Web-based Supplementary Materials for "Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models," by Michael Rosenblum and Mark J. van der Laan.

List of Appendices:

Appendix A: Details of Simulation Studies from Section 6 and Additional Simulation Results

We first describe in full detail the hypothesis testing methods M0-M5 that were compared in Section 6. Next, we present results of simulations showing the Type I error for these methods for various data generating distributions satisfying the null hypothesis (2), for different sample sizes. Then we present results from additional simulations comparing the power of the different methods under various additional working models, data generating distributions, and sample sizes. Next, we discuss selection of coefficients for use in our regression-based method, and the possibility of combining test statistics based on several methods or working models. Lastly, we give the R code used for our simulations. Note that each reported value in the tables below is the result of 100,000 Monte Carlo simulations.

Details of Hypothesis Testing Methods M0-M5

We give the details of hypothesis testing methods M0-M5 below:

Hypothesis Testing Methods:

M0: Regression-based test: This is the hypothesis testing method (*) described in Section 4. The hypothesis test (*) requires that one pre-specify a set of coefficients corresponding to treatment terms in the working model used. In all of the simulation results below and in the paper, we used the estimated coefficients corresponding to all terms in the working model that contain the treatment variable, and combined these into a Wald statistic, as described in Web Appendix B.

M1: Intention-to-treat based test: Reject the null hypothesis whenever the 95% confidence interval for the estimated risk difference excludes 0. This is a standard z-test for comparing two sample means.

M2: Cochran-Mantel-Haenszel test: (Cochran, 1954; Mantel and Haenszel, 1959) First, the baseline variable is discretized, and then the Cochran-Mantel-Haenszel test is run. We discretized into five levels, corresponding to the quintiles of the distribution of the baseline variable. That is, the odds ratio within strata corresponding to each of these levels was computed, and then combined into a single test statistic using the weights specified by the Cochran-Mantel-Haenszel test (see e.g. pg. 130 of (Jewell, 2004)).

M3: Permutation test: (Rosenbaum, 2002) First, the binary outcome Y is regressed on baseline variable V using a logistic regression working model for $P(Y = 1|V)$, which we define in the subsection on working models below. Pearson residuals for each observation are calculated based on the model fit. Then, the residuals for observations in which $A = 1$ are compared to those for $A = 0$ using the Wilcoxon rank sum test.

M4: Targeted Maximum Likelihood based test: (Moore and van der Laan, 2007; van der Laan and Rubin, 2006) The risk difference is estimated, adjusting for the baseline variable

using the targeted maximum likelihood approach; the null hypothesis is rejected if the 95% confidence interval for the risk difference excludes 0. The details of this approach are given in Section 3.1 of (Moore and van der Laan, 2007).

M5: Augmented Estimating Function based test: (Tsiatis et al., 2007; Zhang et al., 2007) The log odds ratio is estimated, using an estimating function that is augmented to adjust for the baseline variable; the null hypothesis is rejected if the 95% confidence interval for the log odds ratio excludes 0. The details are given in Section 4 of (Zhang et al., 2007) (where the "direct method" was used).

Working Models 1, 2, and 3 from the Paper

In Section 6 of the paper, we gave informal descriptions of three working models used in the simulations there. We now formally define these working models. Consider the working models used by methods M0, M4, and M5. (The working models used by method M3 need to be slightly different, and we deal with these next.) These working models are for the probability that outcome $Y = 1$ given treatment A and baseline variable V. Working Model 1, which is correctly specified for all the data generating distributions defined in Section 6, is

$$\text{logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV\right). \tag{A.1}$$

Working Model 2 has a different functional form than the true data generating distributions. Working Model 2 is defined as

$$\text{logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V^2 + \beta_3 AV^2\right). \tag{A.2}$$

Working Model 3 incorporates a "noisy" version of the baseline variable V, to represent measurement error. More precisely, Working Model 3 is obtained by replacing V in (A.1) by the variable $V' = \text{sign}(V) + W$, where $\text{sign}(x)$ is 1 for $x > 0$, 0 if $x = 0$, and -1 for $x < 0$, and where W is a standard normal random variable independent of V, A, Y.

Working Models 2 and 3 are misspecified for the data generating distributions corresponding to those described in Section 6 for which the outcome is generated according to the model $P(Y = 1|A, V) = \text{logit}^{-1}(A + V)$ or the model $P(Y = 1|A, V) = \text{logit}^{-1}(A + V - AV)$. (Working Models 2 and 3 are correctly specified when the outcome is generated according to $P(Y = 1|A, V) = \text{logit}^{-1}(A)$, as is any logistic regression model containing A as a main term.)

Method M3 (the permutation-based test) requires a working model for the probability that outcome $Y = 1$ given just the baseline variable V. In order to make the comparison with the other methods a fair one, we define Working Model 1 used by method M3 so that it is correctly specified. This requires use of a slightly complex logistic regression model: $\text{logit}^{-1}(\beta_0 + \beta_1 f(V))$, where we set

$$f(V) = \text{logit}^{-1}\left\{[\text{logit}(P(Y = 1|1, V)) + \text{logit}(P(Y = 1|0, V))]/2\right\}.$$

Working Model 2 for method M3 is formed by instead using $f(V) = V^2$. Working Model 3 for method M3 is formed by instead using $f(V) = \text{sign}(V) + W$, where W is a standard normal random variable independent of V, A, Y.

Type I error

We now explore the Type I error of methods M0-M5, using simulations. Since all the methods we consider have asymptotically correct Type I error, the remaining question is how large their Type I error is for different sample sizes under various data generating distributions. We consider two cases: first, we look at what happens when we use Working Model 1 defined above, under various data generating distributions for which the null hypothesis is true. Next, we look at what happens to Type I error when working models include a large number of variables. To summarize our findings: in all cases considered, Type I error was either correct (at most 0.05) or nearly correct (at most 0.06) for all methods in all situations we considered.

We consider five different data generating distributions described in Table 1 of this Web

Appendix below. For each of these five distributions, we consider two sample sizes: 200 and 400 subjects.

[Table 1 about here.]

For both these sample sizes and all five of the data generating distributions considered in Table 1 of this Web Appendix, the Type I error behavior of all the methods was either correct ($\leqslant 0.05$) or nearly correct ($\leqslant 0.06$).

We now consider what happens to Type I error when models use a large number of baseline variables. This is of interest since it is useful to know if using a long list of predictors could adversely affect Type I error for moderate sample sizes. We only consider methods M0, M3, M4, M5, since the other methods do not use working models.

We consider data generated as follows: first, we generate 10 baseline variables $V_1, \ldots, V_{10}$, each of which is normally distributed with mean 0 and variance 1 and independent of the others; next, the outcome Y is set to be 1 with probability $\text{logit}^{-1}((V_1 + \ldots + V_{10})/\sqrt{10})$, and 0 otherwise; the treatment variable A is binary and is generated independent of $V_1, \ldots, V_{10}$ and Y. Since in this case the treatment has no effect on the outcome, the null hypothesis (2) is true; thus, any rejections of the null are false rejections (Type I error).

We look at Type I error for methods M0, M3, M4, M5 when they use working models containing different numbers of the baseline variables. We let methods M0, M4, and M5 use the following working model for $P(Y = 1|A, V)$:

$$m(A, V|\beta) = \text{logit}^{-1}(\beta_0 + \beta_0' A + \beta_1 V_1 + \ldots + \beta_j V_j),$$

for each $j \in \{2, 4, 6, 8, 10\}$; this corresponds to five different working models, and we look at Type I error when each one is used. For method $M3$, we use as working model for $P(Y = 1|V)$, the following: $\text{logit}^{-1}(\beta_0 + \beta_1 V_1 + \ldots + \beta_j V_j)$, for each $j \in \{2, 4, 6, 8, 10\}$. The results of simulations using this set of working models are given in Table 2 of this Web Appendix below.

[Table 2 about here.]

For all the methods considered (M0, M3, M4, M5), the Type I error at nominal level 0.05 was always at most 0.06, for all working models considered (ranging from having 4 terms to 12 terms).

