Harvard University
Harvard University Biostatistics Working Paper Series

Year 2012 *Paper* 150

Flexible Covariate-adjusted Exact Tests for Randomized Studies

Alisa J. Stephens* Eric J. Tchetgen Tchetgen†
Victor De Gruttola‡

*Harvard University, alisa.j.stephens@gmail.com
†Harvard University, etchetge@hsph.harvard.edu
‡Harvard University, degrut@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper150

Flexible covariate-adjusted exact tests for randomized studies

Alisa J. Stephens, Eric J. Tchetgen Tchetgen, and Victor De Gruttola

Abstract

Incorporating auxiliary covariates in the analysis of randomized trials can increase power, but questions remain about how to preserve type I error when incorporating such covariates in a flexible way, particularly in small samples. This paper investigates properties of covariate-adjusted tests for both independent and multivariate outcomes. Through simulation, we evaluate several covariate-adjusted tests of intervention effects when baseline covariates are selected adaptively and the number of randomized units is small. We demonstrate that randomization inference preserves type I error under model selection while tests based on asymptotic theory break down. We also demonstrate that covariate adjustment generally increases power, except at extremely small sample sizes using liberal selection procedures. Methods are illustrated by application to data on the *Young Citizens* study, a cluster randomized trial of behavioral HIV intervention.

1 Introduction

In randomized trials the primary goal is to compare the effects of different interventions on some outcome of interest. In addition to the treatment assignment and outcome, data on baseline covariates, such as demographics or biomarkers, are typically collected. To protect type I error, methods for including baseline covariates in analyses, whether as stratification factors or in regression models, are generally precisely defined. Recently, methods have been developed to allow for more flexible model selection without loss of protection of type error, at least asymptotically (Tsiatis et al. (2008); Zhang et al. (2008); Stephens et al. (2012a)). Several studies have demonstrated that new methods permitting flexible use of baseline correlates of the outcome in analysis improve power and efficiency in treatment effect estimation (Tsiatis

et al. (2008); Zhang et al. (2008); Stephens et al. (2012a)). Nonetheless, in small samples, additional variability introduced by flexible model selection may fail to preserve type I error and also result in loss of power and efficiency compared to unadjusted analyses. In this paper, we evaluate several flexible covariate adjustment methods for studies with small numbers of randomized units. We examine the validity of adjusted tests through investigation of type I error and measure improvement over unadjusted tests by comparing power.

Consider a randomized trial in which n independent and identically distributed units $O_i = (Y_i, A_i, \mathbf{X}_i)$ are sampled from a population, where Y_i denotes the outcome of interest, A_i the random treatment assignment such that A_i=1, ..., K, and $\mathbf{X}_i$ the set of baseline covariates. For cluster-randomized or longitudinal trials, $\mathbf{Y}_i$ represents a multivariate outcome vector for individuals within the same randomized group or repeated measurements on a single randomized subject. In the context of multivariate outcomes, we consider settings where the treatment assignment is a scalar shared by measurements within the same cluster or subject. The primary analysis for most randomized trials compares outcomes Y_i among subjects assigned to different levels of treatment A_i. For scalar outcomes, tests comparing some feature of $f_{a^*}(Y)$, the distribution of Y under treatment a^*, are used to assess the statistical significance of observed differences in outcomes across treatment groups. The two-sample t-test, Wilcoxon test, and their extensions for more than two groups are examples of commonly used methods. When outcomes are multivariate, modified versions of these tests are available to adjust standard errors for correlation among multiple measurements within the same randomized unit (Klar and Donner (2000)).

Regression analysis may also be used to evaluate treatment effects. The effect of a binary treatment on the marginal mean of Y may be assessed through assuming the generalized linear model

$$E[Y_i|X_i, A_i] = g(\beta_0 + \beta_1 A_i), \tag{1}$$

where g^{-1} is a link function, and β is estimated through semiparametric estimating equations or fully parametric maximum likelihood inference. The effect of treatment on the marginal mean outcome $E[Y_i|A_i = a]$ is evaluated through testing $H_0 : \beta_1 = 0$. Under randomization, this

test is equivalent to testing for no average causal effect of treatment on Y_i. When outcomes are multivariate, Y_i in (1) is replaced by Y_{ij}, denoting the jth outcome of the ith randomized unit for $i = 1, 2, ..., n$ and $j = 1, ...m_i$, where $M = \sum_{i=1}^{n} m_i$ is the total number of observations. For a semiparametric approach, generalized estimating equations (GEE) that account for correlation in responses may be used to obtain consistent parameter and standard error estimates. Regression methods naturally incorporate baseline covariates by assuming the adjusted mean model (AMM)

$$E[Y_i|\mathbf{X}_i, A_i] = g(\beta_0 + \beta_1^* A_i + \beta_X^{\mathrm{T}} \mathbf{X}_i). \tag{2}$$

When g is the identity link and the true model does not contain any treatment-covariate interactions, independence of A_i and $\mathbf{X}_i$, which results from randomization, guarantees that the adjusted estimator $\hat{\beta}_1^*$ is a consistent estimator of β_1. Moreover, it can be shown that $var(\hat{\beta}_1^*) \leq var(\hat{\beta}_1)$, where $\hat{\beta}_1$ is the unadjusted estimator, even under misspecification of the exact form of $\beta_X^{\mathrm{T}} \mathbf{X}_i$ in (2) (**?**Tsiatis08)). For other link functions $\hat{\beta}_1^*$ is not consistent for β_1, nor does the addition of baseline covariates to the assumed mean model guarantee efficiency improvement. To address this concern when estimating β in marginal model (1), Zhang et al. (2008) advocate using a class of augmented estimators. Augmented estimators are derived from semiparametric theory and involve augmenting standard estimating functions by subtracting their Hilbert space projection onto the span of the scores of the treatment mechanism. Semiparametric theory provides theoretical justification for efficiency improvement of augmented estimators in large samples irrespective of the link function g and only assuming model (1) holds. Stephens et al. (2012a) demonstrated how augmentation may be used for clustered or longitudinal data by augmenting generalized estimating equations. The same authors also presented the locally efficient augmented estimator under model (1) (Stephens et al. (2012b)). Augmented inference relies on asymptotic theory and therefore requires a fairly large number of randomized units. In large samples, model selection variability for baseline covariates is small provided the number of covariates is not large; in small samples, however, flexible covariate selection induces additional variability that may lead to variance underestimation and loss of efficiency.

To avoid reliance on asymptotic theory, Rosenbaum (2002) extended the randomization theory of Fisher (1935) to propose an exact covariate-adjusted test that does not assume a particular distribution for outcomes or that the observed data are a random sample from some unobserved population of independent units. Randomization inference considers the potential outcomes y_{a_i} under each treatment level, where the observed outcome Y_i for a subject assigned to treatment $A_i = a$ is that subject's potential outcome y_{a_i}. The lowercase notation emphasizes that potential outcomes y_{a_i} are fixed. Under the sharp null $H_0 : y_a = y^*$ for all a, and thus $Y_i = y_i^*$, resulting in independent units $\tilde{O}_i = (y_i, A_i, \mathbf{x}_i)$, where only the treatment assignment A_i is random. Rosenbaum (2002) discussed the potential outcomes framework in detail. The null distribution of the test statistic is obtained through permutation of A_i. The test proposed by Gail et al. (1988) approximates the exact test by standardizing the observed test statistic by its randomization-based variance and comparing to the standard normal distribution. Post model-selection inference based on the Gail et al. and Rosenbaum approaches has not been investigated; below, we consider settings where model selection is used to determine covariates that explain variability in y_i. Adaptive selection of baseline covariates may be particularly useful when $\mathbf{x}_i$ is high-dimensional or prior knowledge is not available to inform covariate adjustment. Further improvement in small-sample inference may be possible from higher order approximations of the distribution of a class of randomization test statistics (Bickel and Zwet (1978)), but this theory has not yet been evaluated in practice.

Details of the four covariate-adjusted tests: I) Adjusted mean models (AMM), II) Augmented marginal model, III) Approximate exact, and IV) Exact (permutation) are discussed in Section 2. Inference for independent and correlated outcomes is presented. In Section 3, the small sample properties of covariate-adjusted tests are evaluated through simulation. Methods are illustrated through application to the *Young Citizens* study in Section 4. Finally, we summarize our results and provide recommendations for practical use in Section 5.

2 Methods

We consider four methods of covariate-adjusted hypothesis testing: I) Wald test of β_1^* in the adjusted mean model (2), II) Wald test of β_1 in marginal model (1), in which estimating equations are augmented to include baseline covariates, III) approximate exact test, and IV) the exact test. This list is not comprehensive, but does include widely-recognized classical and modern methods. Each test is first presented for independent outcomes and followed by generalizations for dependent data.

2.1 Independent Outcomes

Method Ia: Wald test of β_1^* in model (2)

Assuming model (2) holds, parameters β and respective standard errors are estimated via maximum likelihood or semiparametric estimating equations. The null hypothesis $H_0 : \beta_1^* = 0$ is evaluated through the test statistic $T_c = \frac{\hat{\beta_1}^*}{SE(\hat{\beta_1}^*)}$.

Method IIa: Wald test of β_1 in model (1) with augmented estimating equations (Tsiatis et al. (2008); Zhang et al. (2008); van der Laan and Robins (2003))

Unlike inference based on the AMM, the augmentation method assumes model (1). Predicted values from a working model for the conditional mean $E[Y_i|\mathbf{X}_i, A_i]$ are incorporated in estimating equations that are solved to estimate β. Consistent estimates of β_1 are obtained even if $E[Y_i|\mathbf{X}, A_i]$ is misspecified, demonstrating a special case of double robustness (van der Laan and Robins (2003)).

Tests of $H_0 : \beta_1 = 0$ are based on the test statistic $T_a = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}$, where $\hat{\beta}_1$ is the solution of the augmented estimating equations

$$\sum_{i=1}^{n} \psi_a(O_i;\beta) = \sum_{i=1}^{n} \Bigg[h(A_i;\beta)\{Y_i - g(A_i;\beta)\} - \sum_{a=1}^{K} \{I(A_i = a) - \pi_a\}\{h(a;\beta)(E[Y_i|X_i, A_i = a] - g(a;\beta))\} \Bigg] = \mathbf{0},$$

and π_a denotes $P(A_i = a)$. In practice ψ_a is evaluated by $\hat{\psi}_a$, where the true regression

$E[Y_i|\mathbf{X}_i, A_i = a]$ is approximated by the working model $E[Y_i|\mathbf{X}_i, A_i = a] = d(\mathbf{X}_i; \eta_a)$ evaluated under an estimate $\hat{\eta}_a$. As implied by the subscript a, the regression for augmented estimators conditions on the treatment assignment. Alternatively, $E[Y_i|\mathbf{X}_i, A = a]$ may be estimated separately in each treatment arm, resulting in K regression models that do not contain indicators for treatment. The variance of $\hat{\beta}_1$ is estimated by the sandwich variance estimator $\hat{Var}(\hat{\beta}_1) = C\left[\left(\sum_{i=1}^{n} \frac{dh(A_i;\beta)}{d\beta^{\mathrm{T}}} D_i\right)^{-1} \sum_{i=1}^{n} \left[\psi_a(O_i;\beta)^{\otimes 2}\right] \left(\sum_{i=1}^{n} \frac{dh(A_i;\beta)}{d\beta^{\mathrm{T}}} D_i\right)^{-1}\right]$, where $D_i = \frac{dg(A_i;\beta)}{d\beta^{\mathrm{T}}}$, and $C = \{(n_0 - p_0 - 1)^{-1} + (n_1 - p_1 - 1)^{-1}\}/\{(n_0 - 1)^{-1} + (n_1 - 1)^{-1}\}$ is incorporated to account for finite-sample variability attributable to augmenting. In C, n_a is the sample size in treatment arm a and p_a is the dimension of η_a for the working covariate-adjustment model.

