinations in $\mathcal{G}_1(m)$; denote the set of combinations corresponding to those rejected hypotheses by $\mathcal{G}_{2,\alpha}(m)$
4. Simultaneously test the subset of $\mathcal{Q}$ corresponding to the combinations in $\mathcal{G}_{2,\alpha}(m)$ at target FWER β using the procedure in Section 5. Denote by $\mathcal{G}'_{2,\alpha,\beta}(m)$ those combinations for which the corresponding $Q_{r_1;\ell_1, \dots, r_k;\ell_k}^{W,i}$ are rejected for $W = 0$ and/or $W = 1$, for all $i = 1, \dots, q(k)$.

Now, set $m = 0$ and execute steps 1., 3. and 4. above with $\{\mathcal{C}_\alpha^1, c_\alpha^0(m)\}$ replaced by $\mathcal{C}_\alpha^1$ in step 1.; let $\mathcal{G}_1(0) = \mathcal{G}_0(0)$ and $\mathcal{G}'_{2,\alpha,\beta}(0)$ denote the output

from step 4. Using the procedure in Section 5, simultaneously test at target FWER α the subset of hypotheses from $\mathcal{H}$ that corresponds to the unique variable/level combinations in $\{\mathcal{C}_\alpha^1, \mathcal{G}'_{2,\alpha,\beta}(0), \dots, \mathcal{G}'_{2,\alpha,\beta}(M')\}$; denote the combinations corresponding to those rejected hypotheses by $\hat{\mathcal{G}}_\alpha$. Denote the subset of $\hat{\mathcal{G}}_\alpha$ corresponding to first-order variable/level pairs by $\hat{\mathcal{G}}_{1,\alpha}$ and to second-order and larger variable/level combinations by $\hat{\mathcal{G}}_{2,\alpha}$. Execute step 4. above with $\mathcal{G}_{2,\alpha}(m)$ replaced by $\hat{\mathcal{G}}_{2,\alpha}$, let $\hat{\mathcal{G}}'_{2,\alpha,\beta}$ denote the output. The final estimate of the set of combinations that have a non-redundant, different prevalence between groups is $\hat{\mathcal{G}}_{0,\alpha,\beta} = (\hat{\mathcal{G}}_{1,\alpha}, \hat{\mathcal{G}}'_{2,\alpha,\beta})$. The probability that $\hat{\mathcal{G}}_{0,\alpha,\beta}$ contains one or more combinations that have equal prevalence between groups, or correspond to a redundant difference, is asymptotically bounded above by $1 - \{(1 - \alpha)(1 - \beta)\}$.

When the set $\mathcal{C}_\alpha^1$ is so large that the above procedure requires too much computer memory, one can replace $\mathcal{C}_\alpha^1$ by say $\mathcal{C}_\alpha^\dagger$, a relatively small subset of $\mathcal{C}_\alpha^1$ corresponding to the most significant marginal group effects. Then, $\mathcal{C}_\alpha^0$ would be defined as all first-order variable/level pairs not in $\mathcal{C}_\alpha^\dagger$. As a result, the required computer memory is decreased but M' is increased. Note that $\mathcal{C}_\alpha^1$ would be used, not $\mathcal{C}_\alpha^\dagger$, when calculating the final estimate $\hat{\mathcal{G}}_{0,\alpha,\beta}$. Also note that the number of k th-order combinations for which at least $\max\{(k - K_0), 0\}$ variable/level pairs are in $\mathcal{C}_\alpha^1$ is larger than for $\mathcal{C}_\alpha^\dagger$. Thus, using $\mathcal{C}_\alpha^1$ instead of $\mathcal{C}_\alpha^\dagger$ results in a closer approximation to the full set of all k th-order combinations.

The steps above leading to $\{\mathcal{C}_\alpha^1, \mathcal{G}'_{2,\alpha,\beta}(0), \dots, \mathcal{G}'_{2,\alpha,\beta}(M')\}$ can be viewed as a dimension reduction technique; asymptotic control of the overall FWER is preserved at the final step when calculating $\hat{\mathcal{G}}_\alpha$ and $\hat{\mathcal{G}}'_{2,\alpha,\beta}$.