Power under Additional Working Models and Data Generating Distributions

Table 3 of this Web Appendix gives the approximate power using the same set of data generating distributions as in Table 1 of the paper, but under several other misspecified working models. Working Model 4 is similar to Working Model 1, except that the logit link is replaced by the probit link. Working Model 5 is similar to Working Model 2, except that we use the function $\sqrt{|V|}$ instead of V^2. Working Model 6 is similar to Working Model 3, except that we replace V in (A.1) by the variable $V' = \text{sign}(V) + 2W$, where W is a standard normal random variable independent of V, A, Y.

We can compare the power of methods M0-M5 between Table 1 in the paper and Table 3 of this Web Appendix, under the various working models considered. Comparing Working Model 4, in which the probit link was used while the data was actually generated using the logit link, to Working Model 1 (which used the correct link function), we see virtually no difference in power; the only difference is that the power of method M0 is slightly *increased* under Working Model 4, under data generating distribution 2.[1] Comparing Working Models 2 and 5, where the functional forms of the working models were wrongly set using V^2 and $\sqrt{|V|}$, respectively, instead of V, we see the power is quite similar. Comparing Working Models 3 and 6, where noise was added to the baseline variable (with more noise added to

[1]This occurs since for data generating distribution 2, under Working Model 4 the expected value of the estimated coefficient $\hat{\beta}_1$ is smaller than the corresponding expected value under Working Model 1, but the robust standard error is also smaller under Working Model 4 than under Working Model 1; it turns out that the magnitude of the reduction in standard error, in this case, is relatively more than the magnitude in reduction in expected values, resulting in a slightly increased power overall. (Under both working models, the expected value of $\hat{\beta}_3$ is 0.)

Working Model 6), we see that the additional noise added to Working Model 6 has almost no effect, except for reducing the power of method M0 under data generating distribution 3 defined in the paper. The power of method M0 decreases when more noise is added since this added noise attenuates the coefficients corresponding to the treatment variables. This indicates that the power of method M0 can be sensitive to measurement error, depending on how much it erodes the coefficients being used in the hypothesis test (*).

[Table 3 about here.]

In Table 4 of this Web Appendix, we further examine for which data generating distributions the regression-based method M0 has more power than the other methods we considered. We consider a different set of data generating distributions than in the above Table 3 of this Web Appendix. Working Model 3 (defined above) is used, and is misspecified for these data generating distributions. The first column in Table 4 of this Web Appendix corresponds to a data generating distribution for which the treatment A always helps (increases probability that $Y = 1$), within all strata of the baseline variable V. The second column corresponds to a data generating distribution for which the treatment sometimes helps (when $V^2 < 1.5$) and sometimes hurts (when $V^2 > 1.5$). The third column corresponds to a data generating distribution for which the treatment always helps or has no effect, within each stratum of V. The regression-based method outperforms the other methods in column 3 and performs as well as the others in column 2. The simulation results from Table 1 of the paper, and Tables 3 and 4 of this Web Appendix, are consistent with the regression method performing well, in comparison to methods M1-M5, when the treatment effect is large in some subpopulations and either negative or null in other populations. However, since this only holds for the data generating distributions we considered, we caution against generalizing this finding to all data generating distributions. In particular, when the regression model is severely misspecified, we expect the regression-based method to perform quite poorly.

[Table 4 about here.]

Coefficient Selection and Combining Test Statistics.

Here we focus on two issues related to the hypothesis testing methods we have considered. First, for the regression-based method of this paper, we consider the question of which coefficients β_i from the working model to use. We then consider ways to combine tests based on several methods or working models. Both of these are open problems, and we only outline several ideas here.

We first turn to the problem of selecting a subset of the coefficients β_i from a given working model to use in our hypothesis test (*). Recall from Section 4 that using any subset of the coefficients β_i corresponding to terms containing the treatment variable A, the hypothesis test (*) based on a Wald statistic combining the estimates of these coefficients (as described in Web Appendix B) will have asymptotically correct Type I error under the assumptions of Sections 3 and 4. Thus, the choice of which coefficients to use can be based solely on power. However, this choice of coefficients must be made prior to looking at the data. In Table 5 of this Web Appendix below, we give the power for the regression-based method M0 of our paper, using correctly specified Working Model 1 (A.1), based on different sets of coefficients. The data generating distributions are those defined in Section 6 and used in Table 1 of the paper.

[Table 5 about here.]

The results in Table 5 of this Web Appendix are consistent with it being advantageous to use both β_1 and β_3 if there are strong main effects and interaction effects (as in the data generating distribution in column 3). The power using both β_1 and β_3 is the same as when using only β_1 for data generating distribution 1, in which the baseline variable is independent of the outcome. For data generating distribution 2, where the baseline variable does influence the outcome, and there is no interaction effect, using β_1 gives more power than using both

β_1 and β_3. It seems quite risky to use only β_3, which corresponds to the interaction term, when there is no interaction effect (columns 1 and 2) in the data generating distribution, since in these cases power is quite low. Thus, β_3 should not be used alone in the hypothesis test (*) if no interaction is suspected. In general the best choice of coefficients is a function of the unknown data generating distribution, and so no general rule for choosing a subset will work in all situations. However, since the test using both β_1 and β_3 does relatively well in the situations in Table 5 of this Web Appendix, and since it is able gain power from both main and interaction effects, it may make sense in many situations to use both of these.

A natural and important question, raised by a referee of this paper, is whether it ever makes sense, when using the regression-based method of our paper, to use a working model but not use all coefficients corresponding to terms containing the treatment variable. That is, since Type I error is asymptotically correct for any working model, rather than use only a subset of the possible coefficients from a working model, why not just use a different working model that omits terms corresponding to unused coefficients? For example, instead of using working model

$$\text{logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV\right), \tag{A.3}$$

and only coefficient β_1, why not just use the following working model:

$$\text{logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V\right), \tag{A.4}$$

with coefficient β_1? Our simulations are consistent with an answer of "yes" to the above question. That is, in our simulations, for the data generating distributions 1, 2, and 3 (defined in Section 6), the working model and corresponding set of coefficients with the most power was always either model (A.3) above using both coefficients β_1, β_3 (for data generating distribution 3) or model (A.4) above using coefficient β_1 (for data generating distributions 1 and 2).[2] However, we do not have a proof that it is always better, in terms of

[2]The power for these methods are given in Table 1 of the paper and Tables 5 and 6 of this Web Appendix.

power, to restrict to using all coefficients corresponding to terms containing the treatment variable in a given working model. This is an important open research question.

We now consider ways to combine test statistics based on several methods or working models. The goal is to construct a single test based on several test statistics, so that the Type I error of the combined test is asymptotically correct. This is appealing since one may be able to simultaneously have adequate power against different sets of alternatives targeted by different tests. Since test statistics based on different methods or working models will likely be correlated with each other, we must use appropriate methods for combining these. For example, we could use the Bonferroni multiple testing correction to combine test statistics from method M0 (regression-based method of this paper) with method M4 (targeted maximum likelihood estimation), where the combined test rejects if the p-value from either of these methods is less than 0.025 (so as to maintain overall nominal level 0.05). The power of this combined test, when both methods use Working Model 1 under the data generating distributions 1, 2, and 3 defined in Section 6, are, respectively, 0.89, 0.77, and 0.88. Comparing this to the power of each of the individual methods M0-M5 under these data generating distributions (see Table 1 of the paper), we see that the combined test has consistently good power for every data generating distribution (though it is never the best), while all the other methods do well for some data generating distributions and poorly in others. Thus, it may be advantageous to use the combined method.

As another example, we look at combining test statistics based on the regression method of our paper, where each test statistic is based on a different working model. Consider the following two working models:

$$\text{logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V\right), \tag{A.5}$$

and

$$\text{logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV\right). \tag{A.6}$$

In this example, we use both coefficients β_1 and β_3 to carry out the hypothesis test using the second working model. Table 6 of this Web Appendix gives the power, at data generating distributions 1, 2, and 3, as in the previous example, based on each working model separately, and then based on a combined test statistic using the Bonferroni correction as in the previous example. The first working model above (with just main terms) gives more power under data generating distributions 1 and 2 (where no interaction effect is present). The second working model above has much more power at data generating distribution 3 where there is a strong interaction effect. The test based on combining test statistics from both working models performs well under all the data generating distributions (though is never the most powerful).

[Table 6 about here.]