Method IIIa: Approximation of the Exact Test

The approximation of the exact test considers the $H_0 : y_a = y_*$ for all a. To test H_0 we construct the test statistic

$$T_s = \frac{S}{\sqrt{Var(S|y, \mathbf{x})}}, \text{ where } S = \sum_{i=1}^{n} (A_i - \pi) w_i,$$

and $Var(S|y, \mathbf{x})$ is shown in (3). Baseline covariates are incorporated by setting $w_i = \hat{\varepsilon}_i = y_i - d(\mathbf{x}_i; \hat{\eta})$, the residual from the working mean model $d(\mathbf{x}_i; \hat{\eta})$, which estimates the true regression model $E[y_i|\mathbf{x}_i] = f(\mathbf{x}_i; \eta)$ under the sharp null. For unadjusted analysis, w_i=y_i. We intentionally omit the subscript a on the regression function as a reminder that under the sharp null, y_i cannot depend on treatment, so A_i is excluded from the proposed working model. The variance of S is calculated by

$$Var(S|y, \mathbf{x}) = \pi(1 - \pi) \sum_{i=1}^{n} w_i^2 + \overbrace{\left(\pi \frac{n/2 - 1}{n - 1} - \pi^2\right) \sum_{i \neq i'} w_i w_{i'}}^{(Q)}, \tag{3}$$

and significance is determined by comparing $|T_a|$ to the standard normal distribution.

Term Q in $Var(S|\mathbf{y}, \mathbf{x})$ is nonzero when the total number of subjects assigned to each treatment is fixed. This typically applies in trials with small samples, where matching and blocked randomization strategies are employed to prevent imbalances in treatment allocation

that may occur with unrestricted random assignment. Under such randomization, the vector $\mathbf{A} = (A_1, A_2, ..., A_n)$ follows a hypergeometric distribution, where the probability of being assigned to treatment for a particular subject is affected by the other subjects' treatment assignments. When w_i is the residual from a working model for $E[y_i|\mathbf{x}_i]$, Q ≈ 0, as $E[\varepsilon_i|\mathbf{x}_i]$=0, and $\varepsilon_i \perp \varepsilon_{i'}$. If considering the unadjusted outcomes Y_i, failure to include Q may result in gross variance overestimation and extremely conservative testing for small n. In large samples, Q ≈ 0 for $w_i = \hat{\varepsilon}_i$ or $w_i = y_i$.

For the class of statistics defined by $T = \sum_{i=1}^{n} A_i c_i$, where c_i is a score, Bickel and Zwet (1978) determined a higher-order approximation for the distribution of the standardized statistic T^*, given by

$$P(T^* < t) = \Phi(t) - \frac{\phi(t)}{\pi(1-\pi)}\left[\frac{\pi(1-\pi)}{2n}H_1(t) + \frac{\sqrt{\pi(1-\pi)}(1-2\pi)}{6}\frac{\sum_{i=1}^{n}(c-c.)^3}{\left\{\sum_{i=1}^{n}(c-c.)^2\right\}^{3/2}}H_2(t) + \left\{\frac{1-6\pi+6\pi^2}{24}\frac{\sum_{i=1}^{n}(c-c.)^4}{\left\{\sum_{i=1}^{n}(c-c.)^2\right\}^2} - \frac{(1-2\pi)^2}{8n}\right\}H_3(t) + \frac{(1-2\pi)^2}{72}\frac{\left\{\sum_{i=1}^{n}(c-c.)^3\right\}^2}{\left\{\sum_{i=1}^{n}(c-c.)^2\right\}^3}H_5(t)\right]$$

The expansion suggests that a higher order accurate quantile of the distribution of the test statistic may be found by solving for Z^*_α such that $P(T < Z^*_\alpha) = 1 - \alpha/2$ for two-sided tests.

Method IVa: Exact Test

The exact test also applies to the hypothesis $H_0 : y_a = y_*$ for all a; the null distribution of $T_p = S$ is calculated by permuting the treatment assignment A_i among subjects. For each permutation, the test statistic T_p is calculated under the permuted treatment assignment A_b, resulting in distribution of statistics $T_p(A_b)$. The exact null distribution is often estimated by conducting B permutations for large B, and a p-value is obtained by $p_B = \frac{1}{B} = \sum_{b=1}^{B} I(|T_p(A_b)| > |T_p|)$. For a level α test, we reject the sharp null of no treatment effect when $p_B < \alpha$.

2.2 Dependent Outcomes

For clustered outcomes, we consider modifications of the univariate tests that accommodate correlation in responses.

Method Ib: Wald test of β_1^* in model (2) using GEE (Liang and Zeger (1986))

To accommodate correlation in outcomes within a cluster, generalized estimating equations may be constructed assuming model (2) holds. The adjusted treatment effect β_1^* is estimated by solving the generalized estimating equations

$$\sum_{i=1}^{n} \mathbf{D}_i \mathbf{V}_i^{-1} [\mathbf{Y}_i - \mathbf{g}(A_i, \mathbf{X}; \beta)] = \mathbf{0}, \tag{4}$$

where $\mathbf{D}_i = \frac{d\mathbf{g}(A_i, \mathbf{X}; \beta))}{d\beta^{\mathrm{T}}}$, $\mathbf{V}_i = V_i(\phi)^{1/2} \mathbf{R} V_i(\phi)^{1/2}$. The working covariance $\mathbf{V}_i$ is determined by the $m_i \times m_i$ correlation matrix $\mathbf{R}$ and diagonal variance matrix $V_i(\phi)$. The variance of $\hat{\beta}$ is calculated by the sandwich variance estimator,

$$\hat{var}(\hat{\beta}) = \left(\sum_{i=1}^{n} \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^{n} \left[\mathbf{D}_i \mathbf{V}_i^{-1} \{\mathbf{Y} - \mathbf{g}(A_i, \mathbf{X}; \beta)\} \right]^{\otimes 2} \right) \left(\sum_{i=1}^{n} \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1}, \tag{5}$$

and T_c is calculated to evaluate $H_0 : E[\mathbf{Y}_i | \mathbf{X}_i, A_i = 1] = E[\mathbf{Y}_i | \mathbf{X}_i, A_i = 0]$.

Method IIb: Wald test of β_1 in model (1) using augmented GEE {Stephens et al. (2012a)}

Assuming marginal model (1), augmented estimating equations are formed by

$$\sum_{i=1}^{n} \psi_a(O_i; \beta, \eta) = \sum_{i=1}^{n} \Bigg\{ \mathbf{D}_i \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mathbf{g}(A_i; \beta)\} - \sum_{a=1}^{K} \{I(A_i = a) - \pi_a\} [\mathbf{D}_i(a) \mathbf{V}_i^{-1}(a) \{\mathbf{d}\{\mathbf{X}_i; \eta_a) - \mathbf{g}(a; \beta)\}] \Bigg\} = \mathbf{0},$$

where $\mathbf{d}(\mathbf{X}_i; \eta_a)$ is an estimate of $E[\mathbf{Y}_i | A_i = a, \mathbf{X}_i]$. To estimate $var(\hat{\beta})$, the standard estimating function is replaced with the augmented estimating function ψ_a in the middle term of (5).

Method IIIb: Approximation to the Exact Test (Multivariate)

Although responses y_{ij} and covariates $\mathbf{x}_{ij}$ are considered fixed for randomization inference, the calculated covariance among y_{ij} in the ith cluster incorporates information on the difference in the between versus within sum of squares, which may increase power in testing. A working covariance $\mathbf{V}_i$ as for GEE is incorporated into the test statistic given by

$$S_D = \sum_{i=1}^{n} (A_i - \pi)\mathbf{1}\mathbf{V}_i^{-1}\mathbf{w}_i, \tag{6}$$

where $\mathbf{w}_i$ is the residual vector $\mathbf{w}_i = (w_{i1}, w_{i2}, ..., w_{im_i})^{\mathrm{T}}$ determined by $w_{ij} = \hat{\varepsilon}_{ij} = y_{ij} - d(\mathbf{x}_{ij}; \hat{\eta})$ and $\mathbf{1}$ is the m_i-dimensional vector of 1s. To estimate correlation parameters, the method of moments is used. We consider the moment estimating equations

$$\sum_{i=1}^{n} \sum_{j<j'} \left\{ \frac{w_{ij}w_{ij'}}{\tau} - r(\gamma) \right\},$$

where $\tau = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}^2$. The weight matrix $\mathbf{V}_i$ is given by $V_i = L^{1/2}UL^{1/2}$, where L is an $m_i \times m_i$ diagonal matrix with τ along the diagonal, and $\mathbf{U}$ is a correlation matrix, where $Q_{j,j'} = r(\gamma)$. For vector-valued outcomes $\mathbf{Y}_i$, the variance is

$$Var(S|\mathbf{y_i}, \mathbf{x_i}) = \pi(1-\pi) \sum_{i=1}^{n} (\mathbf{1}\mathbf{V}_i^{-1}\mathbf{w}_i)^{\otimes 2} + \left(\pi \frac{n/2 - 1}{n-1} - \pi^2 \right) \overbrace{\sum_{i \neq i'} (\mathbf{1}\mathbf{V}_i^{-1}\mathbf{w}_i)(\mathbf{1}\mathbf{V}_{i'}^{-1}\mathbf{w}_{i'})^{\mathrm{T}}}^{Q*}, \tag{7}$$

where $Q*$ is the small-sample correction for fixed treatment allocation. Bickel and Zwet (1978) may be applied to dependent outcomes as well to ensure nominal type I error levels in small samples.

Method IVb: Exact Test (Multivariate)

The null distribution of test statistic (6) is determined by permuting the cluster-level treatment assignment A_i. Because outcomes and covariates are fixed, the residuals

$\hat{\varepsilon}_{ij} = y_{ij} - d(\mathbf{x}_{ij}; \hat{\eta})$ and working covariance $\mathbf{V}_i$ do not depend on the permuted treatment assignment under H_0. Working covariance parameters therefore only need to be estimated once, and $\mathbf{V}_i$ is equal for all permutations A_b. Testing is conducted as in section 2.1.

2.3 Model Selection for Baseline Covariates

When the set of baseline covariates is high dimensional relative to sample size, adjusting for all available covariates may be inefficient. Prior knowledge may suggest the inclusion of some covariates; among other covariates whose impact on Y_i is not well understood model selection may help to determine which covariates to include. Adjusted mean models and augmented estimation require the conditional mean model $E[Y|\mathbf{X}, A]$, whereas randomization inference requires an estimate of $E[Y|\mathbf{X}]$. Current literature provides a wide array of methods for selection of baseline covariates, particularly for univariate outcomes. Stepwise selection procedures based on some entry criterion may be used. Methods based on penalized likelihoods such as LASSO (Tibshirani (1996)), adaptive LASSO (Zou (2006)), SCAD (Fan and Li (2001)), and MC+ (Zhang (2010)) are all applicable. Model selection for multivariate outcomes is less developed, but extensions of available methods are presented and discussed in Sofer et al. (2012). We consider two popular approaches, forward selection by AIC or BIC, and adaptive LASSO, (Zou (2006)) where the tuning parameter is selected by cross validation.

Forward selection is an example of a greedy algorithm, defined as an algorithm that makes the locally optimal choice at each stage in search of a global optimum (Black (2005)). To find the best predictive model, forward selection starts with a generalized linear model containing the intercept and at each step enters a single covariate according to a prespecified criterion. Examples of entry criteria include minimizing p-values or an information criterion such as AIC, or maximizing adjusted r^2.