Suppose one wanted to estimate, for each combination in $\hat{\mathcal{G}}_{0,\alpha,\beta}$, an overall adjusted probability that the combination either has the same prevalence between groups, or corresponds to a redundant group effect. Such a probability would be adjusted for simultaneously testing both types of null hypotheses. First, note that all elements of $\hat{\mathcal{G}}_{1,\alpha}$ are first-order combinations; as such, no conditional independence null hypotheses apply. Thus, for each variable/level pair in $\hat{\mathcal{G}}_{1,\alpha}$, an estimated overall adjusted probability that the variable/level pair has equal prevalence between groups is the corresponding multiple testing-adjusted p-value, as described in Section 5, arising from simultaneously testing the equal prevalence null hypotheses. However, the combinations in $\hat{\mathcal{G}}'_{2,\alpha,\beta}$ are simultaneously tested for equal prevalence at FWER α and conditional independence at FWER β . For a k th-order combination in $\hat{\mathcal{G}}'_{2,\alpha,\beta}$, let $\hat{\pi}^*(0) \leq \alpha$ denote the corresponding estimated adjusted p-value from simultaneously testing the equal prevalence null hypotheses. Recall that $q(k)$ conditional independence null hypotheses are tested for this combination in

each group $W = 0, 1$. In order for this combination to be in $\hat{\mathcal{G}}'_{2,\alpha,\beta}$, for each of the $q(k)$ pairs of null hypotheses, at least one of the two hypotheses was rejected at FWER β . Therefore, for each of the corresponding $q(k)$ pairs of conditional independence null hypotheses, take the smaller of the two adjusted p-value estimates (from simultaneously testing the conditional independence null hypotheses) and let $\hat{\pi}^*(1|\alpha) \leq \beta$ denote the largest of these $q(k)$ minima. The value $\hat{\pi}^*(1|\alpha)$ is the estimated adjusted p-value from testing the conditional independence null hypotheses. Note that this probability is conditional on the target FWER for the equal prevalence null hypotheses, α . An estimator for the overall adjusted probability that this k th-order combination has either equal prevalence between groups or corresponds to a redundant group effect is $1 - [\{1 - \hat{\pi}^*(0)\}\{1 - \hat{\pi}^*(1|\alpha)\}]$. If the equal prevalence null hypotheses are tested at FWER $\hat{\pi}^*(0)$ and the conditional independence null hypotheses at FWER $\hat{\pi}^*\{1|\hat{\pi}^*(0)\}$, then this combination would be declared to have a non-redundant group effect.

7 Simulations

Take $R = 9$ and define Y_r as a binary variable, $r = 1, \dots, 9$. For the independent sample case, the probability distributions (P^0, P^1) are defined in the Appendix. For the paired sample case, take P^0 as for the independent sample case and define $P^1|Y_n^0 \equiv P^1|Y_{n,1}^0$, with $P(Y_1^1 = 1|Y_{n,1}^0 = 0) = 1/5$ and $P(Y_1^1 = 1|Y_{n,1}^0 = 1) = 2/5$; the remaining Y_r , $r = 2, \dots, 9$, are generated as defined by P^1 in the independent sample case. Note that P^1 corresponding to the paired and independent sample case are equivalent here.

Note the following. The marginal (univariate) probability distributions of Y_5 , Y_7 and Y_8 differ between groups. All differences between groups with respect to the prevalence of third-order and larger variable/level combinations are redundant. All second-order combinations arising from (Y_1, Y_2) have a non-redundant difference in prevalence between groups, however, there is no group effect on the corresponding marginal distributions; similarly for (Y_3, Y_4) . All second-order combinations arising from (Y_5, Y_6) have a non-redundant group difference in prevalence, but, only one of the two variable/level pairs in each corresponding combination has a marginal group effect. Each second-order combination arising from (Y_7, Y_8) has a non-redundant group difference in prevalence, and both variable/level pairs in each combination are marginally associated with group membership.