Other methods for combining correlated test statistics can be found in (Hochberg and Tamhane, 1987; Westfall and Young, 1993; van der Laan and Hubbard, 2005). It is an open problem to determine which tests and working models might be combined to form tests that have adequate power at a variety of alternatives of interest, and still have asymptotically correct Type I error regardless of whether the working model is correctly specified.

R Code for Power and Type I Error Simulations

Below, we give the R code for the simulations from this appendix and from Section 6. This particular set of code is for the simulations when data are generated from the logistic regression model: $P(Y = 1|A, V) = \text{logit}^{-1}(A + V - AV)$. Comments in the code below explain how to adapt the code to other data generating distributions. The code below is for Working Model 1 defined above, but can be easily adapted for the other working models considered.

```
# Load library containing function to compute sandwich estimator:
```

```
library(sandwich)
# We give the code for each testing method first.
# Each function returns a (2-sided) p-value.
# M0: Regression-based method
M0 <- function(){
    A_coefficient_estimate <- flm$coefficient[c(2,4)]
    A_SE_estimate <- vcovHC(flm)[c(2,4),c(2,4)]
    pvalue <- 1- pchisq(t(A_coefficient_estimate) %*% solve(A_SE_estimate)
       %*% A_coefficient_estimate,df=2)
    return(pvalue)}
# M1: Intention-to-treat test
M1 <- function(){
    estimated_risk_difference <- p1 - p0
    estimated_SE <- sqrt((p1*(1-p1)/sum(A)) + (p0*(1-p0)/sum(1-A)))
    pvalue <- 2*pnorm(-abs(estimated_risk_difference)/estimated_SE)
    return(pvalue)}
# M2: Cochran-Mantel-Haenszel test
M2 <- function(){
    quintiles <- quantile(V,probs=c(0.2,0.4,0.6,0.8))
    Vstratum <- ifelse(V < quintiles[1], 1, ifelse(V< quintiles[2],2,
        ifelse(V<quintiles[3],3,ifelse(V<quintiles[4],4,5))))
    cmharray <- array(0,c(2,2,5))
    for (i in 1:5) {
        cmharray[1,1,i] <- sum(A*Y*as.numeric(Vstratum==i))
        cmharray[1,2,i] <- sum(A*(1-Y)*as.numeric(Vstratum==i))
```

```
        cmharray[2,1,i] <- sum((1-A)*Y*as.numeric(Vstratum==i))
        cmharray[2,2,i] <- sum((1-A)*(1-Y)*as.numeric(Vstratum==i))}
    cmh<-mantelhaen.test(cmharray,alternative="two.sided",exact=TRUE)
    return(cmh$p.value)}
# M3: Permutation based test (based on Wilcoxon rank sum test)
M3 <- function(){
    logit <- function(x){return(log(x/(1-x)))}
    expit <- function(x){return(1/(1+exp(-x)))}
    #Construct function Z based on V so that regression model
    #logit^(-1)(beta_0 + beta_1 Z) for P(Y=1|V) is correctly specified.
    Z1 <- linearpart(modeltype,rep(1,SampleSize),V)
    Z2 <- linearpart(modeltype,rep(0,SampleSize),V)
    Z <- logit((expit(Z1)+expit(Z2))/2)
    Z[Z>10]<-10; Z[Z< (-10)] <- (-10)  # Truncate extreme values
    logisticmodel2 <- glm(Y ~ 1 + Z,family=binomial)
    # Get Pearson Residuals:
    predicted2 <- predict.glm(logisticmodel2,type="response")
    PearsonResid <- (Y - predicted2)/sqrt(predicted2*(1-predicted2))
    wtest <- wilcox.test(x=PearsonResid[A==1],y=PearsonResid[A==0])
    return(wtest$p.value)}
# M4: Targeted Maximum Likelihood for the risk difference:
M4 <- function(){
    tmle_risk_difference <- sum(Q_1 - Q_0)/SampleSize
    #Influence curve (ic) for computing standard error:
    influencecurve <- 2*A*(Y-Q_1) - 2*(1-A)*(Y-Q_0) + Q_1 - Q_0
```

```
        - tmle_risk_difference
    tmle_estimated_SE <- sqrt(mean(influencecurve^2)/SampleSize)
    pvalue <- 2*pnorm(-abs(tmle_risk_difference)/tmle_estimated_SE)
    return(pvalue)}
# M5: Augmented Estimating Function Based test:
M5 <- function(){
    temp1 <- sum(A*Y -((A-0.5)*(Q_1 - p1)))
    temp2 <- sum(Y - ((A-0.5)*(Q_1 - p1)) + (A - 0.5)*(Q_0-p0)) - temp1
    beta_0_plus_beta_1 <- log( (temp1/sum(A))/(1-(temp1/sum(A))))
    beta_0 <- log((temp2/sum(1-A))/(1-(temp2/sum(1-A))))
    beta_1 <- beta_0_plus_beta_1 - beta_0
    # Get Var(InfluenceCurve) using sandwich estimator:
    # bread^-1 * meat * bread^-1
    bread <- t(cbind(1,A)) %*% (cbind(1,A) * (1/(1+exp(-beta_0 -beta_1*A)))*
        (1-(1/(1+exp(-beta_0 -beta_1*A)))))
    meattemp <- cbind(1,A)*(Y - (1/(1+exp(-beta_0 -beta_1*A)))) -
        cbind(array(1,c(SampleSize,2)))*((A-0.5)*(Q_1 - p1)) +
        cbind(array(c(rep(1,SampleSize),rep(0,SampleSize)),
        c(SampleSize,2)))*((A-0.5)*(Q_0-p0))
    meat <- t(meattemp) %*% meattemp
    se_augmented_estimating_function <- sqrt((solve(bread)
        %*% meat %*% solve(bread))[2,2])
    pvalue <- 2*pnorm(-abs(beta_1)/se_augmented_estimating_function)
    return(pvalue)}
# Initialize counters for how often each method rejects the null hypothesis:
```

```
total_M0 <- 0; total_M1 <- 0; total_M2 <- 0;
total_M3 <- 0; total_M4 <- 0; total_M5 <- 0
# Set number of data generating iterations
iter <- 100000
SampleSize <- 200
for(count in 1:iter)
{
# Get random sample from data generating distribution:
V <- rnorm(SampleSize) + rbinom(SampleSize,1,1/2)
A <- rbinom(SampleSize, 1, 1/2)
# Set logit(P(Y = 1| A,V)) to be the function A + V - AV
### Replacing the following line with another function (such
### as just A or A + V) gives alternative data generating distributions
### used in the simulations.
eta <- A + V - A*V
P <- exp(eta)/(1+exp(eta))
Y <- rbinom(SampleSize,1,P)
# Fit logistic model with data, for use in methods M0,M4,M5
flm <- glm(Y ~ A + V + A*V,family=binomial)
# Calculate quantities used in above methods:
p1 <- ((A %*% Y)/sum(A))
p0 <- (((1-A) %*% Y)/sum(1-A))
tf0 <- as.data.frame(cbind(A=rep(0,SampleSize),V))
tf1 <- as.data.frame(cbind(A=rep(1,SampleSize),V))
Q_0 <- predict.glm(flm,type="response",newdata=tf0)
```

```
Q_1 <- predict.glm(flm,type="response",newdata=tf1)

if(M0() < 0.05) total_M0 <- total_M0 + 1

if(M1() < 0.05) total_M1 <- total_M1 + 1

if(M2() < 0.05) total_M2 <- total_M2 + 1

if(M3() < 0.05) total_M3 <- total_M3 + 1

if(M4() < 0.05) total_M4 <- total_M4 + 1

if(M5() < 0.05) total_M5 <- total_M5 + 1

}

# Print estimates of power for each method:

print("Approximate Power of Methods: M0,M1,M2,M3,M4,M5")

print(c(total_M0,total_M1,total_M2, total_M3,total_M4, total_M5)/iter)
```

Appendix B: Robust Variance Estimators

The robust variance estimators required by the hypothesis test (*) in Section 4 are straightforward to compute using statistical software. For example, in Stata, the option **vce(robust)** gives robust standard errors for the maximum likelihood estimator (Hardin and Hilbe, 2007, Section 3.6.3). In R, the function **vcovHC** in the contributed package {sandwich} gives robust estimates of the covariance matrix of the maximum likelihood estimator (R Development Core Team, 2004); the diagonal elements of this matrix are robust variance estimators for the model coefficients. All of these methods are based on the sandwich estimator (Huber, 1967), which we describe below.

