Published in final edited form as:

Stat Med. 2018 October 15; 37(23): 3357–3372. doi:10.1002/sim.7825.

Multi-Sample Adjusted U-Statistics that Account for Confounding Covariates

Glen A. Satten[1], **Maiying Kong**[2], and **Somnath Datta**[*,3]

[1]Division of Reproductive Health, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

[2]Department of Bioinformatics and Biostatistics, SPHIS, University of Louisville, Louisville, Kentucky, USA

[3]Department of Biostatistics, University of Florida, Gainesville, Florida, USA

Summary

Multi-sample U-statistics encompass a wide class of test statistics that allow the comparison of two or more distributions. U-statistics are especially powerful because they can be applied to both numeric and non-numeric data, e.g., ordinal and categorical data where a pairwise similarity or distance-like measure between categories is available. However, when comparing the distribution of a variable across two or more groups, observed differences may be due to confounding covariates. For example, in a case-control study, the distribution of exposure in cases may differ from that in controls entirely because of variables that are related to both exposure and case status and are distributed differently among case and control participants. We propose to use individually-reweighted data (i.e., using the stratification score for retrospective data or the propensity score for prospective data) to construct adjusted U-statistics that can test the equality of distributions across two (or more) groups in the presence of confounding covariates. Asymptotic normality of our adjusted *U*-statistics is established and a closed form expression of their asymptotic variance is presented. The utility of our approach is demonstrated through simulation studies, as well as in an analysis of data from a case-control study conducted among African-Americans, comparing whether the similarity in haplotypes (i.e., sets of adjacent genetic loci inherited from the same parent) occurring in a case and a control participant differs from the similarity in haplotypes occurring in two control participants.

Keywords

Adjusted U-statistics; Propensity score; Multiple group comparison

[*]**Correspondence**: Somnath Datta, Ph.D., Department of Biostatistics, University of Florida, Gainesville, Florida, USA. somnath.datta@ufl.edu.

DISCLAIMER

The conclusions, findings, and opinions expressed by the authors do not necessarily reflect the official position of the U.S. Department of Health and Human Services, the Public Health Service or the Centers for Disease Control and Prevention.

SOFTWARE AVAILABILITY

R-code for the computing the standardized test statistic is available as "SUPPORTING WEB MATERIAL" to this manuscript.

1 | INTRODUCTION

U-statistics[1,2] are widely used to compare the distribution of a random variable of interest across two or more groups. An appealing feature of U-statistics is that they often rely only on symmetry. For example, given that the distributions of a variable across two or more groups are the same, if the data are pooled and then ranked, we would expect that the average rank of observations from each group should be the same; this forms the basis of the Wilcoxon rank sum test. U-statistics are very general, and can be used for non-numeric data, e.g., ordinal and categorical data where a pairwise similarity or distance-like measure between categories is available. For example, in Section 5 we analyze data from a case-control study, comparing the extent to which the haplotypes (i.e., the alleles that were inherited from the same parent at sets of adjacent genetic loci) of case participants are different from the haplotypes of control participants. Haplotypes can be coded by strings of text, and one measure of the similarity between two haplotypes is the number of letters they share in common at the same location in the string. We show that tests of haplotype similarity can be easily constructed using U-statistics.

When group membership is randomly assigned, we are certain that any difference we observe between the groups must be due to differential treatment of the groups after randomization. For example, after randomization, one group may be given an active drug and another group a placebo. In this case, differences in medical outcome can be attributed to the effect of the drug. However, when group membership is not assigned through randomization, there may be confounding covariates that can cause a spurious association between outcome and group membership. Specifically, if there are covariates that influence both group membership and the outcome variable we are comparing, then an observed difference in the distribution of outcome variable across groups may be due to a difference in the distribution of these covariates across groups. For example, in a study that compares lung capacity among persons who consume alcohol and persons who abstain from alcohol, an observed difference may be due to the presence of more smokers among the persons who consume alcohol. In the haplotype example, the genetic ancestry of cases may differ systematically from controls. In the data from African-Americans we consider, the proportions of African or European ancestry will affect the distribution of haplotypes found in each group. If African-Americans have a different risk of disease than persons of purely European ancestry, then genetic ancestry is a confounder and must be accounted for in the analysis.[3,4]

The usual approach to account for confounding covariates is to model their effect on the outcome of interest using a regression approach. While this direct approach is very useful, it requires a test that can be formulated in a regression setting. Unfortunately, U-statistics are typically not related to regression procedures. As a result, direct adjustment is problematic. For example, it is unclear how the direct approach could be applied in the haplotype example, where the outcome (similarity or sharing) is only defined for pairs of haplotypes. A related approach is to regress the outcome on covariates and then form a U-statistic from the residuals of this regression. This approach is only valid for linear regression, and is limited to the situation where the outcome variable is numeric.

Here we take an alternative approach based on the stratification score[5,6] or the propensity score.[7,8] We model the probability of group membership as a function of confounding covariates, typically using logistic regression for two groups or polyto-mous logistic regression for more than two groups. Then, we inversely weight the sample according to the probability of group membership.[9] Under the null hypothesis of no group effect on the outcome, the weighted outcome distributions should be the same.[5,10] We then construct adjusted U-statistic tests based on these weighted distributions. Since reweighted sample means and estimating equations based on propensity scores have been used in the context of causal inference, it can be anticipated that a similar approach may work for a U-statistic. Although the inverse weighting approach to account for confounding has been proposed over years, it has not been well established in the non-parametric field other than the work by Jiang et al.[11,12] and Rosen-baum.[13] Jiang et al.[11,12] proposed a propensity-score adjusted generalization of Kendall's Tau for estimating the association between genotype and trait in a single population; Rosenbaum[13] proposed a new family of U-Statistics for comparing matched pairs. However, a formal treatment for a general kernel of such U-statistics does not seem to be available. We demonstrate, both theoretically and through simulations, that this approach works not only for propensity scores but also for stratification scores in retrospective studies. More importantly, we obtain a closed form of asymptotic variance estimator for our reweighted U-statistics, which is a novel and useful contribution of our current work. We also generalize our proposed U-Statistics to compare multiple groups. Because this variance estimator is somewhat complex, we have made an R-code available that implements our approach for two- and multi-sample tests.

The rest of the paper is organized as follows. In section 2 we consider stratification-score-based (or propensity-score-based) weighting as a way of adjusting a distribution for confounders, in essence by standardization. In section 3 we consider adjusted two-sample U-statistics including generalizations that compare multiple groups by systematic pairwise comparisons, generalizing both the Kruskal-Wallis test for general alternatives and the Jonckheere-Terpstra test for ordered alternatives. In section 4 we investigate the statistical properties of our adjusted U-statistics using simulated data, focusing on the Rank-Sum test. In section 5 we apply our adjusted U-statistics to genetic data to test the association between haplotypes in the catechol-O-methyltransferase (COMT) gene and schizophrenia among African-Americans. We discuss our results in section 6. Technical details can be found in the appendices.

2 | SCORE-BASED ADJUSTMENT FOR CONFOUNDING AND ADJUSTED U-STATISTICS

To develop adjusted U-statistics that account for confounding covariates, we adopt a marginal approach that standardizes the data by weighting observations so that the distribution of confounding covariates is the same in each group.[5,15] Assume that the ith observation is a member of group g_i, and let Y_i denote the outcome variables with realization y_i. We let n_g denote the total number of observations from group g, $1 \le g \le \mathscr{G}$, and n denote the total number of observations in the entire sample, $n = n_1 + \cdots + n_{\mathscr{G}}$. In the simplest case there are only two groups and we wish to test the null hypothesis that Y from group 1 has the same distribution as Y from group 2, against an alternative hypothesis that the distributions of Y differ by group. In a more general setting, assume $\mathscr{G}$ groups and let $\mathscr{I}$ and $\mathscr{A}$ denote two non-intersecting sets of groups. Then, we may wish to test whether the values of Y from observations having $g_i \in \mathscr{I}$ have a different distribution from values of Y from observations having $g_i \in \mathscr{A}$. As a concrete example, we may wish to test whether the distribution of Y values from group 1 is different from the distribution of Y values in all the other groups. In the absence of confounding covariates, we would expect that hypotheses like these could be tested using the a two-sample U-statistic of order ($\boldsymbol{B}, \boldsymbol{D}$) having the form

$$U = \frac{1}{\binom{m_1}{B}\binom{m_2}{D}} \sum_{\boldsymbol{i} \in \mathscr{S}_B(\mathscr{I})} \sum_{\boldsymbol{j} \in \mathscr{S}_D(\mathscr{A})} K\left(y_{i_1}, \cdots, y_{i_B}; y_{j_1}, \cdots, y_{j_D}\right) \quad (1)$$