The simulations were conducted as follows. For each of three sample size configurations, (N^0, N^1) , and for both the independent and paired sample

case, a random sample $\{(\mathbf{Y}_n^0)_{n=1}^{N^0}, (\mathbf{Y}_n^1)_{n=1}^{N^1}\}$ was generated from the respective P (independent or paired sample case) a total of 1000 independent times. At each of the 1000 simulation iterations, for the independent sample case, $Y_{n,1}^W$ was randomly generated from $P(Y_1^W)$, then $Y_{n,2}^W$ was randomly generated from $P(Y_2^W|Y_{n,1}^W)$, and so on, $W = 0, 1$. Similarly for the paired sample case, except that the distribution for Y_1^1 is conditioned on $Y_{n,1}^0$. For each setting and each simulation iteration, $M = 1000$ independent bootstrap samples were used for each inference procedure. For all cases, $\alpha = \beta = 0.05$, $K = 3$, and $K_0 = 2$; summary statistic (1) was used for equal prevalence null hypotheses, and (3) and (4) were used for conditional independence null hypotheses. Note that a full parsimonious comparison between P^0 and P^1 here requires only $K = K_0 = 2$.

Table 1 presents empirical FWERs, and rejection percentages for the six variable/level pairs whose prevalence differs marginally between groups and the sixteen second-order combinations with a non-redundant group effect. An overall family-wise error is made when a combination from $\hat{\mathcal{G}}_{0,\alpha,\beta}$ either (i) has equal prevalence between groups, or (ii) has a prevalence that can be written as the same product of prevalences of lower-order combinations for both groups $W = 0, 1$. For all configurations, the empirical FWER is below the nominal value of $1 - \{(1 - \alpha)(1 - \beta)\} = 0.0975$. Note that, with 1000 independent simulation iterations, the normal theory-based 95% two-sided confidence interval for a nominal FWER of 9.75% extends $\pm 1.84\%$ of the empirical FWER%. As expected, for both the independent and paired sample case, empirical powers increase with increasing sample size and, for a fixed sample size, are in general greater for larger effect sizes.

The simulations were executed using Matlab version 7.1 and with 64-bit Intel Xeon 3.2 GHz processors. To give an example of the computational demand, the total number of CPU hours required for the 1000 simulation iterations in the paired sample case with 50, 100 and 250 observations per group were 71, 106 and 116, respectively. The corresponding numbers for the independent sample case were 54, 81 and 122 hours. These iterations were split over 10 CPUs and run in parallel, e.g. the effective time for the paired sample case and 50 observations per group was 7.1 hours.

8 Analysis of ACTG 398

At baseline, 450 subjects had an HIV-1 genotype sequence recorded that consisted of all 99 codons from the PR region and 300 codons from the RT region. When sequencing the HIV-1 genome, it is possible that the amino acid value

at a given codon is not the same for all virus in the sample; when there is no majority amino acid value at a codon, the value reported by the assay is a mixture of the values observed in the sample. For the purpose of this analysis, all mixtures were labeled 'other', resulting in 21 possible amino acid levels at each codon.

A total of $N^0 = N^1 = 302$ subjects experienced virologic failure while on study and had an HIV-1 genotype recorded at the time of failure. The median time to virologic failure was 112 days. The following parameters were used for this paired sample analysis: $\alpha = \beta = 0.05$; $K = 3$; $K_0 = 2$; and $M = 1000$ independent bootstrap samples for each inference procedure.

There were 1069 first-order variable/level pairs that were observed in the data; of these, 13 were claimed, at the $\alpha = 0.05$ level, to have different prevalence between baseline and post-baseline. With $K_0 = 2$, $M' = \binom{1056}{2} = 557040$, so that the dimension reduction phase consisted of $557040 + 1$ iterations of the algorithm in Section 6.1. With computing resources as described in Section 7, a total of about 499 CPU hours were required to execute the proposed methodology; split over 10 CPUs, the effective time was about 50 hours. The final estimate of the set of combinations with a non-redundant group effect is given in Table 2.