We give the sandwich estimator (Huber, 1967) in the setting of generalized linear models estimated via maximum likelihood. This includes ordinary least squares estimation in linear models as a special case, since this is equivalent to maximum likelihood estimation using a Gaussian (Normal) generalized linear model. Assume we are using a regression model from Section 5 of the paper. Denote the conditional density (or frequency function for discrete

random variables) implied by the regression model by $p(Y|A,V,\beta)$. (Note that we do not require that the actual data generating distribution belongs to this model.) We denote the corresponding log-likelihood of Y given A,V, for a single subject as $l(\beta;V,A,Y) = \log p(Y|A,V,\beta)$. Assume there exists a finite, unique maximizer β^* of $E(l(\beta;V,A,Y))$, where the expectation here and below is taken with respect to the true (unknown to the experimenter) distribution of the data. (At the end of this Web Appendix (Web Appendix B), we explain why the hypothesis test (*) given in Section 4 of the paper will still have asymptotically correct Type I error, even when no such unique maximizer exists.) Let $\hat{\beta}_n$ denote the maximum likelihood estimator for sample size n. Then by Theorem 5.23 in (van der Vaart, 1998) even when the model is misspecified, the covariance matrix of $\sqrt{n}(\hat{\beta}_n - \beta^*)$ converges to the "sandwich formula":

$$\Sigma = B^{-1}W(B^{-1})^T, \tag{A.7}$$

for B the matrix with i,j entry

$$B_{ij} = E\frac{\partial^2 l}{\partial\beta_i\partial\beta_j}, \tag{A.8}$$

and W the matrix with i,j entry

$$W_{ij} = E\frac{\partial l}{\partial\beta_i}\frac{\partial l}{\partial\beta_j}, \tag{A.9}$$

where all the above derivatives are taken at $\beta = \beta^*$.

The matrix (A.7) can be estimated based on the data. First, one approximates B and W by using the empirical distribution instead of the true data generating distribution in (A.8) and (A.9), where one replaces β^* by the maximum likelihood estimate $\hat{\beta}_n$. One then combines these estimates for B and W as in (A.7) to get an estimated matrix that we denote by $\hat{\Sigma}$. Sometimes this estimated matrix is multiplied by a finite sample adjustment, such as $n/(n-t)$ where n is the sample size and t is the number of terms in the regression model. The classes of regression models defined in Section 5 are sufficiently regular that $\hat{\Sigma}$ converges

to Σ in probability as sample size goes to infinity, even when the model is misspecified. The robust variance estimates for the estimated coefficients $\hat{\beta}$ are the diagonal elements of the matrix $\hat{\Sigma}$.

We said in Assumption (A3) of Section 4 of the paper that one can select more than one coefficient from the regression model for use in the hypothesis test (*). Here we describe how to create and use a Wald statistic to test the null hypothesis given in Section 3 of the paper. Say we are using m coefficients–recall from (A3) that each such coefficient must correspond to a term containing the treatment variable. Denote the $m \times 1$ vector of estimated coefficients one has selected by $\bar{\beta}$. Denote the $m \times m$ covariance matrix resulting from the sandwich estimator restricted to these coefficients by $\bar{\Sigma}$. Then the Wald statistic we use is $w = n\bar{\beta}^T\bar{\Sigma}^{-1}\bar{\beta}$. Under the null hypothesis, this statistic has as asymptotic distribution the χ^2 distribution with m degrees of freedom. This is proved below in Appendix D. We reject the null hypothesis, then, if the statistic w exceeds the 0.95-quantile of the χ^2 distribution with m degrees of freedom (where m is the number of coefficients used).

We now consider the case when there does not exist a unique, finite maximizer β^* of the expected log-likelihood, where the expectation is taken with respect to the true data generating distribution. We prove below in Web Appendix D that for the classes of generalized linear models described in Section 5, the expected log-likelihood is a strictly concave function. This guarantees either the existence of a unique, finite maximizer, or that for large enough sample size the maximum likelihood algorithm will fail to converge. In the latter case, statistical software will issue a warning, and our hypothesis testing procedure, as described in the paragraph above the main theorem in Section 4, is to fail to reject the null hypothesis in such situations. Thus, when there does not exist a unique, finite maximizer β^* of the expected log-likelihood, Type I error converges to 0 as sample size tends to infinity.

Appendix C: Robustness Results for Test of Effect Modification

The results presented in the paper were for tests of the null hypothesis of no mean treatment effect within strata of baseline variables, as formally defined in (2). Here we prove a result for testing a different null hypothesis: that of no effect modification by selected baseline variables. This result only holds when the treatment A is dichotomous, the outcome Y is continuous, and a linear model of the form (5) is used. We now describe a regression-based hypothesis test that, in this setting, can be used to test whether V is an effect modifier on an additive scale, that is, to test the null hypothesis that the treatment effect $E(Y|A=1,V)-E(Y|A=0,V)$ is a constant. This is a weaker null hypothesis than that of no mean treatment effect within strata of V: $E(Y|A=1,V)-E(Y|A=0,V)=0$, which is the null hypothesis (2) that is the focus of the rest of the paper.

To test the null hypothesis that the treatment effect $E(Y|A=1,V)-E(Y|A=0,V)$ is a constant, we can use exactly the hypothesis testing procedure (*) given in Section 4, except that we now additionally require

(1) The treatment A is dichotomous.

(2) The outcome Y is continuous.

(3) The model used is a linear model of the form (5) that must contain the following terms: a main term A and an interaction term AV.

(4) The coefficient β_i pre-specified in assumption A3 must be the coefficient for the interaction term AV.

To illustrate this, consider a hypothesis test of the form (*), using the linear regression model $m(A,V|\beta)=\beta_0+\beta_1 A+\beta_2 V+\beta_3 AV$. We reject the null hypothesis that the treatment effect $E(Y|A=1,V)-E(Y|A=0,V)$ is a constant, whenever the ordinary least squares estimate of β_3 in this linear regression is more than 1.96 (robustly estimated) standard errors from 0.

Our results imply this test for effect modification on an additive scale is guaranteed to have asymptotic Type I error at most 0.05, even when the model is misspecified.

The proof that the above test for effect modification has the robustness property (3) follows similar arguments as the proof of the theorem from Section 4. In Appendix D below, where the proof of this theorem is given, we also show that the above test for effect modification has the robustness property (3).

Appendix D: Proof of Theorem from Section 4.

We prove the theorem from Section 4. We first restate the theorem. Assume the following:

(A1) The data are generated as described in Section 3 of the paper. That is, each subject's data, consisting of a vector of baseline variables V, treatment assignment A, and outcome Y, is an i.i.d. vector. (See below for a modified version of this assumption, in which treatments are randomly assigned to fixed proportions of subjects.) The distribution of A is set by the experimenter, and A is independent of baseline measurements V. All these variables are assumed bounded, that is, there is some $M > 0$ such that Y and all components of V have absolute values less than M with probability 1. We note that the assumption given in Section 3 that treatment variable A takes only a finite set of values, which conforms to the setup of most randomized trials, can be relaxed to allow A to be any bounded random variable independent of V; the one exception is our result for the modified version of assumption A1 below, where we do use the fact that A has a finite set of values. item[(A2)] A regression model $m(A, V|\beta)$ in one of the classes given in Section 5 is used.

(A3) β_i is a pre-specified coefficient of a term that contains the treatment variable A in the linear part of this model. Such coefficients are denoted by $\beta_j^{(0)}$ in (5) and (6) in Section 5.

One can also use more than one of these coefficients; for example, one can use a Wald statistic as described in Web Appendix B.