Model selection by penalized regression is derived by minimizing an objective function

$$\Omega(\beta) = \sum_{i=1}^{n} L\{Y_i, g(A_i, \mathbf{X}; \beta)\} + P_\lambda(\beta), \tag{8}$$

which consists of a loss function $L\{Y_i, g(A_i, \mathbf{X}; \beta)\}$ and a penalty $P_\lambda(\beta)$, where $P_\lambda(\beta)$ is indexed by a nonnegative tuning parameter λ. The form of $P_\lambda(\beta)$ defines various regularized regression methods; for adaptive LASSO $P_\lambda(\beta) = \lambda \sum_{k=1}^{p} \hat{w}_k |\beta_k|$ with weights $\hat{w}_k = 1/|\hat{\beta}_k^\gamma|$ derived from an initial fit of β. We consider an adaptive LASSO-hybrid implementation motivated by the LASSO-OLS hybrid (Efron et al. (2004)), in which LASSO is used to determine the covariates for which $\beta_k \neq 0$, and the selected model is subsequently fit by OLS.

When outcomes are multivariate Sofer et al. (2012) suggest that accounting for correlation improves the efficiency of penalized regression estimates. In small samples, it is especially desirable to reduce the variability in penalized regression since the number of units may not be sufficient to achieve consistency despite estimation under a misspecified independence correlation structure. The authors recommend scaling outcomes and covariates by $\Lambda^{1/2}$, where $\Lambda = \mathbf{V}_i^{-1}$ is a working precision matrix based on an initial estimate of the coefficient vector. The initial estimate may be determined by a model selection method that assumes independence. For validation-based penalized regression, estimation proceeds as in the univariate case on the scaled outcomes $\tilde{\mathbf{Y}}_i = \Lambda^{1/2}\mathbf{Y}_i$ and covariates $\tilde{\mathbf{X}}_i = \Lambda^{1/2}\mathbf{X}_i$. We also consider forward selection of $\tilde{\mathbf{Y}}_i$ on $\tilde{\mathbf{X}}_i$ to evaluate possible improvements in model selection and resulting power for testing treatment effects.

3 Simulation Study

3.1 Univariate

We first consider scalar outcomes Y_i. For each simulated dataset 25 baseline covariates $X_{i_1}, ..., X_{i_{25}}$ were generated from the multivariate lognormal distribution by exponentiating draws from the multivariate normal distribution with mean $\mu = (0, 0, ..., 0)$ and covariance Σ, where Σ was defined such that $corr(\log(X_{i_k}),\log(X_{i_{k'}})) = 0.5$ for $k, k' = 1, ..., 10$, $corr(\log(X_{i_k}),\log(X_{i_{k'}})) = 0.2$ for $k = 1, ..., 10, k' = 11, ..., 20$, $corr(\log(X_{i_k}),\log(X_{i_{k'}}) = 0$ for $k = 1, ..., 20, k' = 2, ..., 25$, and $var(log(X_{i_k})) = 1$ for $k = 1, ..., 25$. Treatment A_i was binary and simulated with a fixed, equal number of subjects assigned to treatment or control. Outcomes were generated from the model $Y_i = \eta_0 + \eta_1 A_i + \eta_2 X_{i_1} + \eta_3 X_{i_2} + \eta_4 X_{i_{10}} + \eta_5 X_{i_{11}} \eta_6 X_{i_{12}} + \varepsilon_i$ with $\log(\varepsilon_i) \sim N(0, 1.9)$, $\eta' = (1, 0, 1, 1, 0.2, 0.2, 0.2)$ under the null and $\eta' = (1, 4, 1, 1, 0.2, 0.2, 0.2)$ under the alternative. Sample sizes of $n_a = 10, 15, 25, 50, 100$ in each treatment arm were considered. Under this design, baseline covariates accounted for roughly 30% of the variability in $Y_i|A_i$.

All four covariate-adjusted methods were applied to each dataset, and various adaptive procedures were used to select among the 25 baseline covariates. Several variations for each covariate-adjusted test were considered, with each variation defined by a different regression model. For adaptive approaches, selection of regression models was based on three different methods: forward selection minimizing AIC, forward selection minimizing BIC, and the adaptive LASSO-OLS hybrid. The adaptive LASSO tuning parameter was selected by l-fold cross validation, where $l = n/10$. For Method Ia, inference was performed by OLS on the model including A_i and covariates suggested by the adaptive model selection procedure. Adaptively selected models were compared to two fixed models: the data generating model, which serves as a benchmark for the largest possible improvement in power, and an incorrect model, $E[Y_i|\mathbf{X}_i, A_i] = \eta_0 + \eta_1 X_{i_1} + \eta_2 X_{i_3} + \eta_3 X_{10} + \eta_4 X_{i_{13}} + \eta_5 X_{i_{21}}$, including two predictive covariates and 3 noisy covariates. Finally, each method was also

applied to the unadjusted outcomes Y_i to assess whether incorporating baseline covariates improved power compared to no adjustment. Treatment was forced into the regression model for Methods Ia and IIa. In investigation of Methods IIIa and IVa, treatment was omitted from covariate selection, as the sharp null excludes any estimated effect of treatment, even if not significant. In addition to assessing type I error and power when the true data-generating model was contained in the set of candidate models, we also assessed power when important transformations for baseline covariates were not included. We modified the data generating mechanism to include squared terms for X_{i_1} and $X_{i_{10}}$ and changed the coefficient of X_{i_1} to $\eta_1 = 0.50$. As in the previous setting, model fitting algorithms for determining predictive covariates only considered linear terms.

Results for type I error are shown below in Figures 1a-1b and Table 1. Method Ia performed poorly for small sample sizes with model selection, leading to type I error rates as large as α=0.2. For fixed models chosen apriori, testing β_1^* preserves type I error, and is even slightly conservative as a result of the skewness in the covariates and outcomes (α=0.0311-0.043). The performance of asymptotically equivalent Method IIa varies over the choice of model selection procedure. For adaptive LASSO, the augmented test resulted in type I errors roughly twice the nominal level at $n_a = 10$. Adaptive selection of covariates by AIC or BIC had even larger type I error inflation (α=0.40 for n_a=10). Type I error was still not preserved when augmenting with fixed models (0.12 for n_a=10). By contrast, Methods IIIa and IVa maintained type I error at all sample sizes considered. The approximate exact test remained slightly conservative due to skewness in the data, while the exact test achieved nominal type I error levels. There are noteworthy differences in the behavior of the various model selection procedures. As expected, BIC favored more parsimonious models than AIC: AIC-based selection resulted in models with 5 to 7 baseline covariates on average; BIC, with 3 to 4 covariates. Adaptive LASSO was the most conservative model selection procedure, and included 1 to 4 covariates on average, with the number of covariates selected increasing with the sample size.

Table 2 provides simulation results demonstrating the impact of model selection procedures on power. For $n_a \leq 50$, covariate adjustment based on AIC and BIC resulted in larger power than did the correct covariate adjustment model for Methods Ia and IIa (Power=0.68-0.91 for AIC and BIC, Power=0.58-0.90 for the correct model), suggesting that the former led to overfitting of the regression. The power of adjustment with adaptive LASSO did not exceed the power of adjustment under the correct model for any covariate-adjusted test statistic considered. In general, Methods IIIa and IVa had lower power than Methods Ia and IIa, reflecting the fact that the randomization-based tests preserve type I error whereas adding covariates to the mean model and augmentation tests do not. For very small sample sizes ($n_a \leq 15$), covariate adjustment by AIC resulted in lower power than the unadjusted test (Approx. Exact AIC = 0.36-0.46 , Approx. Exact Unadjusted 0.41-0.52 ; Exact AIC = 0.49-0.57, Exact Unadjusted = 0.59-0.64). For $n_a \geq 25$, AIC-based adjustment improved power compared to no adjustment. Model selection by BIC and adaptive LASSO, which penalize more severely for model complexity than AIC, improved power over unadjusted test statistics across all simulated sample sizes. Method IVa had higher power than Method IIIa, with the difference in power increasing inversely with sample size. Across all settings considered, Bickel's adjustment for the distribution of the approximate exact test had little impact on resulting inferences, suggesting that even higher order terms may be necessary to recover nominal type I error.

In the second set of power simulations, the data-generating model contained quadratic terms that were not considered in covariate adjustment. Results are shown in Figures 3a-3b and Table 3. The relative performance of adaptive procedures remained the same. At small samples sizes, exact inference AIC resulted in less power improvement than the other adjustment methods. At $n_a = 10$, exact inference based on the AIC-selected model mirrored unadjusted exact inference (Method IVa AIC = 0.27, Method IVa Unadjusted=0.25). Considering Method IIIA, AIC-based inference increased power relative to not adjusting, but gains were limited compared to BIC selection, adaptive LASSO, and

the prespecified incorrect model (AIC =0.245, Unadjusted= 0.18, BIC=0.3166, adaptive LASSO=0.3541, Prespecified=0.3044). Increasing the sample size per arm to $n_a = 25$, power for AIC-selected adjustment was more similar to the BIC and adaptive LASSO. At $n_a \geq 50$, all adaptive procedures resulted in similar power, while the incorrect prespecified model had lower power (Prespecified=0.49-0.75, Adaptive Methods = 0.54-0.84).

3.2 Multivariate

To evaluate clustered outcome data, values for covariates $X_{ij_1}, ..., X_{ij_{25}}$ were generated, with $X_{ij_k} = X_{i_k}$ for $k = 1, ..., 10$. For each cluster, $(\log(X_{i_1}), ..., \log(X_{i_{10}})) \sim MVN(\mathbf{0}, \mathbf{\Sigma_2})$, where Σ_2 was defined such that $corr(\log(X_{i_k}),\log(X_{i_{k'}})) = 0.5$ for $k = 1, ..., 5, k' = 1, ..., 5$ and $k = 6, ..., 10, k' = 6, ..., 10$, $corr(\log(X_{i_k}),\log(X_{i_{k'}})) = 0.2$ for $k = 1, ..., 5, k' = 6, ..., 10$. Each covariate X_{ij_k} for $k = 11, ..., 20$ was simulated from the multivariate lognormal distribution with $corr(log(X_{ij_k}), log(X_{ij'_k}))$=0.2 independently across k. Finally, for $k = 21, ..., 25$, $log(X_{ij_k}) \sim N(0, 25)$ with independence between and within clusters. Binary treatment A_i was generated with $P(A = 1) = 0.5$, with the total number of clusters assigned to each treatment level fixed accordingly. To induce unexplained correlation within clusters, random cluster effects b_i were simulated, with $log(b_i) \sim N(0, \rho\sigma^2)$, where ρ was varied to induce high or low intracluster correlation. Outcomes Y_{ij} were generated from the model $Y_{ij} = \eta_0 + \eta_1 A_i + \eta_2 X_{i_1} + \eta_3 X_{ij_{11}} + \eta_4 X_{i_3} + \eta_5 X_{ij_{12}} \eta_6 X_{ij_{15}} + b_i + \varepsilon_{ij}$, with $\log(\varepsilon_{ij}) \sim N(0, \sigma^2 = 1.9)$. We set the coefficient vector $\eta = (1, 0, 1.25, 1.25, 0.2, 0.2, 0.2)$ under the null hypothesis of no treatment effect, and $\eta = (1, 2.2, 1.25, 1.25, 0.2, 0.2, 0.2)$ under the alternative. Monte Carlo datasets consisted of $n = 10, 15, 25$ clusters of size $m_i = 20, 30$ or $n = 25, 50, 100$ clusters of size $m_i = 4, 6, 8$ per treatment arm. Values of ρ considered were $\rho = 7/19$ under the null, and $\rho = 7/19, 1$ under the alternative, corresponding to $corr(Y_{ij}, Y_{ij'}|\mathbf{X}_i, A_i)$=5% and 50%, respectively. At $\rho = 7/19$, the correlation between Y_{ij} and baseline covariates was 0.28, whereas $\rho = 1$ reduced $corr(Y_{ij}, \mathbf{X}_{ij}|A_i)$ to 0.17.