Note that a positive value of (1) reflects a larger empirical post-baseline prevalence. At the global 5% FWER, there were significant one-way changes from baseline at the following RT codons: 100, 101, 103, 108 and 184. Except for codon 184, the change corresponded to the development of a mutation. The largest magnitude of change was from 103;K to 103;N, a well-known resistance mutation for the drug Efavirenz, a component of the ACTG 398 treatment regimen. The change at codon 184 was from the mutation V to wildtype M. This change is not surprising because the 184;V mutation is known to be associated with resistance to the drug lamivudine, to which these treatment experienced subjects had prior exposure, however, lamivudine was not part of the ACTG 398 treatment regimen and in its absence, 184;V reverted back to the wildtype 184;M. Singly, the RT codon/amino acid pairs 74;L, 190;G and 228;L were more prevalent at baseline.

At the global 9.75% FWER, there were seven second-order combinations and two third-order combinations that were claimed to have a non-redundant different prevalence between baseline and post-baseline. No codon/amino acid pairs from codons 74 and 190 were singly more prevalent post-baseline, however, all two-way combinations arising from 74;V, 181;C and 190;A were more prevalent post-baseline. That is, a greater post-baseline prevalence was observed when 74;V and 190;A either combined together or one combined with

181;C. Notice that, in addition to 74;V, 181;C and 190;A, the RT codon/amino acid pairs 41;M, 41;L, 122;K, 181;Y, 208;H and 210;W were not claimed to possess a marginal group effect, but were members of combinations claimed to have a non-redundant group effect.

The role of codon 181 on the prevalence of 74;V and 190;A post-baseline is intriguing. Post-baseline, the empirical conditional probability of 190;A given 181;C is 0.42; given 181;Y, the empirical probability of 190;A is 0.02, corresponding to a relative risk of about 20. Similarly, the empirical risk of 74;V given 181;C relative to 181;Y is 0.41/0.02. Finally, 103;N is observed to occur post-baseline about half as frequently with 190;A compared to another amino acid at codon 190. Given the 190;A mutation, the post-baseline empirical probability of 103;N is 0.24, while given an amino acid other than A, the empirical probability of 103;N is 0.55.

These results identify changes to known resistance mutations to drugs in ACTG 398 regimens among subjects observed to experience virologic failure. Since HIV-1 virus does not typically mutate in the absence of drug, this evidence suggests the selection of resistant variants as the cause for virologic failure rather than medication non-adherence.

9 Concluding Remarks

Although the statements made in this paper regarding asymptotic control of the FWER hold for any choice of M , larger values will likely reduce variability. As such, one has the option to increase the value of M in calculation of $\mathcal{C}_\alpha^1$, $\hat{\mathcal{G}}_\alpha$ and $\hat{\mathcal{G}}'_{2,\alpha,\beta}$ in the algorithm in Section 6.1. Note that these three calculations are outside of the computationally demanding loop thru M' , so that increasing M for these calculations is feasible. Also, increased precision of these three estimates is important since $(\hat{\mathcal{G}}_{1,\alpha}, \hat{\mathcal{G}}'_{2,\alpha,\beta})$ is the final estimate and $\mathcal{C}_\alpha^1$ determines $\mathcal{C}_\alpha^0$, and thus M' . Finally, note that the M' iterations of the 4-step algorithm in Section 6.1 can be sub-divided over several computer processors and run in parallel.

Generating resamples by permuting group indicators obviously cannot be used for testing the conditional independence null hypotheses. Furthermore, for both the paired and independent sample case, permutation-based resampling does not directly test the equal prevalence null hypotheses of interest. For example, for the paired sample case, the permutation-based approach generates resamples by exchanging group indicators *within* each observation $n = 1, \dots, N$. Given $(\mathbf{Y}_n^0, \mathbf{Y}_n^1)_{n=1}^N$, each resampled vector of test statistics $(S_{m,b}^*)_{b=1}^B$, $m = 1, \dots, M$, is equally likely only under both (i) $P^0 = P^1$ and (ii)

$P^0|Y^1=y = P^1|Y^0=y \forall y$ (Pesarin, 2001, p.20).