where $\boldsymbol{i} = (i_1, i_2, \cdots, i_B)$, $\boldsymbol{j} = (j_1, j_2, \cdots, j_D)$, $\mathscr{S}_B(\mathscr{I}) = \left\{\boldsymbol{i} \middle| i_1 < i_2 < \cdots < i_B, \ g_{i_b} \in \mathscr{I} \text{ for } b = 1, \cdots, \boldsymbol{B}\right\}$, and $\mathscr{S}_D(\mathscr{A}) = \left\{\boldsymbol{j} \middle| j_1 < j_2 < \cdots < j_D, g_{j_d} \in \mathscr{A} \text{ for } d = 1, \cdots, \boldsymbol{D}\right\}$, m_1 is the number of observations having $g_i \in \mathscr{I}$, m_2 is the number of observations having $g_i \in \mathscr{A}$ and $K(\boldsymbol{\cdot}, \boldsymbol{\cdot})$ is the kernel of the U-statistic. We assume without loss of generality that K is symmetric upon interchange of the first B and second D arguments. We let ω denote the expected value of U under the null hypothesis that the distribution of values of Y is the same whether $g_i \in \mathscr{I}$ or $\mathscr{A}$. We assume that the centering of K under the null hypothesis is known and does not depend on the common distribution of Y. For example, if we use the Wilcoxon kernel $K\left(y_1, y_2\right) = \frac{1}{2}I\left[y_1 < y_2\right] + \frac{1}{2}I\left[y_1 \le y_2\right]$ we know that $\omega = \frac{1}{2}$ because the chance that $Y_1 < Y_2$ is $\frac{1}{2}$ (ignoring ties) as long as Y_1 has the same distribution as Y_2. In the presence of confounding covariates Z if we calculated U among persons having the same values of Z, we would still find that the expected value of U was $\frac{1}{2}$ when the null hypothesis is true. However, when we ignore confounding covariates Z when calculating U, we may find that the expected value of (1) differs from $\frac{1}{2}$ even under the null hypothesis. For this reason, we seek a version of (1) that will account for the effect of confounding covariates without requiring that we calculate U separately for each set of covariate values Z.

To develop such a test, consider comparing the CDF of Y between two or more groups in the presence of confounding covariates Z. We wish to test the null hypothesis that $pr[Y \le y/G = g, Z = z] = pr[Y \le y/Z = z]$ or equivalently, $pr[G = g/Y = y, Z = z] = pr[G = g/Z = z]$, where we have implicitly assumed there are no unmeasured confounding covariates. Note that under this null hypothesis we may find that $pr[Y \le y/G = g] \ne pr[Y \le y]$ since $pr[Z = z/G = g] \ne pr[Z = z]$. Thus, it is necessary to properly account for confounding covariates Z when we construct our U-statistics.

Allowing Z to include interactions or powers of measured covariates, a fairly general model is to assume that under the null hypothesis, covariates Z influence group membership according to the multivariate logistic regression model

$$log\left\{\frac{pr[G = g|Z = z]}{pr[G = 1|Z = z]}\right\} = \gamma_g^T \cdot z, \text{ for } g = 2, \cdots, \mathscr{G}, \quad (2)$$

where we assume that the first component of Z is an intercept, and where we note that the choice of reference group in (2) is arbitrary. We assume that model (2) correctly specifies the relationship between group membership and the covariates, and that the usual regularity conditions (see, e.g., Fahrmeir and Kaufmann[14]) for consistency and asymptotic normality of the maximum likelihood estimator of $\boldsymbol{\gamma}$ are satisfied. It should be noticed that the results of inverse weighting approach may not be valid if the stratification and/or propensity scores models in equation (2) are mis-specified. For later use, we note that $\hat{\boldsymbol{\gamma}} = \left(\hat{\gamma}_2, \hat{\gamma}_3, \cdots, \hat{\gamma}_G\right)'$ is obtained by solving the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{S}_i(\boldsymbol{\gamma}) = 0 \quad (3)$$

where $\boldsymbol{S}_i(\boldsymbol{\gamma})$ is the score function from the ith individual under model (2) and the multinomial distribution assumption for group membership. The estimate for $\boldsymbol{\gamma}$ in the estimating equation (3) is obtained by the Fisher-scoring iterative method which is implemented using the R-package VGAM.[18] Once $\hat{\gamma}$ is obtained, we define $\hat{\boldsymbol{S}}_i \equiv \boldsymbol{S}_i(\hat{\gamma})$ and $\hat{\boldsymbol{J}} = \sum_{i=1}^{n} \frac{\partial S_i}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}}$.

Assume that we fit model (2) using data on G and Z only to obtain $\hat{\boldsymbol{\gamma}}$ and thereby obtain $pr\ [G = g|Z = z; \hat{\boldsymbol{\gamma}}]$. For retrospective data, G corresponds to disease status and $pr\ [G = g|Z = z; \hat{\boldsymbol{\gamma}}]$ is the stratification score [5, 11], here generalized to multiple disease categories. For prospective data, G corresponds to an exposure and $pr\ [G = g|Z = z; \hat{\boldsymbol{\gamma}}]$ is the propensity score [6] as generalized to multiple exposure groups by Imbens [7].

Returning to the problem of estimating the distribution of Y, for either retrospective or prospective data, we can estimate a weighted CDF of Y in group g by

$$\frac{1}{n}\sum_{i=1}^{n} w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right)\boldsymbol{I}\left[y_i \le y, g_i = g\right], \quad (4)$$

where the weights in (4) are chosen to standardize the data in each group to the same distribution of Z. We consider here two choices for the weight $w(z_i, g_i; \boldsymbol{\gamma})$, namely

$$w\left(z_i, g_i; \boldsymbol{\gamma}\right) = \frac{1}{pr\left[G = g_i | Z = z_i; \boldsymbol{\gamma}\right]} \qquad (5)$$

and

$$w\left(z_i, g_i; \boldsymbol{\gamma}\right) = \frac{pr\left[G = 1 \middle| Z = z_i; \boldsymbol{\gamma}\right]}{pr\left[G = g_i | Z = z_i; \boldsymbol{\gamma}\right]}. \qquad (6)$$

Choosing (5) corresponds to standardizing data from each group g to the marginal distribution of Z found in the study population, while choosing (6) corresponds to standardizing data from each group to the distribution of Z found in group 1,[5] thus data in group 1 are not reweighted in equation (6). Note that choice of labels for the groups is arbitrary, so (6) can be used to standardize to the distribution of Z found in any of the groups.

Although (4) is normalized in expected value, it may fail to be normalized in practice. Hence, we may prefer to estimate the standardized CDF of Y in group g by

$$\frac{\Sigma_{i=1}^{n} w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right) \boldsymbol{I}\left[y_i \le y, g_i = g\right]}{\Sigma_{i=1}^{n} w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right) \boldsymbol{I}\left[g_i = g\right]}.$$

Finally, under the null hypothesis that the distribution of Y is the same for all groups in $\mathscr{I}$, we can estimate the standardized CDF of Y for those groups in $\mathscr{I}$ by

$$\frac{1}{m_1} \sum_{i=1}^{n} \frac{w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right)}{\widehat{W}_1(\hat{\boldsymbol{\gamma}})} I\left[y_i \le y, g_i \in \mathscr{I}\right],$$

where m_1 is the number of observations for which $g_i \in \mathscr{I}$ and

$$\widehat{W}_1(\hat{\boldsymbol{\gamma}}) = \frac{1}{m_1} \sum_{i=1}^{n} w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right) I\left[g_i \in \mathscr{I}\right].$$

Similarly, we can estimate the standardized CDF of Y for groups in $\mathscr{A}$ by

$$\frac{1}{m_2} \sum_{i=1}^{n} \frac{w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right)}{\widehat{W}_2(\hat{\boldsymbol{\gamma}})} I\left[y_i \le y, g_i \in \mathscr{A}\right].$$

where m_2 is the number of observations for which $g_i \in \mathscr{A}$ and

$$\widehat{W}_2(\hat{\boldsymbol{\gamma}}) = \frac{1}{m_2}\sum_{i=1}^{n} w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right) I\left[g_i \in \mathscr{A}\right].$$

Motivated by the standardized (weighted) CDF estimators just described, we propose the following two-sample adjusted *U*-statistic to account for confounding covariates,

$$\begin{aligned} U_a = {} & \frac{1}{\binom{m_1}{B}} \frac{1}{\binom{m_2}{D}} \sum_{\boldsymbol{i} \in S_B(\mathscr{I})} \sum_{\boldsymbol{j} \in S_D(\mathscr{A})} \left\{ \prod_{b=1}^{B} \widetilde{w}_1\left(z_{i_b}, g_{i_b}; \hat{\boldsymbol{\gamma}}\right) \right\} \\ & \times K\left(y_{i_1}, \dots, y_{i_B}; y_{j_1}, \dots, y_{j_D}\right) \left\{ \prod_{d=1}^{D} \widetilde{w}_2\left(z_{j_d}, g_{j_d}; \hat{\boldsymbol{\gamma}}\right) \right\}, \end{aligned} \qquad (7)$$

where

$$\widetilde{w}_k\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right) = \frac{w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right)}{\widehat{W}_k(\hat{\boldsymbol{\gamma}})}$$

for $k = 1, 2$. In writing (7) we have implicitly assumed the data are *iid* so that the weights assigned to sets of observations are the product of the weights for each observation.