We first adaptively determined predictive models for the mean outcome conditional on baseline covariates without consideration of correlation among outcomes within a cluster. We then compared these results to the Monte Carlo power of adjusted tests when model selection did account for correlation in responses (Section 2.3). Selection of baseline covariates for adjustment included forward selection by AIC, two modifications of BIC for multivariate data, and adaptive LASSO. All regression models were ultimately fit by OLS. For BIC, two regression models were selected, the first considering the number of clusters in the penalty for model complexity(BICn), and the second calculating BIC based on the total number of individual-level observations(BICm). In deriving BIC for mixed models, Pauler (1998) showed that for a random intercept model the true penalty is of the form $\Omega_h = \sum_{k=1}^{p} log(N_k^*)$, where h indexes candidate models, k indexes the p covariates in the hth model, $N_k^* = n$ for between-cluster effects, and $N_k^* = M$ for within-cluster effects. BICm and BICn therefore correspond to models containing only cluster-level covariates or individual-level covariates, respectively. Evaluating the true BIC for models including both types of covariates requires calculating Ω_h for each candidate model in the stepwise procedure by observing its cluster-level and individual-level covariates. To ease computation, BICm and BICn were used. The adaptive LASSO tuning parameter was selected based on five-fold cross validation. The two fixed regression models included the data generating model and an incorrect model, $E[Y_{ij}|\mathbf{X}_{ij}, A_i] = \eta_0 + \eta_1 X_{i_1} + \eta_2 X_{i_2} + \eta_3 X_{i_{10}} + \eta_4 X_{ij_{13}} + \eta_5 X_{ij_{21}}$, including two predictive covariates and 3 noisy covariates. For Methods Ib and IIb, treatment was forced into the regression model; model selection and prespecified models for the randomization tests omitted treatment. The null distribution of the observed test statistic under the exact test was determined by permuting the treatment assignment across clusters $b = 1000$ times. Unadjusted tests were also performed for each method and compared to covariate-adjusted tests. The impact of incorporating the covariance structure on randomization tests was evaluated by conducting each test under both independence and exchangeable

correlation structures for each adjustment model. Specification of a covariance structure for standard GEE and augmented GEE methods have been evaluated elsewhere (Wang and Carey (2003),Stephens et al. (2012a)).

Type I error for each method is presented in Tables 5-7. In small samples ($n_a \leq 25$) GEE methods fail to control type I error for all covariate-adjusted analyses. Inflation of type I error reflects bias in variance estimation of the sandwich estimator in small samples as well as additional variance induced by model selection. Under model selection, type I error was as large as $\alpha = 0.24$ for Method Ib and $\alpha = 0.31$ for Methods IIb. When the number of clusters was large ($n_a \geq 50$), nominal type I error levels of $\alpha = 0.05$ were achieved when covariates were not selected adaptively. Type I error was still inflated under model selection for the large n considered ($\alpha = 0.05 - 0.068$ for $n_a \leq 25$), but inflation was slight compared to that observed for small n (α=0.07-0.31). For testing treatment effects, model selection by AIC resulted in the largest type I error, followed by the BIC methods; the adaptive LASSO had the least type I error inflation. For the randomization tests, the approximate exact test was generally conservative across all outcomes. The Bickel adjustment for defining the rejection region increased type I error levels of the approximate exact test closer to the nominal level. The exact test had nominal type I error across selected and prespecified covariate-adjusted models.

Plots 4a-6b and Tables 8-13 compare power across covariate-adjusted tests for dependent outcomes. In most cases, covariate adjustment improved power compared to the corresponding unadjusted approaches, regardless of the method of model selection used. Precision matrix scaling seemed to reduce overfitting in model selection; adaptive methods tended to select fewer covariates when outcomes and covariates were scaled prior to adjustment in the setting where outcomes were highly correlated (Table 14). Post-selection randomization tests also had larger power when outcomes and covariates were scaled before selection versus not scaled.

Method IVb at $n_a = 10$ AIC and BICn selection strategies had lower power than

did strategies that did not adjust for baseline covariates when the exchangeable working covariance was used and precision matrix scaling was not done prior to model selection (Unadjusted 0.2170, AIC 0.1894, BICn 0.2014). Upon scaling outcomes and covariates prior to model selection, post-selection by AIC or BICn tests were more powerful than unadjusted tests (AIC 0.229, BICn 0.250). Of the adaptive methods considered, forward selection by BICm resulted in the largest power for both levels of intracluster correlation. Exchangeable working covariance specification improved power over working independence only for randomization tests of the unadjusted outcomes $\mathbf{y}_i$.

4 Application

Covariate-adjusted tests were applied to data from the *Young Citizens* study. *Young Citizens* was a cluster-randomized intervention trial designed to evaluate the effectiveness of a behavioral intervention in training adolescents to be peer educators about HIV. Thirty communities were randomized to intervention or control, resulting in 15 communities per arm. Residents in participating communities were surveyed regarding the degree to which they believed adolescents could effectively communicate to their families and peers about HIV transmission dynamics. The outcome Y_{ij} was a child empowerment score from responses given by individuals within each randomized community. Additional covariates characterizing the communities and households of survey respondents were measured.

Predictive models for baseline covariates were first determined by AIC, BICn, BICm, and adaptive LASSO. Covariates selected by AIC include employment status (employment), age of the head of household (age), whether or not the household had a flushing toilet (flushing toilet), number of relatives in the neighborhood (relatives), religion, community population density (density), transportation ownership (transportation), home ownership (home), and interactions of treatment with relatives and density. BICn selected the same covariates as AIC except for transportation and home, which it did not

enter into the model. BIC penalized by the number of total observations (BICm) chose employment, age, and flushing toilet. Finally, adaptive LASSO picked flushing toilet, religion, employment, age, and interactions with treatment and density, relatives, and number of kids in the house. For randomization tests, the AIC-based model contained employment, flushing toilet, age, religion, relatives, home, and wealth deviance for each family from the mean community wealth. BICn selected employment, flushing toilet, age, religion and relatives. Selection by BICm and adaptive LASSO chose employment, flushing toilet, and age.

Table 15 presents results from the *Young Citizens* analysis. Adjusted and augmented GEE methods were associated with highly significant treatment effects ($p < 0.0001$) across covariate-adjusted and unadjusted tests. For the approximate exact tests, all covariate-adjusted methods yielded a significant intervention effect. When unadjusted, however, only the test using exchangeable covariance resulted in significantly different child empowerment between intervention groups ($p = 0.0233$ for exchangeable working covariance, $p = 0.10$ under independence). Applying Bickel's small-sample adjustment to obtain tail probabilities resulted in p-values that were slightly larger than those based on the standard normal distribution. Among permutation tests, significant intervention effects were detected under covariate-adjustment, but not in the absence of such adjustment for either working covariance structure. The data provide sufficient evidence that children who participated in the intervention were significantly more equipped to educate their peers about HIV. The results underscore the importance of using appropriate methodology and utilizing baseline covariate information. Unadjusted tests based on GEE methods were highly significant, but as shown in the simulation studies of Section 3, the validity of such methods is not guaranteed with a fairly small number of clusters. Randomization tests, with guaranteed validity in small samples, showed similar results with covariate adjustment, but conclusions of unadjusted tests were inconsistent.

5 Discussion

We have investigated the dangers and merits of several procedures that allow for flexible covariate adjustment when applied to small samples. Simulation studies showed, as expected, that AMM and augmented methods break down in small samples when the number of baseline covariates is large relative to the sample size. Alternatively, randomization methods, which exploit the fact that outcomes and baseline covariates are regarded as fixed, provide valid tests for treatment effects while flexibly incorporating baseline covariates. Model selection may be used to identify the set of baseline covariates that explain the greatest amount of variability in the outcome while preserving the type I error of the primary test. The central conclusion is that for randomization tests, adjustment models need not be prespecified to preserve the nominal type I error. Furthermore, adjustment generally increases the power of testing for treatment effects over unadjusted methods, with the caveat that in extremely small samples of independent outcomes, such as $n_a = 10, 15$, model selection approaches must be sufficiently conservative. Model selection by BIC and adaptive LASSO, which have stronger penalties and therefore favor more parsimonious models than AIC, resulted in improved power at the smallest sample sizes considered. Further research is needed to formally characterize the power of covariate-adjusted tests under misspecified covariate adjustment and adaptive covariate selection.

Our work has focused on hypothesis testing for evaluating treatment effects. For confidence interval estimation, hypothesis tests may be inverted. When inverting randomization-based hypothesis tests, it is important to note that for each potential value of the treatment effect considered, model selection needs to be repeated, since conditional mean models are estimated by pooling across treated and untreated subjects. Interval estimation may be simplified by a slight modification of the testing procedure. Under the sharp null, the conditional mean model may be estimated using data only for untreated subjects. The model may then be applied to all subjects in conducting the test. Not

pooling the data when estimating the conditional mean model removes the need for its re-estimation with each treatment effect value considered. For small-sample univariate data, it may not be feasible to perform model selection on one treatment group, but for a small number of moderately sized clusters such a strategy may be more reasonable.

References

Bickel, P. J. and Zwet, W. R. V. (1978). Asymptotic expansions for the power of distribution-free tests in the two-sample problem. *The Annals of Statistics* **6** 937–1004.

Black, P. (2005). Greedy algorithm. In *Dictionary of Algorithms and Data Structures* (P. Black, ed.).
URL `http://xlinux.nist.gov/dads//HTML/greedyalgo.html`

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32** 407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconvace penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

Gail, M. H., Tan, W. Y. and Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika* **75** 57–64.

Klar, N. and Donner, A. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Hodder Arnold.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.

Pauler, D. K. (1998). The schwarz criterion and related methods for normal linear models. *Biometrika* **85** 13–27.

Rosenbaum, P. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17** 286–327.

Sofer, T., Dicker, L. and Lin, X. (2012). Variable selection for high-dimensional multivariate outcomes. In preparation.

Stephens, A. J., Tchetgen Tchetgen, E. J. and De Gruttola, V. (2012a). Augmented gee for improving efficiency of inferences in cluster randomized trials by leveraging cluster and individual-level covariates. *Statistics in Medicine* In press.

Stephens, A. J., Tchetgen Tchetgen, E. J. and De Gruttola, V. (2012b). Locally efficient estimation of marginal treatment effects using auxiliary covariates in randomized trials with correlated outcomes. In preparation.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58** 267–288.

Tsiatis, A. A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons for randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27** 4658–4677.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. NY: Springer-Verlag.

Wang, Y. and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika* **90** 29–41.

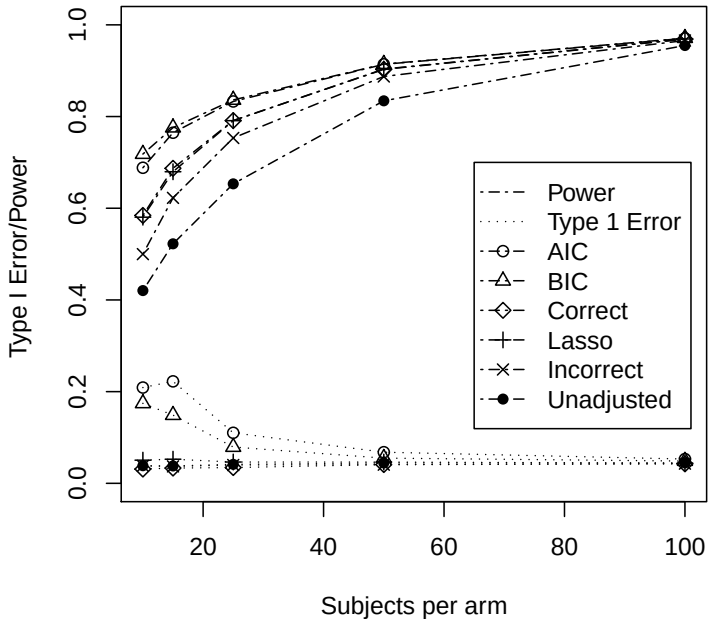
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Journal of the American Statistical Association* **101** 1418–1429.