Consider the following hypothesis test given in Section 4:

Hypothesis Test: (*)

For concreteness, we consider hypothesis tests at nominal level $\alpha = 0.05$. The parameter β is estimated with ordinary least squares estimation if the model used is linear; otherwise it is estimated with maximum likelihood estimation. The standard error is estimated by the sandwich estimator Huber (1967), which can easily be computed with standard statistical software; we describe the sandwich estimator in detail in Web Appendix B. If a single coefficient β_i is chosen in (A3), then the null hypothesis of no mean treatment effect within strata of V is rejected at level 0.05 if the estimate for β_i is more than 1.96 standard errors from 0. Formally, this null hypothesis, defined in (2) in the paper, is that for all treatments a_1, a_2, $E(Y|A = a_1, V) = E(Y|A = a_2, V)$; note that this is equivalent to the single equality $E(Y|A, V) = E(Y|V)$. If several coefficients are chosen in (A3), one can perform a similar test based on a Wald statistic that uses the estimates of these coefficients along with their covariance matrix based on the sandwich estimator; this is described above in Appendix B.

We note that in some cases, the estimators we consider will be undefined. For example, the ordinary least squares estimator will not be unique if the design matrix has less than full rank. Also, the maximum likelihood estimator will be undefined if no finite β maximizes the likelihood of the data; furthermore, statistical software will fail to converge to a finite vector if the maximum of the likelihood is achieved at a finite β, but this β has a component whose magnitude exceeds the maximum allowed by the statistical software. *We therefore specify that regardless of whether the estimate for the coefficient β_i is more than* 1.96 *standard errors*

from 0*, we always fail to reject the null hypothesis if the design matrix has less than full rank or if the maximum likelihood algorithm fails to converge.* Since standard statistical software (e.g. Stata or R) will return a warning message when the design matrix is not full rank or when the maximum likelihood algorithm fails to converge, this condition is easy to check.

We assume that statistical software used to implement maximum likelihood estimation uses the Fisher scoring method as described in Section 2.5 of (McCullagh and Nelder, 1998). We further assume this algorithm will fail to converge to a finite vector if the maximum of the likelihood of the data is not achieved for any finite β, or if this maximum is achieved at a finite β having a component whose magnitude exceeds the maximum allowed by the statistical software. We denote the maximum magnitude allowed for any variable by the statistical software by M'.

Theorem: (A.10)

Under assumptions A1-A3, the hypothesis test (*) has the robustness property (3). That is, it has asymptotic Type I error at most 0.05*, even when the model is misspecified.*

We also prove this theorem under a modified version of assumption A1 above; this modified assumption involves treatments being randomly assigned to fixed proportions of subjects. We call this modified assumption

Assumption A1': Each treatment $a \in \{0, 1, ..., k-1\}$ is randomly assigned to a fixed proportion $p_a > 0$ of the subjects, where $\sum_{a=0}^{k-1} p_a = 1$. For each subject, there is a vector of baseline variables V and a vector of unobserved potential outcomes $[Y(0), \ldots, Y(k-1)]$, representing the outcomes that subject would have had, had he/she been assigned the different possible treatments. The set of values $[V_j, Y_j(0), \ldots, Y_j(k-1)]$ for each subject j is drawn i.i.d. from an unknown distribution. The observed data vector for each subject j is the triple $(V_j, A_j, Y_j(A_j))$, where A_j is the treatment assigned, and A_j is independent

(by randomization) of V_j and the potential outcomes $Y_j(0), \ldots, Y_j(k-1)$. All variables are bounded.

We also prove that the test for effect modification given in Web Appendix C above has the robustness property (3) in the subsection of the proof for linear models below.

Proof of Theorem: We prove this theorem in two parts. First, we prove it for the linear models described in Section 5.1; next, we prove it for the non-linear models described in Section 5.2. In this Web appendix, we prove the above theorem using the assumptions A1-A3 above. The proof of the above theorem under the modified version of assumption 1 above (that is, using assumption A1' instead of assumption A1), is given in our technical report (Rosenblum and van der Laan, 2007); the web-address for this technical report is given in the bibliography.

Throughout the proof, expectations are with respect to the true data generating distribution Q (which is unknown to the experimenter); we make no assumptions on Q beyond (A1)-(A3) above. "Convergence" refers to convergence in probability, unless otherwise stated.

Proof of Theorem for Linear Models

We prove the theorem above for the case when the model $m(A, V|\beta)$ is a linear model of form (5) from Section 5. That is, it is of the form

$$m(A, V|\beta) = \sum_{j=1}^{t} \beta_j^{(0)} f_j(A, V) + \sum_{k=1}^{t'} \beta_k^{(1)} g_k(V), \tag{A.11}$$

where $\{f_j, g_k\}$ can be any square-integrable functions such that for each term $f_j(A, V)$, we have $E(f_j(A, V)|V)$ is a linear combination of terms $\{g_k(V)\}$. We denote the parameter vector $(\beta^{(0)}, \beta^{(1)})$ simply by β. The parameter β of this model is estimated by ordinary least squares. We also prove a converse to the theorem above for linear models: When the hypothesis test (*) uses a linear model not of the form (A.11) (but still such that all terms

are square-integrable and such that the set of terms is linearly independent) then it will not have the the robustness property (3).

We denote the terms in the model $m(A, V|\beta)$ by the column vector

$$\mathbf{x} = [f_1(A, V), f_2(A, V), \ldots, f_t(A, V), g_1(V), g_2(V), \ldots, g_{t'}(V)]^T.$$

Note that this vector has $t + t'$ components. Denote the values of this vector for each subject s by $\mathbf{x}^{(s)}$. If the components of the vector $\mathbf{x}$, considered as random variables, are linearly dependent (that is, if for some non-zero column vector $\mathbf{c}$, we have $\mathbf{c}^T\mathbf{x}$ equals 0 with probability 1), then for sample size $n \geqslant t+t'$, the design matrix $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}]^T$ will not be full rank, with probability 1. Therefore, since the above algorithm, by construction, fails to reject the null hypothesis whenever the design matrix in not of full rank, it will in this case fail to reject the null hypothesis with probability 1. This implies that if the components of $\mathbf{x}$ are linearly dependent, then Type I error for our hypothesis testing procedure will be 0 for any sample size $n \geqslant t + t'$. We can therefore restrict attention to the case in which the components in $\mathbf{x}$ are linearly independent, which implies (by strict convexity of the function x^2) that there is a unique minimizer β^* of $E(Y - m(A, V|\beta))^2$. We therefore restrict to the case in which such a unique minimizer β^* exists for the remainder of this proof for linear models.

Theorem 5.23 in (van der Vaart, 1998, pg. 53) on the asymptotic normality of M-estimators implies that the ordinary least squares estimate of β is asymptotically normal and converges in probability to the minimizer β^* of $E(Y-m(A, V|\beta))^2$. We now show that such a minimizer β^* is 0 for all components corresponding to terms in the model $m(A, V|\beta)$ containing the treatment variable A, when the null hypothesis that $E(Y|A, V) = E(Y|V)$ is true. That is, in the notation of (A.11), we will show that $\beta^{*(0)} = \mathbf{0}$ when the null hypothesis is true. For the remainder of the proof, assume the null hypothesis $E(Y|A, V) = E(Y|V)$ is true.

We start by showing the unique minimizer β^* of $E(Y - m(A, V|\beta))^2$ is also the unique minimizer of $E(E(Y|V) - m(A, V|\beta))^2$. This follows since

$$\begin{aligned} E(Y - m(A,V|\beta))^2 &= E(Y - E(Y|A,V) + E(Y|A,V) - m(A,V|\beta))^2 \\ &= E(Y - E(Y|A,V))^2 + E(E(Y|A,V) - m(A,V|\beta))^2 \\ &= E(Y - E(Y|A,V))^2 + E(E(Y|V) - m(A,V|\beta))^2, \quad \text{(A.12)} \end{aligned}$$

where in the last line we used our assumption of no mean treatment effect within strata of V (that is, $E(Y|A,V) = E(Y|V)$).

We now show the unique minimizer β^* of $E(E(Y|V) - m(A,V|\beta))^2$ must have $\beta^{*(0)} = \mathbf{0}$. This follows immediately from the following lemma, setting $c(A,V,\beta) = (E(Y|V) - m(A,V|\beta))^2$. Note that $(E(Y|V) - m(A,V|\beta))^2$ is integrable for all finite β, by our assumption A1 that V, A, Y are bounded, and our restriction in (A.11) that the functions $\{f_j, g_k\}$ defining $m(A,V|\beta)$ are square-integrable.