Comparing (7) with (1), we see that (7) differs from a standard *U*-statistic in two ways. First, the normalizations $\widehat{W}_k(\hat{\boldsymbol{\gamma}})$ are functions of all the data; second, $\hat{\boldsymbol{\gamma}}$ in $w\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right)$ $(i = 1, \dots, n)$ is also a function of the data. However, U_a is closely related to the standard *U*-statistic having the following adjusted kernel

$$\left\{ \prod_{b=1}^{B} w\left(z_{i_b}, g_{i_b}; \boldsymbol{\gamma}\right) \right\} K\left(y_{i_1}, \dots, y_{i_B}; y_{j_1}, \dots, y_{j_D}\right) \left\{ \prod_{d=1}^{D} w\left(z_{j_d}, g_{j_d}; \boldsymbol{\gamma}\right) \right\}.$$

We assume that the second moment of this adjusted kernel is finite. We show in the appendix that it is possible to develop a linear approximation of U_a using the projection approach. In particular, we show that U_a has an asymptotically normal distribution, and that the asymptotic variance of U_a is consistently estimated by $\hat{v}$ given by

$$\hat{v} = \sum_{g \in \mathscr{I} \cup \mathscr{A}} n_g \hat{\sigma}_g^2 \qquad (8)$$

where n_g is the number of observations having $g_i = g$ and

$$\hat{\sigma}_g^2 = \frac{1}{n_g - 1} \sum_{\{i | g_i = g\}} \left(\xi_i - \bar{\xi}_g\right)^2, \quad (9)$$

and

$$\bar{\xi}_g = \frac{1}{n_g} \sum_{\{i | g_i = g\}} \xi_i .$$

Here $\xi_i = \xi_1(i)\boldsymbol{I}\left[g_i \in \mathscr{I}\right] + \xi_2(i)\boldsymbol{I}\left[g_i \in \mathscr{A}\right]$ with

$$\xi_1(i) = m_1^{-1}\left[-B\,\hat{\mu}\left(\,\tilde{w}_1\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right) - 1\right) + B\left(\,\tilde{h}_1(i) - \hat{\mu}\right)\right] + n^{-1}C_nJ^{-1}\boldsymbol{S}_i\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right) \quad (10a)$$

$$\xi_2(j) = m_2^{-1}\left[-D\,\hat{\mu}\left(\,\tilde{w}_2\left(z_j, g_j; \hat{\boldsymbol{\gamma}}\right) - 1\right) + D\left(\,\tilde{h}_2(j) - \hat{\mu}\right)\right] + n^{-1}C_nJ^{-1}\boldsymbol{S}_j\left(z_j, g_j; \hat{\boldsymbol{\gamma}}\right). \quad (10b)$$

In the above equations, $\hat{\mu}$ is estimated by equation (7), and $\tilde{h}_1(i)$, $\tilde{h}_2(j)$, and C_n are defined in equations (A7–A9) in Appendix A1; $\hat{S}_i$ and $\hat{J}$ were defined right after equations (3). Hypotheses about the distribution of Y can then be tested using the test statistic

$$Z_a = \frac{U_a - \omega}{\sqrt{\hat{v}}},$$

which has an asymptotic Normal distribution with mean zero and unit variance. We provide the asymptotic linear representation in Appendix A1 and show in the Appendix A2 that $U_a - \omega$ constructed using a case-control sample indeed has zero asymptotic mean under the null hypothesis of equal distributions in two groups adjusted for covariates.

3 | COMPARING MORE THAN TWO GROUPS

To account for comparisons involving more than two groups, we consider a vector of two-sample U-statistics $\mathscr{U}_a = \left(U_a^{(1)}, \ldots, U_a^{(R)}\right)$, each component of which has the form (7) and tests the hypothesis that the distribution of Y for observations having $g \in \mathscr{I}_r$ is the same as the distribution of Y for observations having $g \in \mathscr{A}_r$ for $1 \le r \le R$. This approach allows us to account for comparisons involving more than two groups without having to construct multi-sample U-statistics, and leads to generalizations of popular tests such as the Kruskal-Wallis and Jonckheere-Terpstra tests to account for confounding covariates. The r^{th} component of $\mathscr{U}_a$ thus has the form

$$U_a^{(r)} = \frac{1}{\binom{m_1^{(r)}}{B}} \frac{1}{\binom{m_2^{(r)}}{D}} \sum_{\boldsymbol{i} \in S_B(\mathscr{I}_r)} \sum_{\boldsymbol{j} \in S_D(\mathscr{A}_r)} \left\{ \prod_{b=1}^{B} \widetilde{w}_1^{(r)}\left(z_{i_b}, g_{i_b}; \hat{\boldsymbol{\gamma}}\right) \right\}$$
$$\times K\left(y_{i_1}, \ldots, y_{i_B}; y_{j_1}, \ldots, y_{j_D}\right) \left\{ \prod_{d=1}^{D} \widetilde{w}_2^{(r)}\left(z_{j_d}, g_{j_d}; \hat{\boldsymbol{\gamma}}\right) \right\}$$

where

$$\widetilde{w}_1^{(r)}\left(z_i, g_i; \boldsymbol{\gamma}\right) = \frac{w\left(z_i, g_i; \boldsymbol{\gamma}\right)}{\left\{m_1^{(r)}\right\}^{-1} \Sigma_{i:g_i \in \mathscr{I}_r} w\left(z_i, g_i; \boldsymbol{\gamma}\right)}$$

and

$$\widetilde{w}_2^{(r)}\left(z_j, g_j; \boldsymbol{\gamma}\right) = \frac{w\left(z_j, g_j; \boldsymbol{\gamma}\right)}{\left\{m_2^{(r)}\right\}^{-1} \Sigma_{j:g_j \in \mathscr{A}_r} w\left(z_j, g_j; \boldsymbol{\gamma}\right)}.$$

Because each component $U_a^{(r)}$ has the same form as (7), the projection results derived for the two-group situation apply directly. Thus, the sample variance-covariance matrix of $\mathscr{U}_a$ can be consistently estimated by

$$\hat{\boldsymbol{V}} = \sum_{g \in \mathscr{I} \cup \mathscr{A}} n_g \hat{\Sigma}_g$$

where $\hat{\Sigma}_g$ is the variance-covariance matrix of ξ_i (now a vector) calculated among those observations having $g_i = g$. To generalize the Kruskal-Wallis test, we choose $\mathscr{I}_r = \{r\}$ and $\mathscr{A}_r = \left\{j \mid j \neq r\right\}$ for $1 \leq r \leq \mathscr{G}$. General hypotheses about the distribution of Y can then be tested using the test statistic

$$T_a = \left(\mathscr{U}_a - \omega \mathbf{1}\right)' \hat{\boldsymbol{V}}^{-} \left(\mathscr{U}_a - \omega \mathbf{1}\right),$$

where $\hat{\boldsymbol{V}}^{-}$ denotes the generalized inverse of $\hat{\boldsymbol{V}}$ and $\mathbf{1}$ is a vector with all components equal to 1. Asymptotically, T_a has a χ^2 distribution with degrees of freedom given by the rank of the matrix $\hat{\boldsymbol{V}}$.

If the alternative hypothesis is that groups are ordered in their response, then we choose $\mathscr{I}_r = \left\{j \mid j < r\right\}$ and $\mathscr{A}_r = \left\{j \mid j \geq r\right\}$ for $2 \leq r \leq \mathscr{G}$. With this choice, we expect each component of $\mathscr{U}_a - \omega \mathbf{1}$ to be positive under the alternative hypothesis, so we can base testing on

$\boldsymbol{D}'\left(\mathcal{U}_a - \omega\mathbf{1}\right)$ which is normally distributed with variance $\boldsymbol{D}'\hat{\boldsymbol{V}}\boldsymbol{D}$, and where D is a $\boldsymbol{R}$-dimensional vector that specifies the choice of test statistic. To generalize the Jonckheere-Terpstra test[17] we choose $\boldsymbol{D} = \mathbf{1}$ to compute

$$Z_a = \frac{\mathbf{1}'\left(\mathcal{U}_a - \omega\mathbf{1}\right)}{\sqrt{\mathbf{1}'\hat{\boldsymbol{V}}\mathbf{1}}}.$$

Asymptotically, Z_a has a standard normal distribution. If the direction of the ordering is known a priori, a one-sided p-value may be used.

4 | SIMULATION RESULTS

To demonstrate the general properties of our test, we used data on three groups simulated using the model (2) with $Z = (Z_1, Z_2, Z_3, Z_4)$, where Z_1 is the intercept, $Z_2 \sim N(0, 1)$, $Z_3 \sim uniform\left[-\frac{1}{2}, \frac{1}{2}\right]$, and $Z_4 \sim\ Binomial\ \left(1, \frac{1}{2}\right) - \frac{1}{2}$. We used $\gamma_2 = (-2, 0.15, 0.2, 0.1)$ and $\gamma_3 = (-2.5, 0.3, 0.4, 0.2)$. We chose $Y \sim N\left(\frac{2}{11}\left(\mathbf{1}^T \cdot Z\right) - 1, 1\right)$ to ensure substantial confounding. Note that there is no association between Y and G in the presence of Z, so that the p-values of the test should be uniformly distributed. For all results shown here, we used the Wilcoxon kernel $K(x, y) = \frac{1}{2}I[x < y] + \frac{1}{2}I[x \le y]$ for which $\omega = \frac{1}{2}$.