Zhang, M., Tsiatis, A. A. and Davidian, M. (2008). Improving efficiency of inferences in clinical randomized trials using auxiliary covariates. *Biometrics* **64** 707–715.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.

Figure 1: **Type I Error and Power of Univariate AMM and Augmented Tests.** Adaptive regression model selection: AIC, BIC, Adaptive LASSO. Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

(a) AMM



(b) Augmented

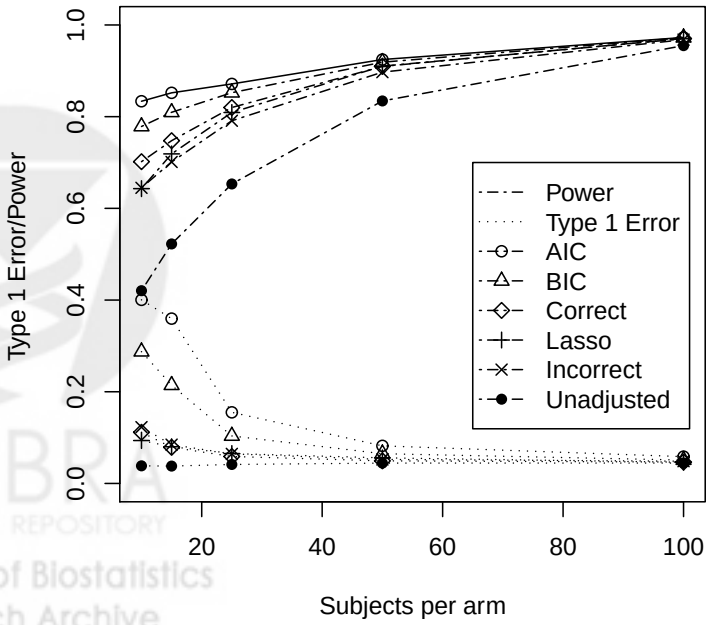
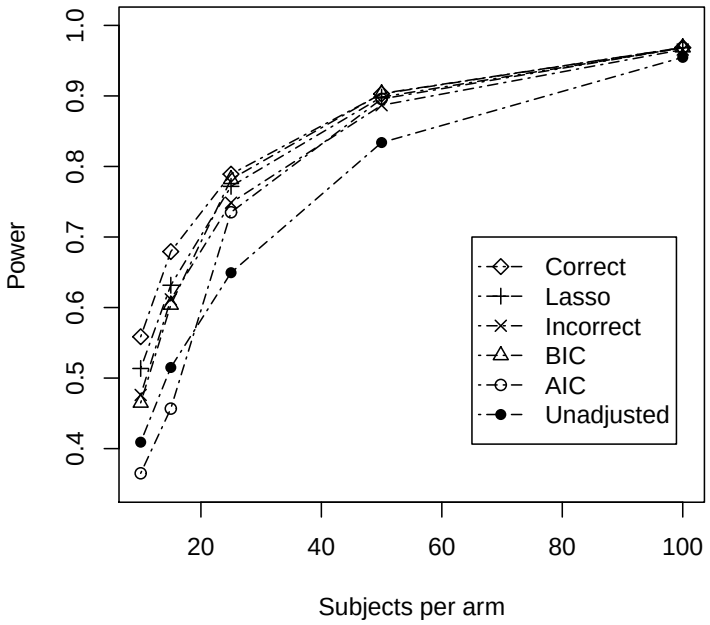


Figure 2: **Power of Univariate Approx. Exact and Exact Tests when the correct model is a candidate model.** Adaptive regression model selection: AIC, BIC, Adaptive LASSO. Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

(a) Approximate



(b) Exact

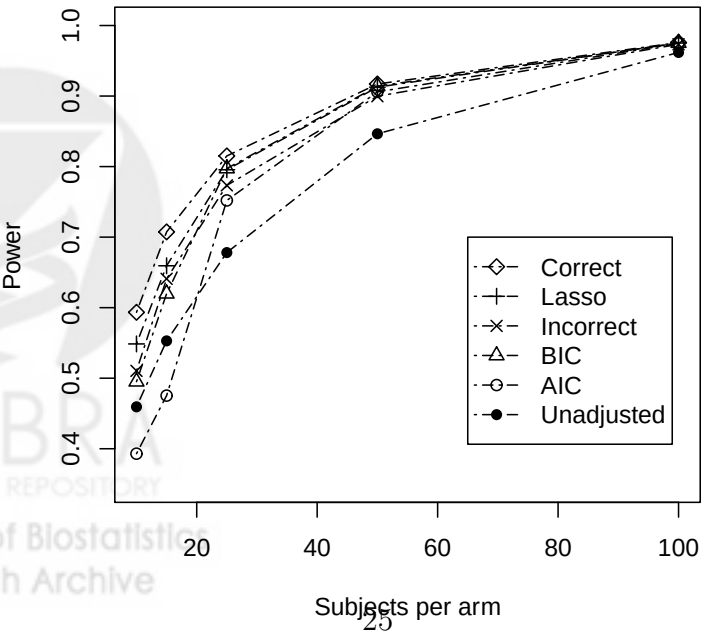
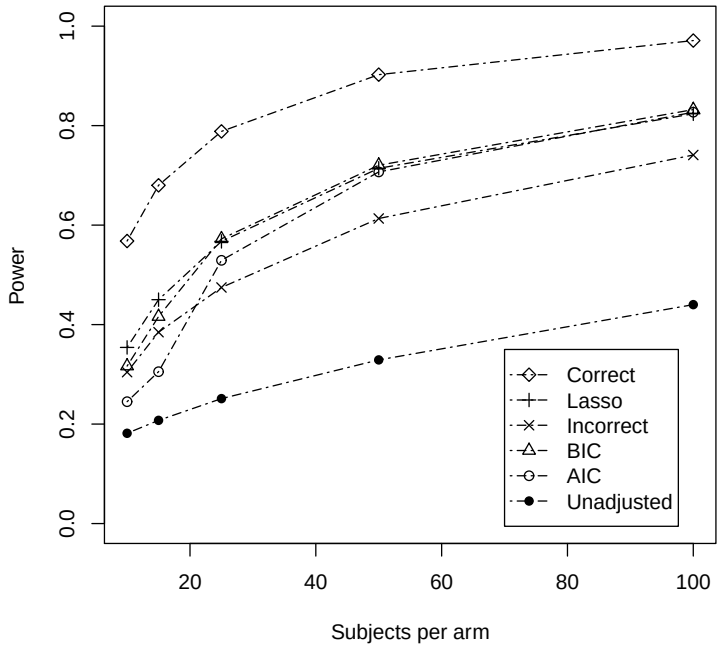


Figure 3: **Power of Univariate Approx. Exact and Exact Tests when the correct model is not a candidate model.** Adaptive model selection: AIC, BIC, Adaptive LASSO. Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

(a) Approximate



(b) Exact

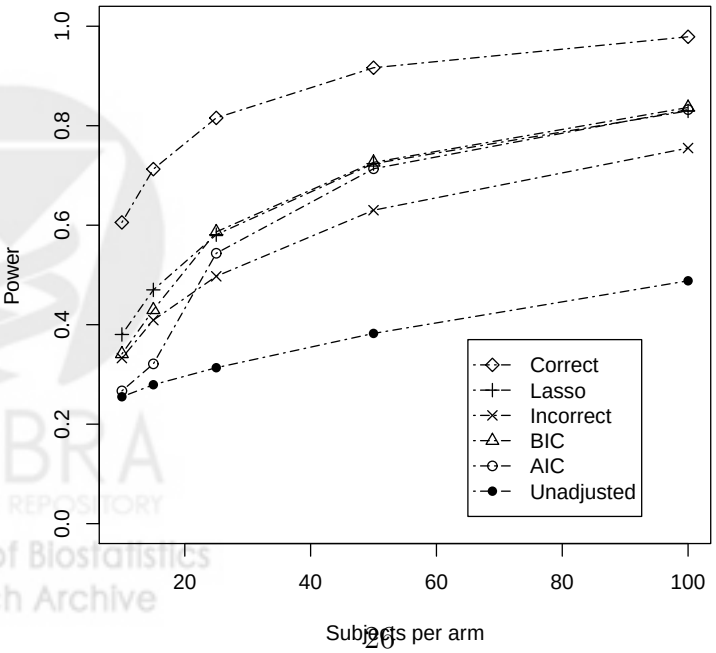
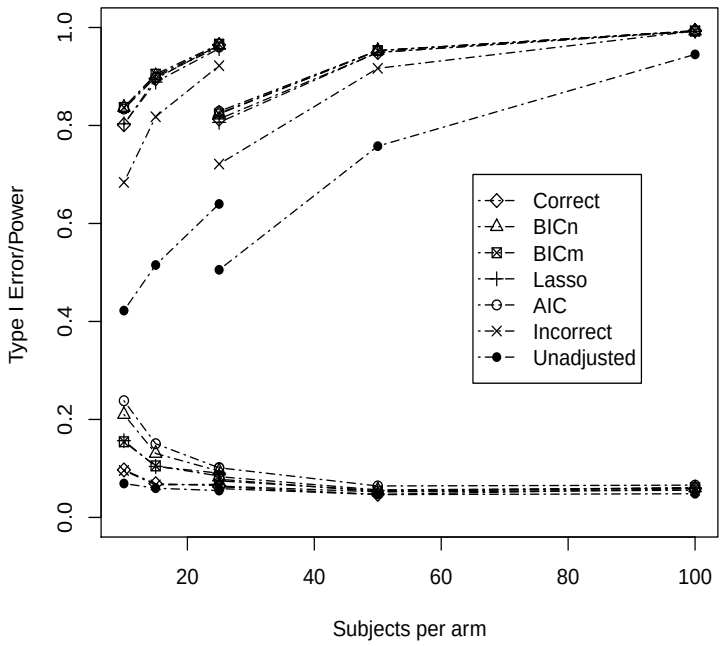


Figure 4: **Type I Error and Power of Multivariate AMM and Augmented Tests.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (Lasso). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