Lemma 1: *Consider any function $h(V)$, and any function $c(A,V,\beta)$ of the form*

$$c(A,V,\beta) = \left(h(V) - \sum_j \beta_j^{(0)} f_j(A,V) - \sum_k \beta_k^{(1)} g_k(V) \right)^2, \quad \text{(A.13)}$$

where for each $f_j(A,V)$, the function $E(f_j(A,V)|V)$ is a linear combination of the terms $\{g_k(V)\}$. Assume $c(A,V,\beta)$ is integrable for any finite β. Assume that A is independent of V and that there is a unique set of coefficients β_{min} achieving the minimum $\min_\beta E(c(A,V,\beta))$. Then $\beta_{min}^{(0)} = \mathbf{0}$.

Proof of Lemma 1: Let Π be the L_2 projection of $h(V)$ on the space of linear combinations of the functions $\{g_k(V)\}$. (See Chapter 6 of (Williams, 1991) for the definition and properties of the space L_2 of square-integrable random variables.) Note that all the functions $E(f_j(A,V)|V)$ are contained in the space of linear combinations of the functions $\{g_k(V)\}$ by the assumptions of the lemma. We will show that for all j, $f_j(A,V)$ is orthogonal to $h(V) - \Pi$, which suffices to prove the lemma. This follows since $E(h(V) - \Pi) f_j(A,V) =$

$EE[(h(V) - \Pi)f_j(A,V)|V] = E[(h(V) - \Pi)E(f_j(A,V)|V)] = 0$, where the last equality follows since $E(f_j(A,V)|V)$ is orthogonal to $h(V) - \Pi$ by our choice of Π. Thus, Π is the L_2 projection of $h(V)$ on the space generated by linear combinations of the set of functions $\{f_j(A,V)\} \cup \{g_k(V)\}$. Since Π, by construction, only involves terms $g_k(V)$, the lemma is proved.

Lemma 1 applied to the function $c(A,V,\beta) = (E(Y|V) - m(A,V|\beta))^2$ gives that the unique minimizer β^* of $E(Y - m(A,V|\beta))^2$ satisfies $\beta^{*(0)} = \mathbf{0}$. Since the argument above implies that the ordinary least squares estimator $\hat{\beta}$ of β converges to β^*, we have that all the components of $\hat{\beta}$ that correspond to functions $f_j(A,V)$ in the model (A.11) converge to 0. The Theorem above for linear models of the form (A.11) then follows since the robust variance estimator given in Web Appendix B is asymptotically consistent, regardless of whether the model is correctly specified. This completes the proof that the above theorem holds for linear models of the form (A.11).

We now prove that for linear models, (A.11) is a necessary condition for the robustness property (3) to hold. That is, when the hypothesis test (*) uses a linear model not of the form (A.11) (but still such that all terms are square-integrable and such that the set of terms is linearly independent[3]) then it will not have the the robustness property (3); more precisely, for any linear model not of the form (A.11), there is a data generating distribution and a coefficient β_i corresponding to a treatment term that if used in hypothesis test (*), will result in asymptotic Type I error greater than the prespecified level α.

Consider a linear model $m(A,V|\beta) = \sum_{j=1}^{J} \beta_j^{(0)} f_j(A,V) + \sum_{k=1}^{K} \beta_k^{(1)} g_k(V)$ that does not satisfy (A.11), but for which the terms are square integrable and linearly independent

[3]We say that a set of real-valued functions $\{h_j(x)\}$ is linearly independent if $\sum c_j h_j(x) = 0$ for all x implies $\forall j, c_j = 0$.

(see footnote below for definition of linearly independent functions). We call the terms $f_j(A,V)$ "treatment terms." Then it must be that for some index j', $E(f_{j'}(A,V)|V) = \sum_a p(a) f_{j'}(a,V)$ is not a linear combination of terms $\{g_k(V)\}$, where $p(a)$ the probability mass function for A (which we could take as assigning, for example, equal probability to each treatment). Let the outcome Y be defined to equal $\sum_a p(a) f_{j'}(a,V)$. We now construct a probability mass function for V such that the minimizer β^* of $E(Y - m(A,V|\beta))^2$ has a non-zero component $\beta_i^{(0)*}$ corresponding to a term that contains the treatment variable; we argue below that this implies the hypothesis test (*) using $\beta_i^{(0)*}$ as prespecified coefficient in Assumption A3, for large enough sample size, will reject the null hypothesis will arbitrarily large probability, completing the proof.

We now turn to constructing a probability mass function (p.m.f.) for V for which the minimizer β^* of $E(Y - m(A,V|\beta))^2$ has a non-zero component $\beta_i^{(0)*}$ corresponding to a treatment term. Using similar arguments as in the proof of Lemma 1, this will be the case if, for Π defined to be the L_2 projection of Y on the subspace generated by $\{g_k(V)\}$, we have $E[(Y - \Pi) f_{j'}(A,V)] \neq 0$. This happens whenever $E(Y - \Pi)^2 > 0$, since

$$E[(Y - \Pi) f_{j'}(A,V)] = E[(Y - \Pi) \sum_{a'} p(a') f_j(a',V)] = E[(Y - \Pi)Y] = E(Y - \Pi)^2,$$

where the first equality follows from A and V being independent and Y being a function of V only, the second equality follows by our having defined Y as $\sum_a p(a) f_{j'}(a,V)$, and the third equality follows from Π being an L_2 projection. Thus, it suffices to construct a p.m.f. for V such that $E(Y - \Pi)^2 > 0$. Because we are assuming $\sum_a p(a) f_{j'}(a,V)$ is not a linear combination of terms $\{g_k(V)\}$, there is a p.m.f. $q(v)$ for V for which the projection Π of $Y = \sum_a p(a) f_{j'}(a,V)$ on $\{g_k(V)\}$ satisfies $E(Y - \Pi)^2 > 0$ and for which all the terms of the model $m(A,V|\beta)$ are linearly independent in the vector space L_2.[4] We now

[4]Such a p.m.f. for V can be constructed as follows. Recall the linear model under consideration is defined as $m(A,V|\beta) = \sum_{j=1}^{J} \beta_j^{(0)} f_j(A,V) + \sum_{k=1}^{K} \beta_k^{(0)} g_k(V)$. Define, for any real-valued v, the column vector $g(v) := [g_1(v), \ldots, g_K(v)]^t$. Then by the assumption above that $g_k(v)$ are linearly independent functions, we can find a finite set of points $v_1, \ldots, v_K$ such

can completely define our data generating distribution for the observation (V, A, Y): V has marginal p.m.f. q(v); A is independent of V and assigns probability $p(a)$ to each treatment a; Y, as defined above, is the following function of V: $E(f_{j'}(A,V)|V) = \sum_a p(a) f_{j'}(a,V)$. The above arguments then imply that for this data generating distribution, the minimizer β^* of $E(Y - m(A,V|\beta))^2$ has a non-zero component $\beta_i^{(0)*}$ corresponding to a term $(f_i(A,V))$ that contains the treatment variable. Since the ordinary least squares estimate of β is asymptotically normal and converges in probability to β^*, we have that the estimator for $\beta_i^{(0)*}$ will converge to a non-zero value. This causes the hypothesis test (*) using the estimate for $\beta_i^{(0)*}$ to reject with arbitrarily large probability as sample size tends to infinity. This completes the proof that when the hypothesis test (*) uses a linear model not of the form (A.11), then it will not have the the robustness property (3).