Error rates other than the FWER can be used in multiple testing procedures. For example, van der Laan, Dudoit and Pollard (2004b) show how any multiple testing procedure that asymptotically controls the FWER can be augmented to control the generalized FWER and the proportion of false positives. Although these other types of error rates can be used instead of the FWER when simultaneously testing the multiple equal prevalence and conditional independence null hypotheses, the resulting impact of their use on the interpretation of the overall error rate associated with the estimator $\hat{\mathcal{G}}_{0,\alpha,\beta}$ is not as straight-forward as when the FWER is used. For example, suppose, for the equal prevalence null hypotheses, that one wanted to asymptotically control the probability of 3 or more false rejections (denoted gFWER(2)) at level α . Furthermore, suppose the conditional independence null hypotheses were tested with asymptotic control of the gFWER(3) at level β . First, note that since several conditional independence null hypotheses correspond to one combination, the number of combinations in $\hat{\mathcal{G}}_{0,\alpha,\beta}$ that are potentially impacted by the falsely rejected conditional independence null hypotheses is not known exactly. For example, if there were 2 falsely rejected conditional independence null hypotheses, then the prevalence of either 0, 1 or 2 combinations in $\hat{\mathcal{G}}_{0,\alpha,\beta}$ can be written as the same product of prevalences of lower-order combinations in each group. Thus, one can conclude the bounded statement that, asymptotically, with probability at least $(1 - \alpha)(1 - \beta)$ and for the specific pair {gFWER(2), gFWER(3)}, the set $\hat{\mathcal{G}}_{0,\alpha,\beta}$ contains 2+3=5 or fewer combinations with either no group effect or a redundant group effect. Note that several combinations of generalized FWERs can lead to the same conclusion, e.g. testing equal prevalence at gFWER(1) (instead of gFWER(2)) and conditional independence at gFWER(4) (instead of gFWER(3)). If one wanted to conclude the more general statement that, asymptotically, with probability at least $(1 - \alpha)(1 - \beta)$ the set $\hat{\mathcal{G}}_{0,\alpha,\beta}$ contains 5 or fewer combinations with either no group effect or a redundant group effect, $\hat{\mathcal{G}}_{0,\alpha,\beta}$ would need to be calculated at each of the six combinations of gFWERs {gFWER(0),gFWER(5)}, ..., {gFWER(5),gFWER(0)}, each at the $\alpha/6$ and $\beta/6$ error levels, and then combined together so that the final result would be, say $\cup_{v=1}^6 \hat{\mathcal{G}}_{0,\alpha/6,\beta/6}(v)$.

Appendix

Conditional independence null hypotheses for fourth-order combinations

Consider the fourth-order variable/level combination $J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3, r_4; \ell_4}$, with $r_i = 1, \dots, R$, $r_i \neq r_j$, $i \neq j$ and $\ell_i = 1, \dots, L_{r_i}$. There are seven conditional independence null hypotheses in each group $W = 0, 1$. To ease notation, define $r_1; \ell_1, r_2; \ell_2, r_3; \ell_3, r_4; \ell_4 \equiv \text{comb}$.

$$Q_{\text{comb}}^{W,1} : P(J_{\text{comb}} = 1|W) = P(J_{r_1; \ell_1} = 1|W)P(J_{r_2; \ell_2, r_3; \ell_3, r_4; \ell_4} = 1|W),$$

$$Q_{\text{comb}}^{W,2} : P(J_{\text{comb}} = 1|W) = P(J_{r_2; \ell_2} = 1|W)P(J_{r_1; \ell_1, r_3; \ell_3, r_4; \ell_4} = 1|W),$$

$$Q_{\text{comb}}^{W,3} : P(J_{\text{comb}} = 1|W) = P(J_{r_3; \ell_3} = 1|W)P(J_{r_1; \ell_1, r_2; \ell_2, r_4; \ell_4} = 1|W),$$