To confirm the asymptotic normality of our adjusted U-statistics, we generated 1,000 data prospectively from model (2) with three groups, using the first 334 observations to fall into group 1, the next 333 observations to fall into group 2, and the last 333 observations to fall into group 3 as our data. The results are shown in Figure 1 Panel A for our generalization of the Kruskal-Wallis test and in Panel B for our generalization of the Jonckheere-Terpstra test. There is good agreement between the empirical and theoretical (uniform) p-values for the two adjusted tests, while the naive tests consistently have smaller p-values than are expected under the null. This is consistent with the notion that, absent adjustment for the confounder Z, there is an actual difference between the distribution of Y in the two groups. We show the empirical size for our simulations in Table 1. Drake[19] investigated the effects of misspecification of the propensity score on estimators of treatment effect, and conclude that the bias of the estimator of the treatment effect is large if the covariates are omitted. The naive approach (i.e., the unadjusted U-Statistics) can be considered as a special case of the propensity score models with all covariates omitted. The results presented in Table 1 are coincident with those from Drake,[19] indicating that the naive approach has a large bias for assessing the treatment effect.

Next, to investigate power, we considered simulating from the model

$$log\left\{\frac{pr[G = g \mid Z = z, Y = y]}{pr\ [G = 1 \mid Z = z, Y = y]}\right\} = \gamma_g^T \cdot z + \beta_g y,\ g = 2, 3 \quad (11)$$

for values (β_2, β_3) = α(0.025, 0.5), where α varies from 0 to 1. Note that under model (11) there is a true association between Y and G even after controlling for Z so long as $\alpha \neq 0$. We compared the power of our approach to detect this association with a Wald test of (β_2, β_3) = (0,0) calculated using the R package VGAM.[18] The results are given in Figure 2 Panel A. To compare with the Jonckheere-Terpstra test, we considered a Wald test of the hypothesis (β_2, β_3) = (0,0) for the same model as specified in equation (11). The results, given in Figure 2 Panel B, indicate that standardization to the study population outperforms standardization to group 1. The parametric model (11) outperforms the adjusted U-statistic, as to be expected. The efficiency loss when comparing to the parametric test could be large. Recall that this setting assumes that the parametric model is exactly correctly specified; we anticipate that the non-parametric approach will be more adventageous when the model for the group membership is misspecified.

To examine the power of the Wald test and U-statistics based test when the regression model is mis-specified, we carried out the exactly same simulations as above except that the simulated data were generated from the following model:

$$log\left\{\frac{pr[G = g \mid Z = z, Y = y]}{pr[G = 1 \mid Z = z, Y = y]}\right\} = \gamma_g^T \cdot z + \beta_g y^{\frac{1}{3}}, \ g = 2, 3.$$

As before, the Wald test is based on the regression model specified in (11). The power of the generalized Kruskal-Wallis U-statistics and Wald-test for different α are shown in Figure 2 Panel C, and the power of the generalized Jonckheere-Terpstra U-statistics and Wald-test for different α are shown in Figure 2 Panel D. In both Panels C and D, the adjusted U-statistic test when standardized to the study population performs better than when standardized to group 1; furthermore, it also outperforms the Wald test, albeit slightly. In addition, we carried out simulation study with group membership specified by a probit model, the results showed that the estimator still constitutes an improvement over the naive approach.

5 | COMT HAPLOTYPES AND THE RISK OF SCHIZOPHRENIA IN AFRICAN AMERICANS

To illustrate the wide variety of analyses that can be done with the adjusted U-statistics we describe here, we analyze data on the association between genetic haplotypes and the risk of schizophrenia in African Americans. Haplotypes (i.e., the adjacent alleles that were contributed by the same parent, e.g. the adjacent paternally-derived alleles) in the catechol-O-methyltransferase (COMT) gene have been associated with Schizophrenia in an Ashkenazi population,[20] and deletions of the region containing COMT cause velocardiofacial syndrome, a syndrome that is associated with a high rate of schizophrenia.[21] Here we test the hypothesis that haplotypes of COMT are associated with schizophrenia using data from the GAIN network study of Schizophrenia, a genome-wide association study with data from 885 African-American case participants and 830 African-American control participants.

Because genotypes, not haplotypes, are observed, it is necessary to exercise some care when making inference about haplotypes. Here we avoid these issues by comparing the similarity between the haplotypes of a case and a control participant to the similarity of haplotypes between two control participants. While it would seem that the unobserved haplotypes are required to measure this similarity, Tzeng et al.[22] showed that the "counting measure" which compares the number of alleles that the haplotypes in one person share in common with the haplotypes of another person, can be calculated using genotype data alone. The similarity Δ_{ij} between the ith and jth individual is then measured by counting the number of alleles the two individuals have in common at each locus, and then summing over the entire region. We consider here the U-statistic of order (1, 2) defined by the kernel

$$K(i;j,j') = \frac{1}{2}I\left[\ \Delta_{ij}\ >\ \Delta_{jj'}\right] + \frac{1}{2}I\left[\ \Delta_{ij}\ \geq\ \Delta_{jj'}\right] + \frac{1}{2}I\left[\ \Delta_{ij'}\ >\ \Delta_{jj'}\right] + \frac{1}{2}I\left[\ \Delta_{ij'}\ \geq\ \Delta_{jj'}\right] \tag{12}$$

which compares whether the similarity between the haplotypes of a case and a control participant differs in similarity from the haplotypes of two control participants. For this kernel, we have $\omega = 1$ by symmetry.

For this analysis, we define the region of interest when calculating similarity to be the 15 SNPs that are genotyped in these data and lie between rs737865 and rs165599 inclusive (the region identified by Shifman et al.[22]). Our null hypothesis is that the distribution of haplotypes among case participants is the same as that among control participants; the alternative is that case participants have a different haplotype distribution, implying that COMT haplotypes are risk factors for schizophrenia.

Unlike the Ashkenazi population, African-Americans are genetically heterogeneous, and individuals vary in their proportion of African and European ancestry. Ancestry is a confounder because it affects both haplotype frequencies and the risk of disease. While ancestry is typically unmeasured, it is well established[3,4] that principal components of genotype data can be used to control for confounding by ancestry. The details of the calculation of these confounding covariates in the GAIN schizophrenia study is described in Allen and Satten,[5] who concluded that 3 principal components were sufficient to control for confounding by population stratification. Thus, we adjust for ancestry using principal components as confounding covariates when calculating the U-statistic just described.

To confirm the performance of our method with the higher-order kernel (12), we conducted a small simulation study. We first confirmed that our approach gave the proper size by generating datasets by sampling with replacement from the GAIN data, and assigning disease status according to the model (11) with Z corresponding to an intercept plus the 3 principal components. When resampling we used $\gamma_g = \hat{\gamma}_g$, the MLE of γ_g, and $\beta_g = 0$. In this scheme, the association between Z and group membership found in the original data is preserved in each replicate dataset, but group membership is unrelated to genotype at the 15 loci we are considering. We used rejection sampling to generate 1,000 datasets each having 100 case and 100 control participants. We found that the empirical size was 5.5% for tests having a nominal size of 5%, indicating good performance of our asymptotics for this sample size (Table 2). The q-q plot of the p-values is also very close to linear (Figure 3).

To ensure that the kernel (12) can discriminate between cases and controls when they truly differ in their genetic similarity, we first decomposed the matrix of similarities using multidimensional scaling with one dimension. We then used the resulting variable u (scaled to have unit variance) as a predictor of case status in model (11), with Y replaced by u. As β_g increases, cases will be more likely to have larger values of u (and hence be increasingly similar to each other), controls will be more likely to have smaller values of u (and hence be increasingly similar to each other) while cases and controls will be increasingly dissimilar. The estimated power of our adjusted U-statistic for datasets having 100 case and 100 control participants using the kernel (12) with $\beta_g = 0.5$, 1.0 and 1.5 is presented in Table 2. Estimates of power were based on 500 simulated datasets for each value of β_g. It is clear that when cases and controls differ in their allele-sharing characteristics, the U-statistic based on kernel (12) can detect these differences.

Using the GAIN data, we tested the association of COMT haplotypes and case status using the kernel described above. Standardizing to the study population, we obtained a test statistic of 0.996, corresponding to a p-value of 0.318. These results suggest that COMT haplotypes are not associated with Schizophrenia in the GAIN study.

6 | DISCUSSION

U-statistics are a powerful tool for statistical analysis for a variety of data types. However, the standard U-statistic that compare samples from two or more populations do not allow for differences in confounding covariates in these populations. Using stratification- or propensity-score based weights, we have introduced adjusted U-statistics that adjust for confounding covariates. Using simulated data, we have shown that our adjusted U-statistics have appropriate size when the only association is spurious (due to confounding covariables) and maintain good efficiency against a properly-specified parametric model when a true association is present. We have also developed a closed form variance estimate for the adjusted U-statistics and provided an R-code for implementing our procedure. Finally, we have demonstrated the use of our adjusted U-statistics using genotype data, testing for genetic association between haplotypes in the COMT gene and schizophrenia in an African-American population in which adjustment for confounding by the proportion of African and European ancestry is required for valid inference.