Before proving Theorem (A.10) for the class of generalized linear models given in Section 5.2, we show how the above lemma implies that the hypothesis test for effect modification given in Web Appendix C has the robustness property (3). Consider a linear model containing at least the main term A and interaction term V of the form $m(A,V|\beta) = \beta_0 + \beta_1 A + \beta_2 AV + \sum_i \beta_i' f_i(A,V)$. The proof above for linear models can be used unchanged up until (A.12). This is the point in the proof in which the null hypothesis (2) was used. Using instead the null hypothesis that $E(Y|A=1,V) - E(Y|A=0,V)$ is a constant b, we can replace

---

that the vectors $g(v_1), \ldots, g(v_K)$ are linearly independent. Define G to be the $K \times K$ matrix $[g(v_1), \ldots, g(v_K)]$. Define $\tilde{\beta} := G^{-1}[h(v_1), \ldots, h(v_K)]^t$, for $h(v) := \sum_a p(a) f_{j'}(a,v)$. Since we are assuming $\sum_a p(a) f_{j'}(a,V)$ is not a linear combination of terms $\{g_k(V)\}$, we can find an additional point v_{K+1} for which $\tilde{\beta}^t g(v_{K+1}) \neq h(v_{K+1})$. The set of points $v_1, \ldots, v_{K+1}$, then, is such that no linear combination of $\{g_k(v)\}$ equals $h(v)$ at all these points (since the only linear combination of $\{g_k(v)\}$ that equals $h(v)$ at $v_1, \ldots, v_K$ is $\tilde{\beta}^t g(v)$, and this does not equal $h(v)$ at $v = v_{K+1}$). Therefore, defining V to have mass $1/(1+K)$ at each such point, the projection Π of $Y = h(V)$ on the subspace spanned by $\{g_k(V)\}$ is such that $E(Y - \Pi)^2 > 0$. An extension of the above argument can be used to add more mass points in such a way that we ensure that all the terms of the model $m(A,V|\beta)$ are linearly independent in the vector space L_2, and that $E(Y - \Pi)^2 > 0$.

the expression in (A.12) by $E(Y - E(Y|A,V))^2 + E(E(Y|A=0,V) + bA - m(A,V|\beta))^2$. This shows that minimizing $E(Y - m(A,V|\beta))^2$ is equivalent to minimizing $E(E(Y|A = 0,V) + bA - m(A,V|\beta))^2$ under this null hypothesis of no effect modification. Let β^* be the unique, finite minimizer of this function (if no such minimizer exists, then as argued above Type I error goes to 0 as sample size goes to infinity). Then by Lemma 1, we have under this null hypothesis that β_2^* (which corresponds to the AV term in $m(A,V|\beta)$) is 0 and that β_1^* equals b. Since we are using robust variance estimators, we then have that the test for effect modification that rejects the null hypothesis when the design matrix has full rank and the estimate for the coefficient β_2 is more than 1.96 standard errors from 0 has asymptotic Type I error at most 0.05.

Proof of Theorem for Generalized Linear Models

We prove the theorem (A.10) above for the large class of generalized linear models defined in Section 5.2 of the paper. We repeat this definition here. Consider the following types of generalized linear models: logistic regression, probit regression, binary regression with complementary log-log link function, and Poisson regression (using the log link function). We define our class of generalized linear models to be any generalized linear model from the previous list, coupled with a linear part of the following form:

$$\eta(A,V|\beta) = \sum_{j=1}^{t} \beta_j^{(0)} f_j(A) g_j(V) + \sum_{k=1}^{t'} \beta_k^{(1)} h_k(V), \tag{A.14}$$

for any measurable functions $\{f_j, g_j, h_k\}$ such that for all j, there is some k for which $g_j(V) = h_k(V)$; we also assume the functions $\{g_j, h_k\}$ are bounded on compact subsets of $\mathbb{R}^q$, where V has dimension q. We denote the parameter vector $(\beta^{(0)}, \beta^{(1)})$ simply by β. Let $\mathbf{x}$ denote the column vector of random variables corresponding to the terms in $\eta(A,V|\beta)$. That is, denote

$$\mathbf{x} := [f_1(A)g_1(V), f_2(A)g_2(A), \ldots, f_t(A)g_t(V), h_1(V), h_2(V), \ldots, h_{t'}(V)]^T.$$

We can restrict attention to the case in which the components in $\mathbf{x}$ are linearly independent random variables, by virtue of the same arguments given above for the case of linear models; we assume the components of $\mathbf{x}$ are linearly independent for the remainder of the proof.

As described in (McCullagh and Nelder, 1998), the log-likelihood for any of the above generalized linear models can be represented, for suitable choices of functions b, d, g, as

$$l(\beta; V, A, Y) = Y\theta - b(\theta) + d(Y), \tag{A.15}$$

where θ is called the canonical parameter of the model, and is related to the parameter β through the following equality: $\dot{b}(\theta) = g^{-1}(\eta(A, V|\beta))$, where $\dot{b}(\theta) := \frac{db}{d\theta}$. and $g(\mu)$ is called the link function. For binary outcomes, the function $b(\theta) = \log(1 + e^{\theta})$ and $d(y) = 0$; for Poisson regression, in which the outcome is a nonnegative integer, $b(\theta) = e^{\theta}$ and $d(y) = -\log y!$. Note that in both cases, $\ddot{b}(\theta) := \frac{d^2b}{d\theta^2} > 0$ for all θ. Also note that $\dot{b}$ is invertible in both cases. The Theorem (A.10) also holds when a dispersion parameter is included, and holds for other families of generalized linear models, such as the Gamma and Inverse Gaussian families with canonical link functions; however, in these cases additional regularity conditions on the likelihood functions are required. For the class of generalized linear models given in Section 5.2, no additional regularity conditions beyond the linear part having the form (A.14) with all terms measurable and the functions $\{g_j, h_k\}$ bounded on compact subsets of $\mathbb{R}^q$ are needed for the theorem to hold.

Before giving the detailed proof that Theorem (A.10) above holds for models of the above type, we give an outline of the main steps in the proof. First, we show that due to the strict concavity of the expected log-likelihood for models from the exponential family, it suffices to consider the case in which the expected log-likelihood has a unique, finite maximum. Next, we show that when there is a unique, finite maximizer β^* of the expected log-likelihood, all components of β^* corresponding to terms in (A.14) that contain the treatment variable A are 0. Lastly we apply Theorem 5.39 on the asymptotic normality of maximum likelihood

estimators in (van der Vaart, 1998), completing the proof of Theorem (A.10) for the class of generalized linear models given in Section 5.2.

We now turn to proving the strict concavity of the expected log-likelihood for the class of models defined in Section 5.2. This is equivalent to showing the Hessian matrix $\mathbf{H} := \frac{\partial^2}{\partial\beta_j\partial\beta_k}E(l(\beta;V,A,Y))$ is negative definite, or in other words, that for any β and any non-zero column vector $\mathbf{a}$ of length $t+t'$, we have $\mathbf{a}^T\mathbf{H}\mathbf{a} < 0$.

Consider the case in which the link function in the generalized linear model is the canonical link for that family (that is, assuming the canonical parameter $\theta = \eta(A,V|\beta)$), which is the case for logistic regression and for Poisson regression with log link. We then have from (A.15)

$$\mathbf{H}_{ij} = E\frac{\partial^2 l}{\partial\beta_i\partial\beta_j} = -E\ddot{b}(\eta)x_i x_j.$$

Since as noted above, $\ddot{b}(\eta) > 0$ for all η, and since we have restricted to the case in which the random variables in the vector $\mathbf{x}$ are linearly independent, we have

$$\mathbf{a}^T\mathbf{H}\mathbf{a} = -E\ddot{b}(\eta)\left(\mathbf{a}^T\mathbf{x}\right)^2 < 0.$$

Thus, we have shown that the expected log-likelihood $E(l(\beta;V,A,Y))$ is a strictly concave function of β, whenever the canonical link function is used in a generalized linear model.

We now give a similar argument to that given above, but now applied to the generalized linear models from Section 5.2 that have non-canonical link functions. More precisely, we will show the expected log-likelihood is a strictly concave function of β for binary regression models using either of the following link functions:

(1) $\Phi^{-1}(\mu)$, for Φ the distribution function of a standard normal random variable, which corresponds to probit regression,

(2) $\log(-\log(1-\mu))$, called the complementary log-log link.