$$Q_{\text{comb}}^{W,4} : P(J_{\text{comb}} = 1|W) = P(J_{r_4; \ell_4} = 1|W)P(J_{r_1; \ell_1, r_2; \ell_2, r_3; \ell_3} = 1|W),$$

$$Q_{\text{comb}}^{W,5} : P(J_{\text{comb}} = 1|W) = P(J_{r_1; \ell_1, r_2; \ell_2} = 1|W)P(J_{r_3; \ell_3, r_4; \ell_4} = 1|W),$$

$$Q_{\text{comb}}^{W,6} : P(J_{\text{comb}} = 1|W) = P(J_{r_1; \ell_1, r_3; \ell_3} = 1|W)P(J_{r_2; \ell_2, r_4; \ell_4} = 1|W),$$

$$Q_{\text{comb}}^{W,7} : P(J_{\text{comb}} = 1|W) = P(J_{r_1; \ell_1, r_4; \ell_4} = 1|W)P(J_{r_2; \ell_2, r_3; \ell_3} = 1|W).$$

Probability distributions for simulation study

Denote the probability mass function $p_{Y_i|Y_j=y}^W = P(Y_i = 1|Y_j = y, W)$ for the binary variables Y_i , $i \neq j$.

$$\begin{aligned} \text{The non-zero components of } P^0 \text{ are } p_{Y_1}^0 &= \frac{1}{4}, p_{Y_2|Y_1=y}^0 = \begin{cases} 1/2; y=0 \\ 9/10; y=1 \end{cases}, \\ p_{Y_3}^0 &= \frac{1}{2}, p_{Y_4|Y_3=y}^0 = \begin{cases} 3/10; y=0 \\ 7/10; y=1 \end{cases}, p_{Y_5}^0 = \frac{3}{4}, p_{Y_6|Y_5=y}^0 = \begin{cases} 3/5; y=0 \\ 4/5; y=1 \end{cases}, p_{Y_7}^0 = \\ \frac{1}{4}, p_{Y_8|Y_7=y}^0 &= \begin{cases} 1/5; y=0 \\ 4/5; y=1 \end{cases}, p_{Y_9}^0 = \frac{3}{4}. \end{aligned}$$

$$\begin{aligned} \text{The non-zero components of } P^1 \text{ are } p_{Y_1}^1 &= \frac{1}{4}, p_{Y_2|Y_1=y}^1 = \begin{cases} 7/10; y=0 \\ 3/10; y=1 \end{cases}, \\ p_{Y_3}^1 &= \frac{1}{2}, p_{Y_4|Y_3=y}^1 = \begin{cases} 7/10; y=0 \\ 3/10; y=1 \end{cases}, p_{Y_5}^1 = \frac{1}{4}, p_{Y_6|Y_5=y}^1 = \begin{cases} 4/5; y=0 \\ 3/5; y=1 \end{cases}, \\ p_{Y_7}^1 &= \frac{3}{4}, p_{Y_8|Y_7=y}^1 = \begin{cases} 1/2; y=0, 1 \end{cases}, p_{Y_9}^1 = \frac{3}{4}. \end{aligned}$$

References

- Cox, D.R., and Reid, N. (2004), "A note on pseudolikelihood constructed from marginal densities," *Biometrika*, 91, 729-737.
- Dudoit, S., van der Laan, M. and Pollard, K. (2004). Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology* 1, article 13. URL www.bepress.com/sagmb/vol3/iss1/art13.
- Pesarin, F. (2001), *Multivariate Permutation Tests: With applications in Biostatistics*. John Wiley & Sons, Chichester, England.
- Pollard, K.S. and van der Laan, M.J. (2003). "Multiple testing for gene expression data: an investigation of null distributions with consequences for the permutation test." *Proceedings of the 2003 International MultiConference in Computer Science and Engineering*, METMBS'03 Conference, pp. 3-9.
- Romano, J., and Wolf, M. (2005), "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100, 94-108.
- Troendle, J. (1995), "A Stepwise Resampling Method of Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 90, 370-378.
- van der Laan, M., Dudoit, S., and Pollard, K. (2004a), "Multiple Testing. Part II. Step-Down Procedures for Control of the Family-Wise Error Rate," *Statistical Applications in Genetics and Molecular Biology* 1, article 14. URL www.bepress.com/sagmb/vol3/iss1/art14.
- van der Laan, M., Dudoit, S., and Pollard, K. (2004b), "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* 1, article 15. URL sagmb/vol3/iss1/art15.
- van der Laan, M., and Hubbard, A. (2005), "Quantile-Function Based Null Distribution in Resampling Based Multiple Testing," U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 198. URL