Although a few studies on adjusted U-Statistics have been appeared in the literature.[11,12,13], there are some fundamental differences between our approach and their methods in terms of the context, scope, and the basic approaches. Jiang et al.[11,12] deal only with the question of estimating the association between genotype and trait in a single population; their starting point is a one-sample U statistic that has the special form of a product of a kernel involving only trait information and a kernel involving only genotype information. Because the genotype kernel is linear in genotype G, they can replace G by $G/E(G|Z)$ to give a test that has mean zero in the presence of confounding covariates Z. However, this trick only works when the kernel is linear in the genotype. In general, if a kernel that is a function of G is not linear in G, it is not possible to replace G by $G/E(G|Z=z)$ (or by $G/P(G|Z)$ for that matter) and get a statistic that is properly centered in the presence of confounding. Finally, because Jiang et al.[11,12] consider only testing a correlation between two variables in a single population, their approach does not generalize to popular two- or multi-sample U-statistics such as the Wilcoxon test we considered here. Rosenbaum[13] proposes a rank-based U-Statistics for matched pairs, where a set of one-sample quantities (the differences between the case value and control value for each matched set) are used. Thus, in effect, it is a one sample problem. It may be possible to use a permutation approach with our test for small sample in certain situations (e.g., when the confounding model is correctly specified). One needs to ensure that the amount of confounding in each permuted dataset remains the same. Further, the weighting model would have to be re-fit for each permutation (see, e.g., Epstein et al.[23]).

APPENDIX A1:: ASYMPTOTIC LINEAR REPRESENTATION OF UA.

We derive here an iid representation of U_a that facilitates calculation of its asymptotic distribution. We use simplified notations whenever possible:

$\boldsymbol{i} = \left(i_1, \ldots, i_B\right), i_1 < i_2 < \cdots < i_B \in \mathscr{I}, \boldsymbol{w}_1(i; \boldsymbol{\gamma}) = \prod_{b=1}^{B} w\left(z_{i_b}, g_{i_b}; \boldsymbol{\gamma}\right), K(\boldsymbol{i}, \boldsymbol{j}),$

$= K\left(y_{i_1}, \ldots, y_{i_B}; y_{j_1}, \ldots, y_{j_D}\right)$

$\boldsymbol{w}_2(\boldsymbol{j}; \boldsymbol{\gamma}) = \prod_{d=1}^{D} w\left(z_{j_d}, g_{j_b}; \boldsymbol{\gamma}\right)$, and $c(\boldsymbol{m}) = \frac{1}{\binom{m_1}{B}} \frac{1}{\binom{m_2}{D}}$. Also let $\mu^* = c(\boldsymbol{m})E[\sum \boldsymbol{w}_1(\boldsymbol{i}, \boldsymbol{\gamma})K(\boldsymbol{i}, \boldsymbol{j})\boldsymbol{w}_2(\boldsymbol{j}, \boldsymbol{\gamma})]$ and $\mu = \mu^*/[\mu_{1,W}]^B[\mu_{2,W}]^D$, where $\mu_{1,W} = plim\, m_1^{-1} \sum_{i: g_i \in \mathscr{I}} w\left(z_i, g_i; \boldsymbol{\gamma}\right), \mu_{2,W} = plim\, m_2^{-1} \sum_{j: g_j \in \mathscr{A}} w\left(z_j, g_j; \boldsymbol{\gamma}\right)$. For any random variable W, W^c denotes its mean corrected version $W - E(W)$. Equalities up to $o_p(n^{-1/2})$ terms will be denoted by $\approx$, where $n = m_1 + m_2$.

Suppose the group membership is related to covariate by the logistic regression model (2) and that the usual regularity conditions (see, e.g., Fahrmeir and Kaufmann[14]) for consistency and asymptotic normality of the maximum likelihood estimator of γ are satisfied. Also assume $E\,[\boldsymbol{w}_1(\boldsymbol{i}, \boldsymbol{\gamma})K(\boldsymbol{i}, \boldsymbol{j})\boldsymbol{w}_2(\boldsymbol{j}, \boldsymbol{\gamma})]^2 < \infty$.

By using a first-order Taylor series expansion, write

$$\begin{aligned} U_a - \mu &= \frac{c(\boldsymbol{m})\sum_{\boldsymbol{i},\boldsymbol{j}} w_1(\boldsymbol{i},\hat{\boldsymbol{\gamma}})K(\boldsymbol{i},\boldsymbol{j})w_2(\boldsymbol{j},\hat{\boldsymbol{\gamma}})}{\left[m_1^{-1}\sum_{i:g_i\in\mathscr{I}} w\left(z_i,g_i;\hat{\boldsymbol{\gamma}}\right)\right]^B\left[m_2^{-1}\sum_{j:g_j\in\mathscr{A}} w\left(z_j,g_j;\hat{\boldsymbol{\gamma}}\right)\right]^D} - \frac{\mu^*}{\left[\mu_{1,W}\right]^B\left[\mu_{2,W}\right]^D} \\ &\approx \ell_1\left\{m_1^{-1}\sum_{i:g_i\in\mathscr{I}} w\left(z_i,g_i;\hat{\boldsymbol{\gamma}}\right) - \mu_{1,W}\right\} + \ell_2\left\{m_2^{-1}\sum_{j:g_j\in\mathscr{A}} w\left(z_j,g_j;\hat{\boldsymbol{\gamma}}\right) - \mu_{2,W}\right\} \qquad \text{(A1)} \\ &\quad + \ell_3\left\{c(\boldsymbol{m})\sum_{\boldsymbol{i},\boldsymbol{j}} w_1(\boldsymbol{i},\hat{\boldsymbol{\gamma}})K(\boldsymbol{i},\boldsymbol{j})w_2(\boldsymbol{j},\hat{\boldsymbol{\gamma}}) - \mu^*\right\}, \end{aligned}$$

where

$$\begin{aligned} \ell_1 &= \frac{-B\mu^*}{\left[\mu_{1,W}\right]^{B+1}\left[\mu_{2,W}\right]^D} = \frac{-B\mu}{\mu_{1,W}}, \\ \ell_2 &= \frac{-D\mu^*}{\left[\mu_{1,W}\right]^{B}\left[\mu_{2,W}\right]^{D+1}} = \frac{-D\mu}{\mu_{2,W}}, \\ \ell_3 &= \frac{1}{\left[\mu_{1,W}\right]^{B}\left[\mu_{2,W}\right]^{D}}. \end{aligned}$$

Next, note that

$$\begin{aligned} & m_1^{-1}\sum_{i:g_i\in\mathscr{I}} w\left(z_i,g_i;\hat{\boldsymbol{\gamma}}\right) - \mu_{1,W} \\ &= m_1^{-1}\sum_{i:g_i\in\mathscr{I}} w\left(z_i,g_i;\hat{\boldsymbol{\gamma}}\right) - m_1^{-1}\sum_{i:g_i\in\mathscr{I}} w\left(z_i,g_i;\boldsymbol{\gamma}\right) + m_1^{-1}\sum_{i:g_i\in\mathscr{I}} w\left(z_i,g_i;\boldsymbol{\gamma}\right) - \mu_{1,W} \qquad \text{(A2)} \\ &\approx \ell_4(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + m_1^{-1}\sum_{i:g_i\in\mathscr{I}}\left(w\left(z_i,g_i;\boldsymbol{\gamma}\right) - \mu_{1,W}\right), \end{aligned}$$

and likewise

$$\begin{aligned} & m_2^{-1}\sum_{j:g_j\in\mathscr{A}} w\left(z_j,g_j;\hat{\boldsymbol{\gamma}}\right) - \mu_{2,W} \\ &= m_2^{-1}\sum_{j:g_j\in\mathscr{A}} w\left(z_j,g_j;\hat{\boldsymbol{\gamma}}\right) - m_2^{-1}\sum_{j:g_j\in\mathscr{A}} w\left(z_j,g_j;\boldsymbol{\gamma}\right) + m_2^{-1}\sum_{j:g_j\in\mathscr{A}} w\left(z_j,g_j;\boldsymbol{\gamma}\right) - \mu_{2,W} \qquad \text{(A3)} \\ &\approx \ell_5(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + m_2^{-1}\sum_{j:g_j\in\mathscr{A}}\left(w\left(z_i,g_i;\boldsymbol{\gamma}\right) - \mu_{2,W}\right), \end{aligned}$$

where

$$\ell_4 = plim\, m_1^{-1} \sum_{i: g_i \in \mathscr{I}} \frac{\partial w\left(z_i, g_i; \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}} \text{ and } \ell_5 = plim\, m_2^{-1} \sum_{j: g_j \in \mathscr{A}} \frac{\partial w\left(z_j, g_j; \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}.$$