(a) AMM



(b) Augmented

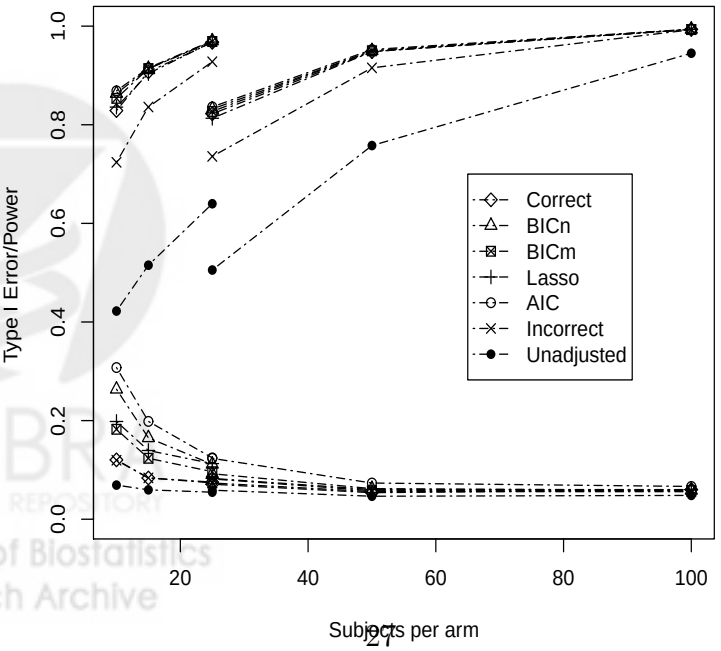
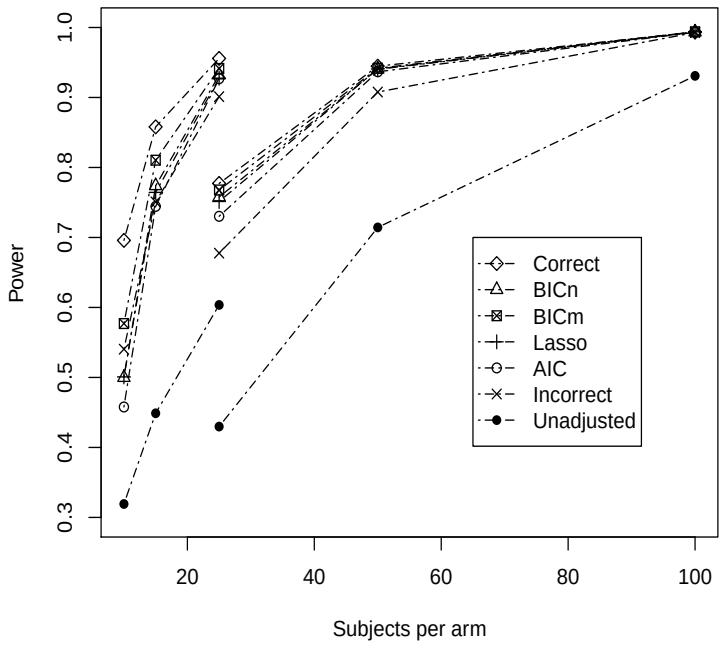


Figure 5: **Power of Multivariate Approx. Exact and Exact Tests: low correlation.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (Lasso). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

(a) Approximate



(b) Exact

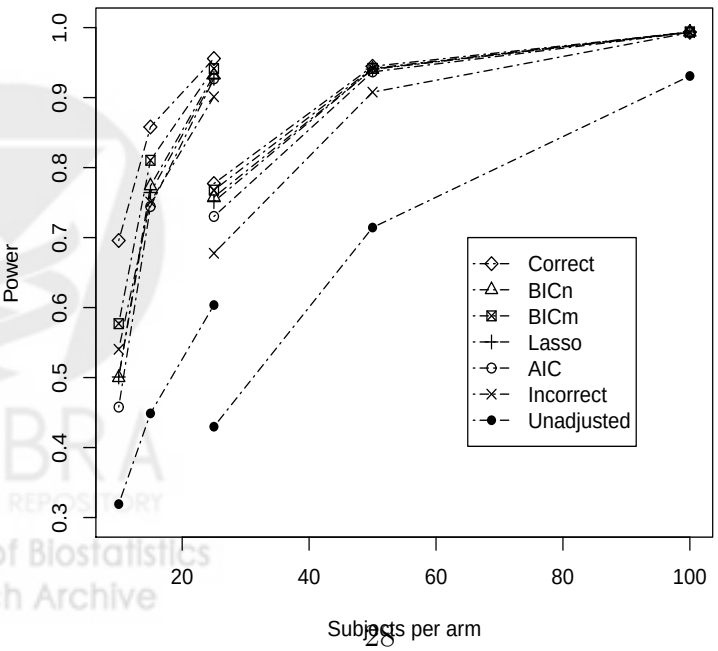
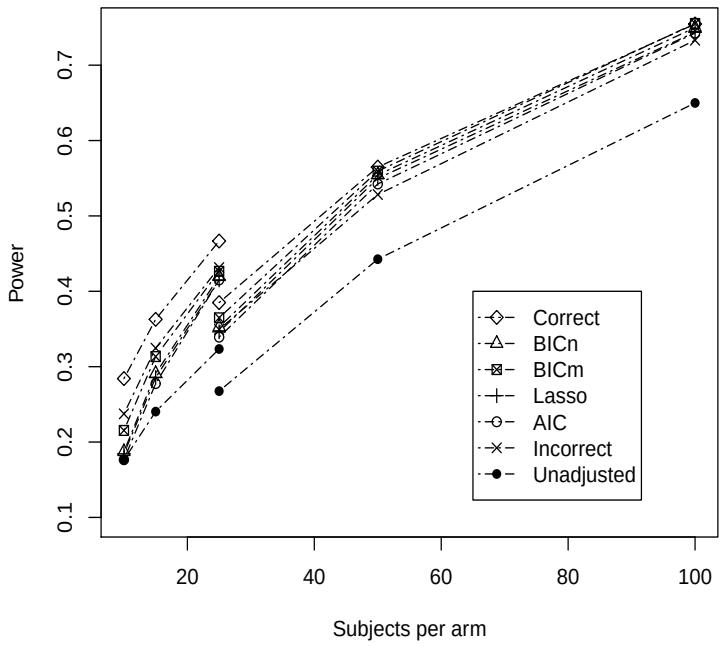


Figure 6: **Power of Multivariate Approx. Exact and Exact Tests: high correlation.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (Lasso). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

(a) Approximate



(b) Exact

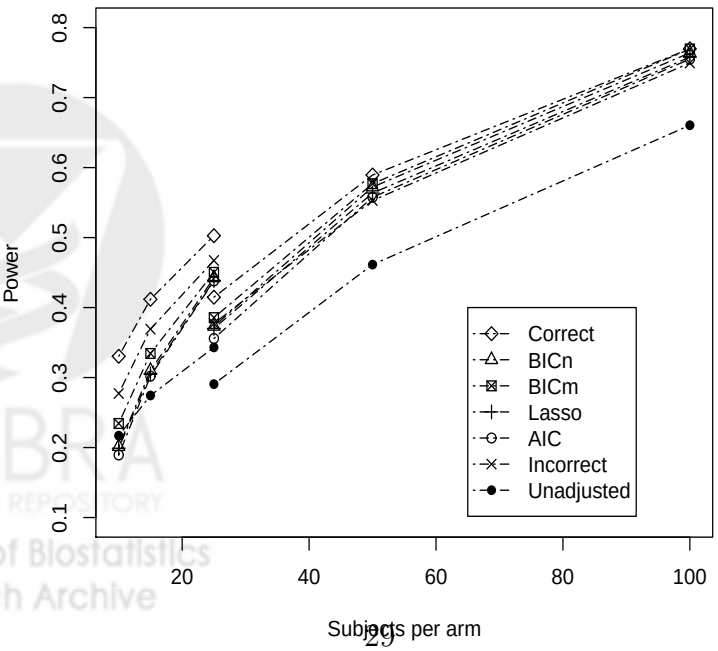


Table 1: **Type I Error of Univariate Covariate-adjusted Tests.** Adjusted mean model (AMM), Augmented, Approx. Exact (without Bickel adjustment), Approx. Exact (Sm) (with Bickel adjustment) and Exact tests. Adaptive regression model selection: AIC, BIC, Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.0384	0.2089	0.1744	0.0505	0.0381	0.0311
15	0.0379	0.2224	0.1488	0.0526	0.037	0.0333
25	0.0414	0.1102	0.0792	0.0465	0.04	0.0344
50	0.0444	0.0679	0.055	0.0464	0.0407	0.0409
100	0.0445	0.053	0.0486	0.044	0.043	0.0425

Augmented

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.0384	0.4005	0.2874	0.0936	0.1228	0.1116
15	0.0379	0.3595	0.2143	0.0801	0.0846	0.0788
25	0.0414	0.1551	0.1036	0.0652	0.0645	0.0588
50	0.0444	0.082	0.0649	0.0559	0.0524	0.0493
100	0.0445	0.0585	0.051	0.0462	0.0486	0.0466

Approx. Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.0346	0.0368	0.0383	0.0356	0.0368	0.0356
15	0.0354	0.0446	0.0406	0.0375	0.0347	0.0375
25	0.0398	0.0375	0.0388	0.039	0.0389	0.039
50	0.0438	0.0415	0.0423	0.0417	0.0398	0.0417
100	0.0442	0.0421	0.0438	0.0418	0.043	0.0418

Approx. Exact (Sm)

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.033	0.035	0.036	0.034	0.0347	0.034
15	0.0354	0.0442	0.0396	0.037	0.0345	0.037
25	0.0412	0.0384	0.0398	0.0403	0.0394	0.0403
50	0.0456	0.0433	0.0443	0.0432	0.0424	0.0432
100	0.0454	0.0432	0.0453	0.0433	0.0442	0.0433

Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.0498	0.0487	0.0491	0.0489	0.0519	0.0486
15	0.0499	0.0543	0.0511	0.0491	0.0481	0.0495
25	0.0518	0.0456	0.0491	0.0492	0.0509	0.0494
50	0.0515	0.0517	0.0529	0.0541	0.0524	0.0546
100	0.0505	0.0483	0.0524	0.0489	0.0513	0.0504

Table 2: **Power of Univariate Covariate-adjusted Tests when the correct model is a candidate model.** Adjusted mean model (AMM), Augmented, Approx. Exact (without Bickel adjustment), Approx. Exact (Sm) (with Bickel adjustment) and Exact tests. Adaptive regression model selection: AIC, BIC, Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4204	0.6883	0.7182	0.5805	0.4999	0.5843
15	0.5224	0.7647	0.7758	0.6796	0.6226	0.6871
25	0.6532	0.8329	0.8362	0.791	0.7532	0.7912
50	0.8343	0.9139	0.9144	0.9035	0.8874	0.9029
100	0.9549	0.9706	0.971	0.9692	0.9658	0.9687

Augmented

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4204	0.7991	0.7786	0.643	0.6448	0.7018
15	0.5224	0.8244	0.8095	0.7188	0.7012	0.7476
25	0.6532	0.8573	0.8523	0.8091	0.7911	0.82
50	0.8343	0.9206	0.9188	0.9102	0.8971	0.9096
100	0.9549	0.9722	0.9722	0.97	0.9679	0.9705

Approx. Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4091	0.365	0.4649	0.5136	0.4761	0.5586
15	0.515	0.4567	0.6038	0.6316	0.6116	0.6793
25	0.6494	0.7351	0.7819	0.7718	0.7486	0.7891
50	0.8339	0.8957	0.9034	0.8983	0.8868	0.9029
100	0.9547	0.9683	0.9686	0.9682	0.9657	0.9686

Approx. Exact (Sm)

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4051	0.3567	0.4552	0.5056	0.4681	0.5549
15	0.5139	0.4528	0.5996	0.6297	0.6107	0.6807
25	0.6516	0.7366	0.7831	0.7741	0.7515	0.7922
50	0.8358	0.898	0.9055	0.9014	0.8901	0.9054
100	0.9562	0.9695	0.971	0.9696	0.9676	0.97

Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4594	0.393	0.4951	0.5486	0.5104	0.5934
15	0.5529	0.4753	0.6198	0.6594	0.6409	0.7074
25	0.6781	0.752	0.7973	0.7955	0.7734	0.8151
50	0.8465	0.9059	0.914	0.9126	0.8998	0.9171
100	0.9618	0.9747	0.9752	0.9752	0.9734	0.9759

Table 3: **Power of Univariate Covariate-adjusted Tests when the correct model is not a candidate model.** Adjusted mean model (AMM), Augmented, Approx. Exact (without Bickel adjustment), Approx. Exact (Sm) (with Bickel adjustment) and Exact tests. Adaptive regression model selection: AIC, BIC, Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.1880	0.5805	0.6094	0.4500	0.5883	0.3231
15	0.2132	0.6545	0.6557	0.5359	0.6894	0.3947
25	0.2544	0.6809	0.6692	0.6150	0.7919	0.4793
50	0.3305	0.7613	0.7554	0.7343	0.9030	0.6154
100	0.4413	0.8417	0.8412	0.8295	0.9714	0.7419