Table 1: *Simulation results at $\alpha = \beta = 5\%$*

			INDEPENDENT			PAIRED			
			N^0 :	54	92	265	50	100	250
			N^1 :	45	105	243	50	100	250
			FWER(%):	8.3	7.0	6.9	9.1	7.2	8.0
COMBIN- ATION	P^1	P^0	EMPIRICAL POWER (%)						
5;0	3/4	1/4	99.5	100	100	99.6	100	100	
5;1	1/4	3/4	99.5	100	100	99.6	100	100	
7;0	1/4	3/4	99.7	100	100	99.6	100	100	
7;1	3/4	1/4	99.7	100	100	99.6	100	100	
8;0	1/2	13/20	18.3	34.4	78.4	18.1	31.9	79.4	
8;1	1/2	7/20	18.3	34.4	78.4	18.1	31.9	79.4	
2;0, 1;0	9/40	3/8	0	10.8	84.8	0	11.7	87.7	
2;1, 1;0	21/40	3/8	0.2	11.5	78.5	0.1	13.0	82.3	
2;0, 1;1	7/40	1/40	4.2	11.7	98.9	3.4	14.7	98.7	
2;1, 1;1	3/40	9/40	1.0	19.8	99.0	2.0	23.8	99.3	
4;0, 3;0	3/20	7/20	2.1	65.3	100	2.5	58.7	99.7	
4;1, 3;0	7/20	3/20	3.1	61.7	99.8	3.1	58.2	99.8	
4;0, 3;1	7/20	3/20	2.9	60.1	100	3.3	57.2	99.8	
4;1, 3;1	3/20	7/20	2.5	64.2	99.9	2.9	58.5	99.9	
6;0, 5;0	3/20	1/10	0	0	0.4	0	0	0.4	
6;1, 5;0	3/5	3/20	0	0.6	27.6	0	1.4	22.9	
6;0, 5;1	1/10	3/20	0	0	0.9	0	0	0.8	
6;1, 5;1	3/20	3/5	0	0.6	27.6	0	1.4	22.9	
8;0, 7;0	1/8	3/5	2.7	54.4	100	2.0	62.8	100	
8;1, 7;0	1/8	3/20	0.2	1.2	3.9	0.1	1.5	4.2	
8;0, 7;1	3/8	1/20	2.6	54.4	100	2.0	62.8	100	
8;1, 7;1	3/8	1/5	0.9	31.6	96.6	0.5	37.6	97.3	

Note: P^W here denotes each combination's prevalence in group $W = 0, 1$.

Table 2: *Results from ACTG 398 at $\alpha = \beta = 5\%$*

COMBINATION	$\hat{S}(\%)$
103;K*	-43
103;N	37
108;V*	-9
108;V*/I	8
100;L*	-9
100;L*/I	7
101;K*	-7
101;K*/E	6
184;V	-7
184;M*	6
190;G*	-13
74;L*	-7
228;L*	-4
190;G*, 101;K*	-13
74;L*, 41;L	-6
181;Y*, 103;K*/N	5
181;C, 74;V	4
181;C, 190;A	4
122;K*, 41;M*	4
190;A, 74;V	3
103;K*, 190;G*, 181;C	-7
103;K*, 210;W, 208;H*	-5

Note: Each variable/level pair in a combination is of the form: RT codon;amino acid. The column $\hat{S}(\%)$ denotes the observed value of $(1) \times 100$. '*' denotes the consensus wildtype amino acid. For amino acids A and B, A/B denotes their mixture (possibly with other amino acids).