Also using a similar triangulation we get

$$\begin{aligned} & c(\boldsymbol{m}) \sum_{\boldsymbol{i}, \boldsymbol{j}} w_1(\boldsymbol{i}, \hat{\boldsymbol{\gamma}}) K(\boldsymbol{i}, \boldsymbol{j}) w_2(\boldsymbol{j}, \hat{\boldsymbol{\gamma}}) - \mu^* \\ & = c(\boldsymbol{m}) \sum_{\boldsymbol{i}, \boldsymbol{j}} w_1(\boldsymbol{i}, \cdot) K(\boldsymbol{i}, \boldsymbol{j}) w_2(\boldsymbol{j}, \cdot) |_{\boldsymbol{\gamma}}^{\hat{\boldsymbol{\gamma}}} + c(\boldsymbol{m}) \sum_{\boldsymbol{i}, \boldsymbol{j}} w_1(\boldsymbol{i}, \boldsymbol{\gamma}) K(\boldsymbol{i}, \boldsymbol{j}) w_2(\boldsymbol{j}, \boldsymbol{\gamma}) - \mu^* \\ & \approx \ell_6(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + \frac{B}{m_1} \sum_{i: g_i \in \mathscr{I}} \left(h_1(i) - \mu^*\right) + \frac{D}{m_2} \sum_{j: g_j \in \mathscr{A}} \left(h_2(j) - \mu^*\right), \end{aligned} \qquad \text{(A4)}$$

by the first order Taylor series expansion and the standard projection argument applied to the generalized U-statistic with kernel

$$h(\boldsymbol{i}, \boldsymbol{j}) = w_1(\boldsymbol{i}, \boldsymbol{\gamma}) K(\boldsymbol{i}, \boldsymbol{j}) w_2(\boldsymbol{j}, \boldsymbol{\gamma}),$$

where $h_1(i) = E\left(h(\boldsymbol{i}, \boldsymbol{j}) \middle| Y_{i_1} = \mid y_i\right)$, $h_2(j) = E\left(h(\boldsymbol{i}, \boldsymbol{j}) \middle| Y_{j_1} = y_j\right)$, and

$$\ell_6 = plim\, c(\boldsymbol{m}) \sum_{\boldsymbol{i}, \boldsymbol{j}} \frac{\partial\left[w_1(\boldsymbol{i}, \boldsymbol{\gamma}) K(\boldsymbol{i}, \boldsymbol{j}) w_2(\boldsymbol{j}, \boldsymbol{\gamma})\right]}{\partial \boldsymbol{\gamma}}.$$

Assume that $\hat{\boldsymbol{\gamma}}$ is estimated by solving the (unbiased) estimating equation

$$\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{g}; \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \boldsymbol{S}_i\left(z_i, g_i; \hat{\boldsymbol{\gamma}}\right) = 0.$$

Then,

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \approx -\boldsymbol{J}^{-1} n^{-1} \sum_{i=1}^{n} \boldsymbol{S}_i\left(z_i, g_i; \boldsymbol{\gamma}\right), \qquad \text{(A5)}$$

where

$$J = plim\, n^{-1} \sum_{i=1}^{n} \frac{\partial \boldsymbol{S}_i\left(z_i, g_i; \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}.$$

Finally, combining (A1)-(A5), we obtain the desired asymptotically linear representation in terms of mean zero independent summands:

$$\begin{aligned} U_a - \mu \approx\ & \ell_1 m_1^{-1} \sum_{i: g_i \in \mathscr{I}} \left(w\left(z_i, g_i; \boldsymbol{\gamma}\right) - \mu_{1,W}\right) + \ell_2 m_2^{-1} \sum_{j: g_j \in \mathscr{A}} \left(w\left(z_i, g_i; \boldsymbol{\gamma}\right) - \mu_{2,W}\right) \\ & + B\ell_3 m_1^{-1} \sum_{i: g_i \in \mathscr{I}} \left(h_1(i) - \mu^*\right) + D\ell_4 m_2^{-1} \sum_{j: g_j \in \mathscr{A}} \left(h_2(j) - \mu^*\right) \\ & -\left(\ell_1\ell_4 + \ell_2\ell_5 + \ell_3\ell_6\right) J^{-1} n^{-1} \sum_{i=1}^{n} \boldsymbol{S}_i\left(z_i, g_i; \boldsymbol{\gamma}\right) \\ & = -B\,\mu\, m_1^{-1} \sum_{i: g_i \in \mathscr{I}} \left(\widetilde{w}_1\left(z_i, g_i; \boldsymbol{\gamma}\right) - 1\right) - D\,\mu\, m_2^{-1} \sum_{j: g_j \in \mathscr{A}} \left(\widetilde{w}_2\left(z_j, g_j; \boldsymbol{\gamma}\right) - 1\right) \\ & + B\, m_1^{-1} \sum_{i: g_i \in \mathscr{I}} \left(\tilde{h}_1(i) - \mu\right) + D\, m_2^{-1} \sum_{j: g_j \in \mathscr{A}} \left(\tilde{h}_2(j) - \mu\right) \\ & + C_n J^{-1} \left(m_1/n\right) m_1^{-1} \sum_{i: g_i \in \mathscr{I}} \boldsymbol{S}_i\left(z_i, g_i; \boldsymbol{\gamma}\right) + C_n J^{-1} \left(m_2/n\right) m_2^{-1} \sum_{j: g_j \in \mathscr{A}} \boldsymbol{S}_j\left(z_j, g_j; \boldsymbol{\gamma}\right) \\ & = \sum_{i=1}^{n} \left(\xi_1(i) I\left[g_i \in \mathscr{I}\right] + \xi_2(i) I\left[g_i \in \mathscr{A}\right]\right) = \sum_{i=1}^{n} \xi_i, \end{aligned} \tag{A6}$$

where $\boldsymbol{C}_n = -(\ell_1\ell_4 + \ell_2\ell_5 + \ell_3\ell_6)$. Hence U_a follows an asymptotically normal distribution with mean μ and the estimated variance given by (8)–(10) with

$$\xi_1(i) = m_1^{-1}\left[-B\,\mu\left(\widetilde{w}_1(z_i, g_i; \boldsymbol{\gamma}) - 1\right) + B\left(\tilde{h}_1(i) - \mu\right)\right] + C_n J^{-1}(1/n) \boldsymbol{S}_i\left(z_i, g_i; \boldsymbol{\gamma}\right)$$

$$\xi_2(j) = m_2^{-1}\left[-D\,\mu\left(\widetilde{w}_2\left(z_j, g_j; \boldsymbol{\gamma}\right) - 1\right) + D\left(\tilde{h}_2(j) - \mu\right)\right] + C_n J^{-1}(1/n) \boldsymbol{S}_j\left(z_j, g_j; \boldsymbol{\gamma}\right).$$

Here

$$\begin{aligned} \tilde{h}_1(i) = & \frac{\widetilde{w}_1\left(z_i, g_i; \boldsymbol{\gamma}\right)}{\binom{m_1 - 1}{B - 1}\binom{m_2}{D}} \sum_{\left\{i_2, \ldots, i_B \middle| g_{i_b} \in \mathscr{I}\right\}} \sum_{\left\{j_1, \ldots, j_D \middle| g_{j_d} \in \mathscr{A}\right\}} \left\{\prod_{b=2}^{B} \widetilde{w}_1\left(z_{i_b}, g_{i_b}; \boldsymbol{\gamma}\right)\right\} \\ & K\left(y_i, y_{i_2}, \ldots, y_{i_B}; y_{j_1}, \ldots, y_{j_D}\right)\left\{\prod_{d=1}^{D} \widetilde{w}_2\left(z_{j_d}, g_{j_d}; \boldsymbol{\gamma}\right)\right\}, \end{aligned} \tag{A7}$$

when $i \in \mathscr{I}$ and $\tilde{h}_1(i) = 0$ otherwise; and

$$\tilde{h}_2(j) = \frac{\widetilde{w}_2\left(z_j, g_j; \boldsymbol{\gamma}\right)}{\binom{m_1}{B}\binom{m_2 - 1}{D - 1}} \sum_{\left\{i_1, \ldots, i_B \middle| g_{i_b} \in \mathscr{I}\right\}} \sum_{\left\{j_2, \ldots, j_D \middle| g_{j_d} \in \mathscr{A}\right\}} \left\{\prod_{b=1}^{B} \widetilde{w}_1\left(z_{i_b}, g_{i_b}; \boldsymbol{\gamma}\right)\right\} \\ K\left(y_{i_1}, y_{i_2}, \ldots, y_{i_B}; y_j, y_{j_2}, \ldots, y_{j_D}\right)\left\{\prod_{d=2}^{D} \widetilde{w}_2\left(z_{j_d}, g_{j_d}; \boldsymbol{\gamma}\right)\right\}, \tag{A8}$$

and

$$\boldsymbol{C}_n \equiv -\left(\ell_1\ell_4 + \ell_2\ell_5 + \ell_3\ell_6\right) \\ = \frac{B\mu}{m_1} \sum_{i: g_i \in \mathscr{I}} \widetilde{w}_1\left(z_i, g_i; \boldsymbol{\gamma}\right) \frac{\partial\, log\, w\left(z_i, g_i; \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}} + \frac{D\mu}{m_2} \sum_{j: g_j \in \mathscr{A}} \widetilde{w}_2\left(z_j, g_j; \hat{\boldsymbol{\gamma}}\right) \frac{\partial\, log\, w\left(z_j, g_j; \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}} \\ - \frac{1}{\binom{m_1}{B}\binom{m_2}{D}} \sum_{\boldsymbol{i}, \boldsymbol{j}} \widetilde{w}_1(\boldsymbol{i}, \boldsymbol{\gamma}) K(\boldsymbol{i}; \boldsymbol{j})\, \widetilde{w}_2(\boldsymbol{j}, \boldsymbol{\gamma}) \left(\frac{\partial\, log\, w_1(\boldsymbol{i}; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} + \frac{\partial\, log\, w_2(\boldsymbol{j}; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right). \tag{A9}$$