Augmented

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.1880	0.7467	0.7057	0.5427	0.6054	0.4729
15	0.2132	0.7556	0.7143	0.6067	0.6397	0.4889
25	0.2544	0.7297	0.7078	0.6588	0.7134	0.5329
50	0.3305	0.7820	0.7701	0.7508	0.8196	0.6386
100	0.4413	0.8480	0.8476	0.8367	0.9071	0.7512

Approx. Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.1815	0.2450	0.3166	0.3541	0.5680	0.3044
15	0.2075	0.3053	0.4161	0.4501	0.6800	0.3847
25	0.2512	0.5292	0.5724	0.5673	0.7884	0.4746
50	0.3290	0.7069	0.7204	0.7142	0.9025	0.6133
100	0.4401	0.8269	0.8322	0.8238	0.9710	0.7409

Approx. Exact (Sm)

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.1792	0.2380	0.3101	0.3452	0.5629	0.2995
15	0.2088	0.3020	0.4114	0.4479	0.6820	0.3829
25	0.2569	0.5284	0.5719	0.5676	0.7915	0.4760
50	0.3360	0.7075	0.7219	0.7153	0.9059	0.6166
100	0.4499	0.8289	0.8328	0.8250	0.9726	0.7442

Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.2669	0.3412	0.3803	0.6056	0.3329	0.2551
15	0.3212	0.4298	0.4700	0.7127	0.4092	0.2793
25	0.5436	0.5866	0.5810	0.8157	0.4973	0.3135
50	0.7135	0.7263	0.7238	0.9165	0.6301	0.3824
100	0.8324	0.8367	0.8299	0.9788	0.7551	0.4882

Table 4: **Average Number of Baseline Covariates** selected by AIC, BIC, and Adaptive LASSO by sample size when candidate models include the correct model. First entry - number of baseline covariates selected when treatment was forced into the model. Second entry - number of baseline covariates when treatment was omitted from the model.

n_a	AIC	BIC	A. LASSO
10	6.45	3.93	1.84
	5.75	3.60	1.61
15	8.65	4.14	2.63
	7.96	3.93	2.26
25	6.13	3.19	3.11
	5.95	3.15	2.87
50	5.46	2.94	3.69
	5.41	2.93	3.57
100	5.49	3.01	3.92
	5.48	3.00	3.82

Table 5: **Type I Error of Multivariate AMM and Augmented tests.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO). Pre-specified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

			Adjusted Mean Model					
	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0692	0.2382	0.2100	0.1544	0.1566	0.0970	0.0958
	15	0.0596	0.1504	0.1306	0.1052	0.1040	0.0688	0.0664
	25	0.0548	0.1012	0.0946	0.0846	0.0904	0.0650	0.0676
Small m_i	25	0.0589	0.1014	0.0831	0.0779	0.0747	0.0627	0.0639
	50	0.0466	0.0642	0.0562	0.0526	0.0550	0.0470	0.0522
	100	0.0483	0.0659	0.0601	0.0607	0.0601	0.0586	0.0556
			Augmented					
	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0692	0.3076	0.2636	0.1824	0.1982	0.1204	0.1196
	15	0.0596	0.1984	0.1650	0.1236	0.1394	0.0836	0.0838
	25	0.0548	0.1244	0.1114	0.0964	0.1128	0.0752	0.0738
Small m_i	25	0.0589	0.1234	0.0923	0.0817	0.0827	0.0710	0.0734
	50	0.0466	0.0734	0.0620	0.0578	0.0602	0.0538	0.0560
	100	0.0483	0.0665	0.0586	0.0580	0.0601	0.0601	0.0559

Table 6: **Type I Error of (Multivariate) Approximate Exact Tests.** Results based on Bickel's adjusted cdf are indicated by (Sm). Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Approximate Exact (Ind)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0406	0.0426	0.0384	0.0418	0.0380	0.0400	0.0438
	15	0.0460	0.0378	0.0404	0.0406	0.0392	0.0382	0.0378
	25	0.0430	0.0524	0.0514	0.0500	0.0496	0.0444	0.0484
Small m_i	25	0.0443	0.0469	0.0443	0.0451	0.0471	0.0439	0.0413
	50	0.0432	0.0408	0.0392	0.0404	0.0396	0.0386	0.0434
	100	0.0428	0.0501	0.0516	0.0531	0.0528	0.0531	0.0492

Approximate Exact (Ind-Sm)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0412	0.0436	0.0400	0.0434	0.0392	0.0428	0.0454
	15	0.0478	0.0400	0.0416	0.0432	0.0416	0.0392	0.0390
	25	0.0444	0.0532	0.0522	0.0512	0.0516	0.0468	0.0496
Small m_i	25	0.0453	0.0479	0.0473	0.0475	0.0488	0.0455	0.0429
	50	0.0444	0.0422	0.0406	0.0414	0.0414	0.0400	0.0458
	100	0.0431	0.0519	0.0537	0.0543	0.0549	0.0549	0.0507

Approximate Exact (Exch)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0392	0.0394	0.0384	0.0430	0.0384	0.0402	0.0434
	15	0.0432	0.0396	0.0418	0.0412	0.0402	0.0384	0.0384
	25	0.0430	0.0522	0.0518	0.0510	0.0510	0.0478	0.0480
Small m_i	25	0.0439	0.0463	0.0455	0.0453	0.0477	0.0447	0.0447
	50	0.0406	0.0412	0.0392	0.0404	0.0394	0.0390	0.0458
	100	0.0434	0.0486	0.0525	0.0525	0.0528	0.0534	0.0495

Approximate Exact (Exch-Sm)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0394	0.0418	0.0404	0.0448	0.0402	0.0430	0.0456
	15	0.0446	0.0406	0.0430	0.0428	0.0414	0.0402	0.0390
	25	0.0440	0.0538	0.0530	0.0526	0.0528	0.0486	0.0490
Small m_i	25	0.0451	0.0481	0.0475	0.0475	0.0496	0.0467	0.0461
	50	0.0410	0.0430	0.0408	0.0418	0.0422	0.0410	0.0470
	100	0.0443	0.0504	0.0528	0.0525	0.0534	0.0537	0.0510

Table 7: **Type I Error of Multivariate Exact Tests.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Exact (Ind)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0494	0.0496	0.0454	0.0480	0.0428	0.0490	0.0478
	15	0.0526	0.0450	0.0450	0.0466	0.0434	0.0464	0.0434
	25	0.0474	0.0568	0.0556	0.0528	0.0574	0.0510	0.0510
Small m_i	25	0.0486	0.0512	0.0492	0.0500	0.0524	0.0488	0.0498
	50	0.0466	0.0446	0.0396	0.0408	0.0420	0.0452	0.0474
	100	0.0416	0.0553	0.0543	0.0556	0.0586	0.0562	0.0522

Exact (Exch)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0482	0.0460	0.0454	0.0486	0.0444	0.0494	0.0512
	15	0.0500	0.0470	0.0474	0.0456	0.0456	0.0464	0.0436
	25	0.0484	0.0558	0.0564	0.0560	0.0570	0.0530	0.0502
Small m_i	25	0.0481	0.0520	0.0494	0.0492	0.0522	0.0518	0.0510
	50	0.0444	0.0436	0.0408	0.0416	0.0432	0.0446	0.0476
	100	0.0464	0.0534	0.0556	0.0556	0.0580	0.0565	0.0522

Table 8: **Power of Multivariate AMM and Augmented Tests: low correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. (*) indicates model selection on precision matrix-transformed covariates and outcomes. Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Incorr.
10	0.422	0.834	0.832	0.837	0.832	0.837	0.829	0.803	0.797	0.802	0.684
15	0.515	0.899	0.901	0.901	0.903	0.905	0.905	0.889	0.884	0.895	0.818
25	0.640	0.960	0.962	0.963	0.965	0.966	0.967	0.957	0.954	0.964	0.922
25	0.505	0.829	0.830	0.825	0.826	0.823	0.822	0.806	0.806	0.813	0.721
50	0.758	0.953	0.950	0.953	0.952	0.953	0.952	0.949	0.948	0.949	0.917
100	0.945	0.993	0.994	0.993	0.993	0.993	0.993	0.992	0.993	0.994	0.993

Augmented

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Incorr.
10	0.422	0.869	0.863	0.863	0.858	0.854	0.847	0.836	0.833	0.829	0.724
15	0.515	0.915	0.914	0.914	0.915	0.914	0.912	0.904	0.905	0.907	0.836
25	0.640	0.970	0.969	0.970	0.970	0.969	0.968	0.968	0.965	0.967	0.928
25	0.505	0.836	0.837	0.832	0.832	0.827	0.824	0.813	0.812	0.822	0.736
50	0.758	0.953	0.952	0.950	0.949	0.950	0.950	0.948	0.947	0.948	0.915
100	0.945	0.994	0.994	0.993	0.994	0.993	0.994	0.993	0.994	0.993	0.993

Table 9: **Power of Multivariate Approximate Exact Tests: low correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. (*) indicates model selection on precision matrix-transformed covariates and outcomes. Results based on Bickel's adjusted CDF are indicated by (Sm). Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Approximate Exact (Ind)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Incorr.
10	0.221	0.453	0.482	0.496	0.530	0.566	0.604	0.495	0.494	0.686	0.529
15	0.325	0.740	0.777	0.769	0.806	0.802	0.829	0.759	0.762	0.853	0.738
25	0.465	0.923	0.935	0.930	0.943	0.939	0.948	0.925	0.927	0.952	0.897
25	0.322	0.725	0.735	0.754	0.760	0.763	0.766	0.748	0.748	0.769	0.671
50	0.564	0.933	0.935	0.938	0.939	0.939	0.939	0.935	0.937	0.941	0.905
100	0.827	0.992	0.993	0.992	0.993	0.993	0.993	0.993	0.993	0.993	0.992

Approximate Exact (Ind-Sm)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Incorr.
10	0.226	0.460	0.491	0.503	0.539	0.574	0.612	0.501	0.500	0.692	0.536
15	0.328	0.744	0.780	0.773	0.810	0.807	0.833	0.764	0.768	0.856	0.743
25	0.467	0.925	0.937	0.931	0.944	0.940	0.949	0.926	0.928	0.953	0.901
25	0.326	0.730	0.741	0.759	0.767	0.769	0.772	0.753	0.752	0.776	0.675
50	0.568	0.936	0.938	0.941	0.942	0.943	0.942	0.938	0.940	0.943	0.907
100	0.831	0.993	0.994	0.993	0.994	0.993	0.994	0.994	0.994	0.994	0.992

Approximate Exact (Exch)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Incorr.
10	0.315	0.450	0.484	0.493	0.531	0.568	0.606	0.493	0.495	0.690	0.536
15	0.445	0.739	0.777	0.769	0.809	0.806	0.832	0.760	0.763	0.855	0.748
25	0.602	0.925	0.938	0.930	0.944	0.940	0.950	0.927	0.927	0.955	0.898
25	0.425	0.726	0.733	0.753	0.760	0.762	0.766	0.746	0.747	0.771	0.674
50	0.712	0.935	0.936	0.937	0.939	0.939	0.940	0.937	0.937	0.942	0.906
100	0.930	0.992	0.994	0.993	0.994	0.993	0.994	0.993	0.994	0.993	0.992

Approximate Exact (Exch-Sm)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Incorr.
10	0.319	0.458	0.489	0.500	0.539	0.577	0.613	0.501	0.502	0.696	0.540
15	0.449	0.744	0.781	0.774	0.812	0.810	0.837	0.764	0.768	0.858	0.751
25	0.604	0.927	0.940	0.932	0.946	0.941	0.951	0.928	0.929	0.956	0.901
25	0.430	0.730	0.739	0.758	0.766	0.768	0.772	0.751	0.752	0.777	0.678
50	0.714	0.937	0.938	0.940	0.941	0.942	0.942	0.941	0.940	0.945	0.908
100	0.931	0.993	0.994	0.993	0.994	0.994	0.994	0.994	0.994	0.993	0.993

Table 10: **Power of Multivariate Exact Tests: low correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Exact (Ind)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.246	0.472	0.502	0.512	0.548	0.587	0.622	0.510	0.513	0.705	0.550
15	0.338	0.751	0.785	0.776	0.815	0.811	0.836	0.767	0.770	0.862	0.751
25	0.473	0.927	0.938	0.934	0.945	0.940	0.950	0.929	0.929	0.956	0.902
25	0.335	0.735	0.744	0.763	0.771	0.773	0.776	0.759	0.759	0.785	0.681
50	0.570	0.940	0.940	0.943	0.944	0.944	0.944	0.942	0.943	0.948	0.909
100	0.830	0.994	0.995	0.994	0.995	0.995	0.995	0.995	0.995	0.994	0.992

Exact (Exch)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.347	0.470	0.504	0.512	0.550	0.587	0.622	0.512	0.515	0.709	0.553
15	0.465	0.750	0.785	0.780	0.817	0.813	0.837	0.769	0.772	0.861	0.756
25	0.614	0.929	0.940	0.935	0.948	0.943	0.953	0.931	0.931	0.957	0.902
25	0.443	0.737	0.744	0.761	0.771	0.771	0.774	0.758	0.758	0.784	0.684
50	0.717	0.941	0.941	0.944	0.945	0.946	0.946	0.943	0.943	0.949	0.911
100	0.930	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995	0.994	0.993

Table 11: **Power of Multivariate AMM and Augmented tests: high correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. (*) indicates model selection on precision matrix-transformed covariates and outcomes. Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.252	0.524	0.482	0.527	0.477	0.519	0.467	0.503	0.487	0.409	0.356
15	0.297	0.511	0.486	0.515	0.485	0.514	0.479	0.498	0.491	0.449	0.412
25	0.350	0.544	0.532	0.547	0.531	0.549	0.526	0.537	0.527	0.504	0.477
25	0.308	0.487	0.470	0.477	0.462	0.468	0.455	0.459	0.448	0.431	0.395
50	0.466	0.611	0.606	0.605	0.603	0.604	0.603	0.600	0.599	0.590	0.558
100	0.663	0.768	0.769	0.771	0.766	0.770	0.766	0.769	0.761	0.765	0.742

Augmented

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.252	0.630	0.547	0.527	0.459	0.493	0.434	0.607	0.566	0.449	0.401
15	0.297	0.591	0.523	0.507	0.481	0.491	0.465	0.575	0.544	0.479	0.442
25	0.350	0.583	0.551	0.539	0.533	0.532	0.523	0.578	0.557	0.524	0.488
25	0.308	0.515	0.486	0.470	0.463	0.462	0.455	0.487	0.467	0.453	0.414
50	0.466	0.623	0.608	0.602	0.603	0.602	0.601	0.613	0.605	0.598	0.563
100	0.663	0.771	0.767	0.766	0.769	0.763	0.767	0.770	0.766	0.764	0.745

Table 12: **Power of Multivariate Approximate Exact Tests: high correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. (*) indicates model selection on precision matrix-transformed covariates and outcomes. Results based on Bickel's adjusted CDF are indicated by (Sm). Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Approximate Exact (Ind)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.140	0.170	0.200	0.181	0.217	0.211	0.248	0.181	0.191	0.278	0.232
15	0.197	0.270	0.316	0.280	0.328	0.306	0.340	0.279	0.299	0.355	0.322
25	0.268	0.411	0.438	0.414	0.453	0.421	0.458	0.412	0.422	0.463	0.430
25	0.213	0.328	0.351	0.342	0.365	0.352	0.367	0.340	0.344	0.375	0.342
50	0.355	0.532	0.545	0.541	0.554	0.547	0.557	0.536	0.542	0.554	0.522
100	0.557	0.733	0.744	0.740	0.749	0.743	0.749	0.734	0.736	0.744	0.717

Approximate Exact (Ind-Sm)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.142	0.173	0.205	0.185	0.223	0.216	0.252	0.185	0.194	0.284	0.235
15	0.197	0.274	0.320	0.284	0.333	0.310	0.344	0.281	0.303	0.359	0.324
25	0.270	0.413	0.441	0.417	0.454	0.423	0.460	0.415	0.425	0.466	0.432
25	0.215	0.332	0.354	0.345	0.369	0.356	0.371	0.344	0.349	0.379	0.344
50	0.357	0.535	0.549	0.546	0.558	0.550	0.561	0.540	0.548	0.558	0.525
100	0.559	0.734	0.746	0.740	0.751	0.744	0.751	0.736	0.738	0.746	0.719

Approximate Exact (Exch)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.174	0.172	0.198	0.183	0.215	0.212	0.247	0.181	0.191	0.281	0.234
15	0.239	0.274	0.320	0.287	0.333	0.311	0.345	0.283	0.300	0.359	0.322
25	0.321	0.413	0.443	0.416	0.453	0.423	0.458	0.413	0.427	0.466	0.430
25	0.266	0.334	0.356	0.348	0.371	0.360	0.374	0.341	0.350	0.380	0.346
50	0.442	0.538	0.553	0.550	0.562	0.556	0.563	0.546	0.548	0.561	0.526
100	0.649	0.740	0.748	0.748	0.751	0.754	0.752	0.742	0.742	0.753	0.732

Approximate Exact (Exch-Sm)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.176	0.177	0.201	0.187	0.219	0.215	0.252	0.184	0.193	0.284	0.237
15	0.240	0.278	0.323	0.291	0.337	0.313	0.348	0.286	0.304	0.363	0.325
25	0.323	0.415	0.445	0.419	0.456	0.426	0.459	0.415	0.430	0.467	0.431
25	0.268	0.339	0.360	0.353	0.374	0.365	0.377	0.346	0.354	0.385	0.349
50	0.443	0.542	0.558	0.554	0.565	0.559	0.566	0.550	0.552	0.565	0.528
100	0.650	0.743	0.751	0.750	0.752	0.755	0.754	0.744	0.746	0.755	0.733

Table 13: **Power of Multivariate Exact Tests: high correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Exact (Ind)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.175	0.188	0.231	0.200	0.250	0.234	0.283	0.199	0.215	0.330	0.275
15	0.222	0.294	0.347	0.304	0.367	0.331	0.380	0.302	0.330	0.410	0.366
25	0.295	0.436	0.472	0.442	0.483	0.450	0.487	0.437	0.451	0.505	0.463
25	0.231	0.351	0.379	0.365	0.392	0.378	0.398	0.367	0.374	0.409	0.370
50	0.369	0.552	0.571	0.566	0.579	0.569	0.578	0.556	0.565	0.583	0.546
100	0.571	0.748	0.758	0.754	0.762	0.759	0.761	0.752	0.752	0.761	0.732

Exact (Exch)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.217	0.189	0.229	0.201	0.250	0.234	0.284	0.199	0.215	0.330	0.277
15	0.274	0.302	0.355	0.310	0.372	0.334	0.385	0.305	0.329	0.412	0.369
25	0.343	0.438	0.474	0.442	0.483	0.450	0.490	0.439	0.457	0.503	0.467
25	0.291	0.356	0.384	0.373	0.398	0.386	0.399	0.370	0.380	0.415	0.377
50	0.461	0.558	0.574	0.572	0.585	0.577	0.587	0.564	0.572	0.589	0.553
100	0.661	0.755	0.762	0.764	0.768	0.769	0.768	0.758	0.758	0.770	0.749

Table 14: **Average Number of Baseline Covariates** selected by AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO) by sample size when outcomes were multivariate. Rows 1-3 contain results for cluster size $m_i = (20, 30)$; rows 4-6 for $m_i = (4, 6, 8)$. Results are shown for estimating $E[\mathbf{Y}_i|\mathbf{X}_i]$ considering untransformed (U) and transformed (T) covariates and outcomes.

Low Correlation

	AIC		BICn		BICm		A. LASSO	
n_a	U	T	U	T	U	T	U	T
10	8.61	9.55	6.65	7.67	3.95	5.12	7.66	8.57
15	9.02	9.33	6.48	7.05	4.10	4.98	8.22	8.79
25	9.45	9.62	6.37	7.01	4.29	5.31	8.77	9.51
25	6.84	7.65	4.11	5.08	3.13	4.15	4.51	5.28
50	7.27	7.96	3.98	4.98	3.22	4.28	4.86	5.52
100	7.82	8.49	4.23	5.28	3.55	4.67	5.93	6.50

High Correlation

	Aic		BICn		BICm		Adap Lasso	
n_a	U	T	U	T	U	T	U	T
10	10.93	9.70	8.95	7.79	5.84	5.24	11.52	9.87
15	11.30	9.44	8.76	7.28	5.99	5.26	12.34	9.65
25	11.69	9.69	8.53	7.30	6.06	5.70	13.01	9.74
25	8.06	7.81	4.99	5.35	3.70	4.41	6.81	5.70
50	8.51	8.61	4.72	5.73	3.72	4.92	7.31	5.94
100	8.86	9.67	4.66	6.46	3.80	5.75	7.94	6.43

Table 15: **Analysis of the** *Young Citizens* **study.** Covariate-adjusted method (Method), regression (OR) {AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO)}, test statistic (T) and p-value (p), with each test statistic evaluated under independence (Ind) and exchangeable (Exch) working covariance. P-values for Approx. Exact tests are calculated under Bickel's cdf for randomization test statistics. 'Unadjusted' denotes the unadjusted test.

		Ind		Exch	
Method	OR	Test Stat	P	Test Stat	P
Adjusted	AIC	58.7003	< 0.0001	53.5700	< 0.0001
	BICm	59.9557	< 0.0001	54.5695	< 0.0001
	BICm	4.5046	< 0.0001	4.6231	< 0.0001
	A. LASSO	112.0423	< 0.0001	103.4147	< 0.0001
	Unadjusted	4.1415	< 0.0001	4.3186	< 0.0001
Augmented	AIC	5.1136	< 0.0001	5.2477	< 0.0001
	BICM	5.1845	< 0.0001	5.2321	< 0.0001
	BICN	4.6400	< 0.0001	4.6565	< 0.0001
	Adaptive LASSO	5.3805	< 0.0001	5.3756	< 0.0001
Approx. Exact	AIC	3.1326	0.0017	3.3316	0.0009
	BICm	3.1431	0.0017	3.3836	0.0007
	BICn	3.1223	0.0018	3.3280	0.0009
	A. LASSO	3.1223	0.0018	3.3280	0.0009
	Unadjusted	1.6172	0.1058	2.2682	0.0233
Approx. Exact (Sm)	AIC	3.1326	0.0017	3.3316	0.0008
	BICm	3.1431	0.0017	3.3836	0.0007
	BICn	3.1223	0.0018	3.3280	0.0009
	A. LASSO	3.1223	0.0018	3.3280	0.0009
	Unadjusted	1.6170	0.1060	2.2682	0.0233
Exact	AIC	89.8329	0.0003	37.0575	0.0003
	BICm	91.9124	0.0007	36.5084	0.0003
	BICn	88.8094	0.0007	36.5876	0.0007
	A. LASSO	88.8094	0.0007	26.5876	0.0007
	Unadjusted	434.8410	0.1043	71.4085	0.1200