$h_1(i)$ and $h_2(j)$ in (A4) and (A6) have the form as $\tilde{h}_1(i)$ and $\tilde{h}_2(j)$ in (A7) and (A8) except that $\widetilde{w}_1(z, g; \boldsymbol{\gamma})$ and $\widetilde{w}_2(z, g; \boldsymbol{\gamma})$ are replaced by $w(z, g; \boldsymbol{\gamma})$. To compute $\xi(i)$ for the purpose of estimating variance, $\boldsymbol{\gamma}$ is plugged by its estimate $\hat{\boldsymbol{\gamma}}$, and μ is replaced by its sample analogue. Note, however, in the centering of test statistics in Sections 2 and 3, the mean of U_a, say ω, is taken under the null hypothesis. In the case that the weight is chosen as (5) standardizing Z to the distribution in the study population, then

$$\frac{\partial\, log\, w\left(z_i, g_i; \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}_g} = pr\left(G = g \middle| Z = z_i, \gamma\right) z_i - I_{\left[g_i = g\right]} z_i \cdot \ (g = 2, \ldots, \mathscr{G})$$

In the case that the weight is chosen to standardize to group 1 as defined in (6), then

$$\frac{\partial\, log\, w\left(z_i, g_i; \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}_g} = -I_{\left[g_i = g\right]} z_i \cdot \ (g = 2, \ldots, \mathscr{G}).$$

APPENDIX A2:: PROOF THAT THE ADJUSTED U-STATISTIC HAS ASYMPTOTICALLY MEAN ω UNDER THE NULL HYPOTHESIS FOR STRATIFIED SAMPLING

First, consider a weighted mean

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f\left(y_i, z_i\right)}{pr\left(G = g_i \middle| Z = z_i\right)} = \sum_{g=1}^{\mathscr{G}}\frac{n_g}{n}\left\{\frac{1}{n_g}\sum_{i \in \mathscr{S}_g}\frac{f\left(y_i, z_i\right)}{pr\left(G = g \middle| Z = z_i\right)}\right\}, \quad \text{(A10)}$$

for a function f with $E(|f(Y, Z)|) < \infty$. Assume we have sampled by group (i.e., cases and controls for a retrospective study, exposed and unexposed people for a follow-up study). Then the null distribution of (Y, Z) may differ for different values of G. As a result, the weighted sum in (A10) may not converge to the same quantity if we change the proportion of persons from each group. In order to show that RHS (A10) is independent of how the groups are assembled, we need to consider the conditional distribution of (Y, Z) given G under the null hypothesis, which we write as

$$pr(Y, Z|G) = pr(Y|Z, G)pr(Z|G) = pr(Y|Z)pr(Z|G),$$

where we have used the null hypothesis for the first factor in the middle equality. Now express

$$pr(Z|G) = \int pr(Z|G, S)dF(S|G) = \int pr(Z|S)dF(S|G),$$

where the last equality uses the balancing score property that $Z \perp G|S$ which holds for both the stratification score[5] when G is disease status, and the propensity score[7] when G is exposure. Under the null then we have

$$pr(Y, Z|G) = pr(Y|Z)\int pr(Z|S)dF(S|G)\,. \quad \text{(A11)}$$

Note that only dependence on G in the RHS (A11) is through the distribution of S in each group. Thus, if we can weight the data so that the distribution function $F(S|G)$ is the same in each group, we will find that $pr(Y, Z|G)$ is the same in each group as well. Since S is a function of Z, a sufficient condition to ensure that $pr(S|G)$ is the same across groups is that after weighting, the distribution of Z be the same in each group. Thus, we consider whether $pr_w(Z|G)$, the effective distribution of Z in group G after weighting, is independent of group, where

$$pr_w(Z|G) \propto \frac{pr(Z|G)}{pr(G|Z)} = \frac{pr(Z)}{pr(G)}, \quad \text{(A12)}$$

where $pr(Z)$ and $pr(G)$ are the distributions of Z and G in the study population, i.e., $pr(G = g) = \frac{n_g}{n}$ and

$$pr(Z) = \sum_{g} pr(Z \mid G = g) pr(G = g) \,. \qquad \text{(A13)}$$

Thus, after normalizing, we see that inverse-probability-of-group-membership weighting gives the same distribution of covariates Z in each group. Therefore, as argued before, the distribution of (Y, Z) is the same in each group and hence the expected value is the same in each group.

We can now immediately extend this argument to kernels of order (1,1). First write

$$\frac{1}{m_1}\frac{1}{m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{K\left(y_i, y_j\right)}{pr_1\left(G_1 = g_i \middle| Z = z_i\right)pr_2\left(G_2 = g_j \middle| Z = z_j\right)} =$$

$$\sum_{g_1 \in \mathscr{I}}\sum_{g_2 \in \mathscr{A}}\frac{m_{g_1}}{m_1}\frac{m_{g_2}}{m_2}\left\{\frac{1}{m_{g_1}}\frac{1}{m_{g_2}}\sum_{i \in \mathscr{S}_{g_1}}\sum_{j \in \mathscr{S}_{g_2}}\frac{K\left(y_i, y_j\right)}{pr_1\left(G_1 = g_1 \middle| Z = z_i\right)pr_2\left(G_2 = g_2 \middle| Z = z_j\right)}\right\}$$

where the subscript s on pr_s denotes the probability law that applies to sample $s = 1,2$ used to calculate the two-sample U-statistic. As before, we need to show that the conditional expected values of each term is the same under the null model. If we can do this, then the result is proved. However, since the data from the two observations are independent, the one-sample results shown above apply immediately. Hence, the inverse-probability-of-group-membership weights completely account for differences in the distribution of Y and Z across groups, and hence our adjusted U-statistic has asymptotic mean ω under the null. Further, the distribution of Z (and hence S) that characterizes the population that characterizes each group after weighting is given in (A13). If we additionally weight data from each population by $pr(G = 1|Z = z)$, the argument leading to (A12) easily shows that $pr_w(Z|G) \propto pr(Z|G = 1)pr(G = 1)/pr(G)$, indicating that weighting by $pr(G = 1|Z = z)/pr(G = g|Z = z)$ corresponds to standardizing each group to have the same distribution of Z values found in group 1.

References

1. Mann HB & Whitney DR (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 18, 50–60.
2. Hoeffding W (1948). A class of statistics with asymptotically normal distribution. Ann. Math. Statist 19, 293–325.
3. Chen HS, Zhu X, Zhao H & Zhang S (2003). Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 67, 250–264. [PubMed: 12914577]
4. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, & Reich D(2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–909. [PubMed: 16862161]

5. Allen AS & Satten GA (2011). Control for confounding in case-control studies using the stratification score, a retrospective balancing score. Am J Epidemiol 173, 752–760. [PubMed: 21402731]

6. Epstein MP, Allen AS & Satten GA (2007). A simple and improved correction for population stratification in case-control studies. Am J Hum Genet 80, 921–930. [PubMed: 17436246]

7. Rosenbaum PR & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

8. Imbens GW(2000). The role of the propensity score in estimating dose-response functions.. Biometrika 87, 706–710.

9. Rosenbaum PR(1987). Model-based direct adjustment. J Amer Stat Assoc 82, 387–394.

10. Lunceford JK & Davidian M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 23, 2937–2960. [PubMed: 15351954]

11. Jiang Y, Li N & Zhang H (2014). Identifying genetic variants for addiction via propensity score adjusted generalized Kendall's tau. J Amer Stat Assoc 109, 905–930. [PubMed: 25382885]

12. Jiang Y& Zhang H (2011). Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. Genetic Epidemiology 35, 125–132. [PubMed: 21254220]

13. Rosenbaum PR (2011). A new u-Statistic with superior design sensitivity in matched observational studies. Biometrics 67, 1017–1027. [PubMed: 21175557]

14. Fahrmeir L & Kaufann H (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Annals of Statistics 13, 342–368.

15. Sato T & Matsuyama Y (2003). Marginal structural models as a tool for standardization. Epidemiology 14, 680–686. [PubMed: 14569183]

16. Epstein MP, Allen AS & Satten GA(2007). A simple and improved correction for population stratification in case-control studies. American Journal of Human Genetics 80, 921–930. [PubMed: 17436246]

17. Randles RH & Wolfe DA(1979). Introduction to the theory of nonparametric statistics, vol.1 Wiley New York.

18. Yee TW (2010). The VGAM package for categorical data analysis. Journal of Statistical Software 32.

19. Drake C (1993). Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 49, 1231–1236.

20. Shifman S, Bronstein M, Sternfeld M, Pisanté-Shalom A, Lev-Lehman E, Weizman A, Reznik I, Spivak B, Grisaru N, Karp L, Schiffer R, Kotler M, Strous RD, Swartz-Vanetik M, Knobler HY, Shinar E, Beckmann JS, Yakir B, Risch N, Zak NB & Darvasi A (2002). A highly significant association between a COMT haplotype and schizophrenia. Am J Hum Genet 71, 1296–1302. [PubMed: 12402217]

21. Coman IL, Gnirke MH, Middleton FA, Antshel KM, Fremont W, Higgins AM, Shprintzen RJ & Kates WR (2010). The effects of gender and catechol O-methyltransferase (COMT) Val108/158Met polymorphism on emotion regulation in velo-cardio-facial syndrome (22q11.2 deletion syndrome): An fMRI study. Neuroimage 53,1043–50. [PubMed: 20123031]

22. Tzeng JY, Devlin B, Wasserman L & Roeder K (2003). On the identification of disease mutations by the analysis of haplotype similarity and goodness-of-fit. The American Journal of Human Genetics 72, 891–902. [PubMed: 12610778]

23. Epstein MP, Duncan R, Jiang Y, Conneelly KN, Allen AS & Satten GA (2012). A novel permutation procedure to correct for confounders in case-control studies, including tests of rare variation. The American Journal of Human Genetics 91(2), 215–223. [PubMed: 22818855]

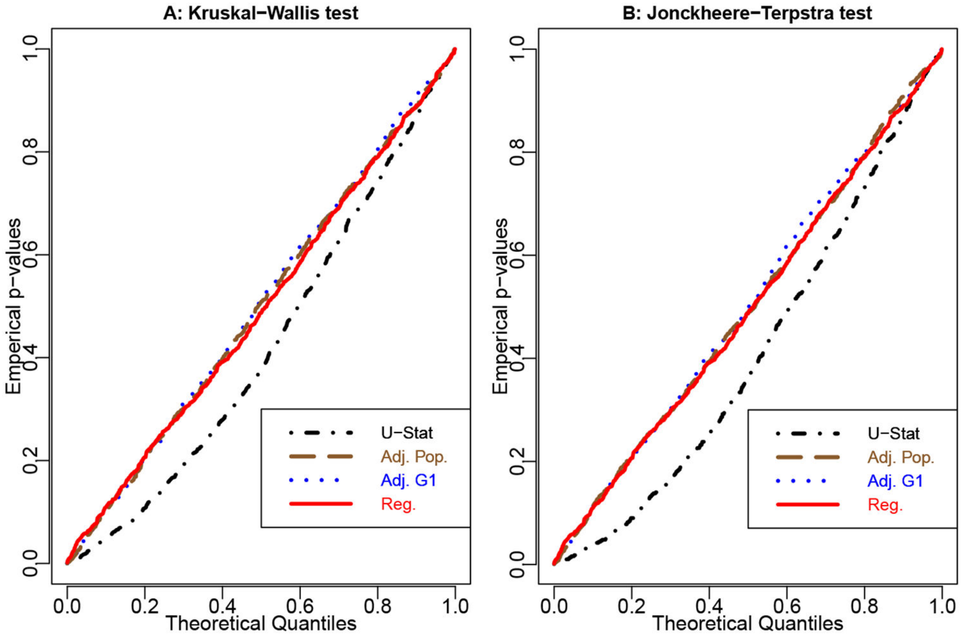

FIGURE 1.

Comparison of empirical p-values and theoretical (uniform) p-values for the Kruskal-Wallis type test (Panel A) and Jonckheere-Terpstra type test (Panel B). Brown (long-dashed curve) corresponds to standardization to the study population, blue (dotted curve) is standardization to the group 1, red (solid curve) is the parametric model, and black (dash-dotted curve) is the naive U-statistic that does not account for confounding.

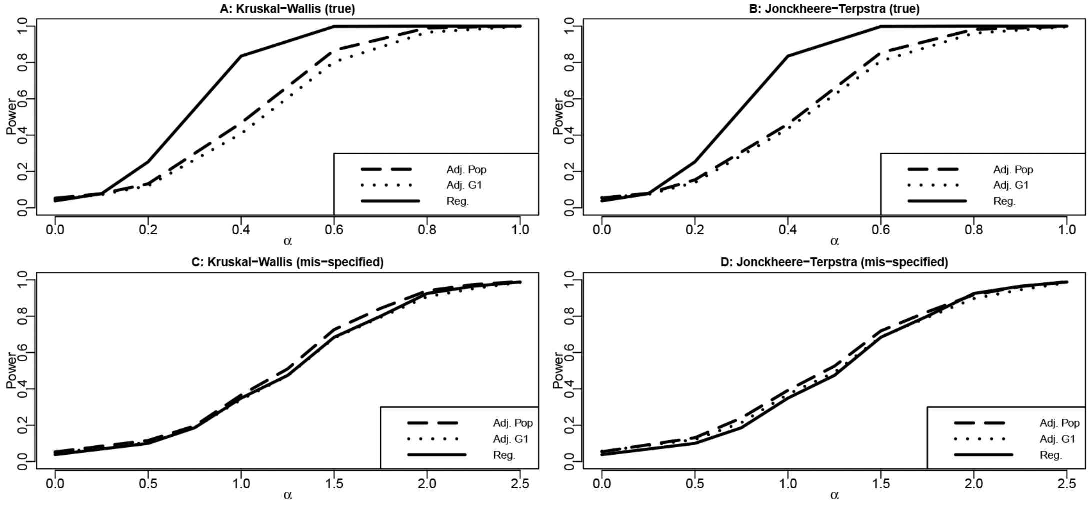

FIGURE 2.

Power of the adjusted U-statistic for the Kruskal-Wallis type test (Panel A) and Jonckheere-Terpstra type test (Panel B) when the parametric model is correctly specified, and the power of the adjusted U-statistic for the Kruskal-Wallis type test (Panel C) and Jonckheere-Terpstra type test (Panel D) when the parametric model is mis-specified. Solid curve is the Wald test for the parametric model. Long-dashed and dotted curves are adjusted U-statistics that standardize to the study population and group 1, respectively. The parameter α determines strength of the association. When $\alpha = 0$ (no association) the power corresponds to the size of the test. All tests have 3 groups and 2 degrees of freedom.

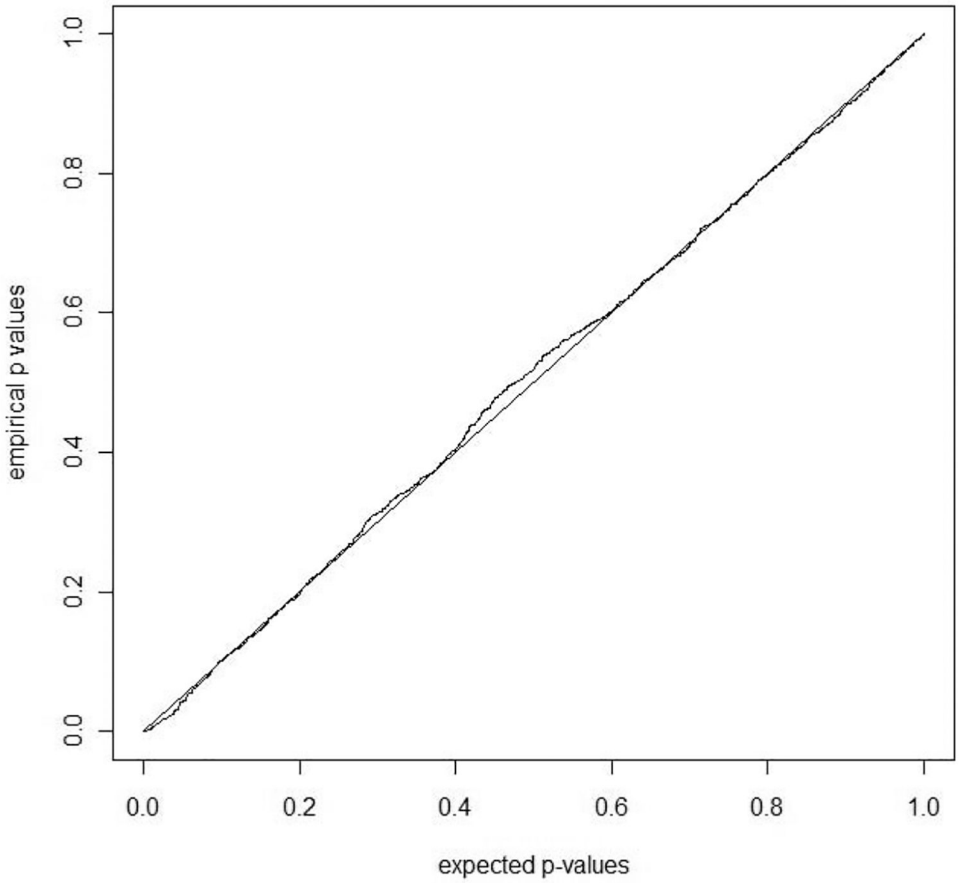

FIGURE 3.

Expected vs. empirical p-values under null hypothesis using Kernel (12) and simulation data based on the COMPT study.

TABLE 1

Empirical size from 1,000 simulated data sets for tests having a nominal size of 5%.

Analysis	**Standardization**	**Size**
Kruskal-Wallis test (U-statistic)	None	0.137
Kruskal-Wallis test (Adjusted U-statistic)	Study Population	0.053
Kruskal-Wallis test (Adjusted U-statistic)	Group 1	0.047
Jonckheere-Terpstra test (U-statistic)	None	0.155
Jonckheere-Terpstra test (Adjusted U-statistic)	Study Population	0.054
Jonckheere-Terpstra test (Adjusted U-statistic)	Group 1	0.057
Wald Test, Logistic Regression	Not applicable	0.038

TABLE 2

Size and power for simulated data sets for the genetic example.

β_g	**Power**
0.0	0.055
0.5	0.146
1.0	0.806
1.5	0.976