For a generic link function g, let γ denote its inverse. For a binary outcome Y, taking values in $\{0,1\}$, the log-likelihood for a single subject is:

$$l(\beta; V, A, Y) = \log\left(\gamma(\eta)^Y (1-\gamma(\eta))^{1-Y}\right), \tag{A.16}$$

where $\eta = \eta(A, V|\beta)$. The Hessian matrix of the expected log-likelihood is then

$$\mathbf{H} = E\mathbf{x}\mathbf{x}^T z, \tag{A.17}$$

$$\text{for } z = Y\frac{d^2}{d\eta^2}\left(\log\gamma(\eta)\right) + (1-Y)\frac{d^2}{d\eta^2}\left(\log(1-\gamma(\eta))\right),$$

and $\mathbf{x}$ containing the terms in $\eta(A, V|\beta)$ as defined in (A.14) above. Thus, to show that the expected log-likelihood is strictly concave, it suffices to show that $\log\gamma$ and $\log(1-\gamma)$ are strictly concave; this follows since when these two functions are strictly concave, z defined above is strictly negative, and so for any non-zero vector $\mathbf{a}$, we have

$$\mathbf{a}^T\mathbf{H}\mathbf{a} = E(\mathbf{a}^T\mathbf{x})^2 z < 0.$$

We now verify that for the two link functions listed above, their inverses are such that $\log\gamma$ and $\log(1-\gamma)$ are strictly concave.

(1) For link function $g(\mu) = \Phi^{-1}(\mu)$, its inverse is Φ, and $\frac{d^2}{d\eta^2}\left(\log\Phi(\eta)\right) = -1$, and $\frac{d^2}{d\eta^2}\left(\log(1-\Phi(\eta))\right) = \frac{d^2}{d\eta^2}\left(\log(\Phi(-\eta))\right) = -1$.

(2) For link function $g(\mu) = \log(-\log(1-\mu))$, its inverse is $\gamma(\eta) = 1 - e^{-e^\eta}$, and

$$\frac{d^2}{d\eta^2}\left(\log(1-e^{-e^\eta})\right) = \{(1-e^\eta-e^{-e^\eta})e^{\eta-e^\eta}\}/\left(1-e^{-e^\eta}\right)^2 < 0,$$

since the term $1-e^\eta-e^{-e^\eta}$ is strictly negative for all η, which follows by substituting $-e^\eta$ for x in the well-known inequality $e^x > 1+x$ for all $x \neq 0$; also, $\frac{d^2}{d\eta^2}\left(\log e^{-e^\eta}\right) = -e^\eta < 0$.

Thus, for both of the above link functions, $\log\gamma$ and $\log(1-\gamma)$ are strictly concave, which by (A.17) implies that the expected log-likelihood is strictly concave.

The purpose of having shown above that the expected log-likelihood is strictly concave for each of the models in the class defined in Section 5.2, is that this implies one of the following two cases must hold:

Case 1: There is no finite β that maximizes $E(l(\beta; V, A, Y))$.

Case 2: There is a unique maximizer β^* of $E(l(\beta; V, A, Y))$.

Consider what happens if Case 1 holds. Let Q_n be the distribution of the maximizer $\hat{\beta}_n$ of the log-likelihood $\sum_{j=1}^{n} l(\beta; V_j, A_j, Y_j)$. By Theorem 5.7 of (van der Vaart, 1998), we have for any $w > 0$ that $Q_n(|\hat{\beta}_n| > w)$ converges to 1. Thus, the maximum likelihood algorithm will fail to converge to a finite vector that is within the bounds allowed by the algorithm (as explained just before (A.10) above), with probability tending to 1, as sample size n tends to infinity. Since, by construction, the hypothesis testing algorithm given above fails to reject the null hypothesis when the maximum likelihood estimation procedure fails to converge, we have that under Case 1 the asymptotic Type I error of the above hypothesis test converges to 0. Thus, it suffices to restrict attention to when Case 2 above holds, and we assume this case holds for the rest of the proof.

When Case 2 above holds, by Theorem 5.39 on the asymptotic normality of maximum likelihood estimators in (van der Vaart, 1998), the maximum likelihood estimator for β is asymptotically normal, and converges to the unique value of β maximizing the expected log-likelihood $E(l(\beta; V, A, Y))$. Call this maximizer $\beta^* = (\beta^{*(0)}, \beta^{*(1)})$, where we partition the components of β^* as described just below (A.14). We will show $\beta^{*(0)} = \mathbf{0}$. Using (A.15), under the null hypothesis that $E(Y|A, V) = E(Y|V)$, the expected log-likelihood can be expressed as follows:

$$\begin{aligned} E(l(\beta; V, A, Y)) &= EE(l(\beta; V, A, Y)|A, V) \\ &= E\left(E(Y|A, V)\theta - b(\theta) + E(d(Y)|A, V)\right) \\ &= E\left(E(Y|V)\theta - b(\theta) + E(d(Y)|A, V)\right) \end{aligned}$$

Thus, the parameter β that maximizes $E(l(\beta; V, A, Y))$ also maximizes $E(E(Y|V)\theta - b(\theta))$. (Recall that $\theta = \dot{b}^{-1}(g^{-1}(\eta(A, V|\beta)))$, and so is a function only of A, V, and β.) Using assumption A1 above that all variables V, A, Y are bounded and our restriction in (A.14) that the terms in $\eta(A, V|\beta)$ are bounded on compact sets, it is straightforward to show for each of the

generalized linear models in the class defined in Section 5.2, that the corresponding expected log-likelihood is integrable for any finite β. The lemma below, analogous to Lemma 1, implies that β^* must have $\beta^{*(0)} = \mathbf{0}$. This lemma is the main technical contribution of this paper.

Lemma 2*: Consider any function $c(A, V, \beta)$ of the form*

$$c(A, V, \beta) = s\left(V, \sum_j \beta_j^{(0)} f_j(A) g_j(V) + \sum_k \beta_k^{(1)} h_k(V)\right). \tag{A.18}$$

where for all j, there is some k for which $g_j(V) = h_k(V)$. Assume $c(A, V, \beta)$ is integrable for any finite β. Assume that A is independent of V and that there is a unique set of coefficients β_{min} achieving the minimum $\min_\beta E(c(A, V, \beta))$. Then $\beta_{min}^{(0)} = \mathbf{0}$; that is, β_{min} assigns 0 to all coefficients in (A.18) of terms containing the variable A.

Proof of Lemma 2:

Since from elementary probability theory

$$Ec(A, V, \beta_{min}) = EE[c(A, V, \beta_{min})|A],$$

we must have for some a_0 in the range of A that

$$Ec(A, V, \beta_{min}) \geqslant E[c(A, V, \beta_{min})|A = a_0]. \tag{A.19}$$

We now construct a set of coefficients $\bar{\beta} = (\bar{\beta}^{(0)}, \bar{\beta}^{(1)})$ with $\bar{\beta}^{(0)} = \mathbf{0}$ and that we will later show attains the minimum $\min_\beta E(c(A, V, \beta))$. We leverage the assumed property of $\sum_j \beta_j^{(0)} f_j(A) g_j(V) + \sum_k \beta_k^{(1)} h_k(V)$ that for all j, there is some k for which $g_j(V) = h_k(V)$; this property allows us to group all terms together that correspond to the same $h_k(V)$ as follows:

$$\sum_j \beta_j^{(0)} f_j(A) g_j(V) + \sum_k \beta_k^{(1)} h_k(V) = \sum_k \left(\sum_{j: g_j(V) = h_k(V)} \beta_j^{(0)} f_j(A) + \beta_k^{(1)} \right) h_k(V).$$

This motivates defining $\bar{\beta}$ as follows: $\bar{\beta}_j^{(0)} := 0$, for all j, and

$$\bar{\beta}_k^{(1)} := \sum_{j: g_j(V) = h_k(V)} \beta_{min,j}^{(0)} f_j(a_0) + \beta_{min,k}^{(1)}, \text{ for all } k.$$

Note that $\bar{\beta}$ has the following property:

$$\begin{aligned} Ec(a_0, V, \bar{\beta}) &= Ec(a_0, V, \beta_{min}) \\ &= E[c(A, V, \beta_{min})|A = a_0] \end{aligned} \tag{A.20}$$

where the first equality follows by construction of $\bar{\beta}$ that for all v, $c(a_0, v, \bar{\beta}) = c(a_0, v, \beta_{min})$, and the second equality follows by the independence of A and V.

We now show that $\bar{\beta}$ minimizes $E(c(A, V, \beta))$. Since by definition $\bar{\beta}$ assigns 0 to the coefficients of all terms containing the variable A, we have $c(a, V, \bar{\beta})$ has the same value for all a. We then have

$$\begin{aligned} Ec(A, V, \bar{\beta}) &= Ec(a_0, V, \bar{\beta}) \\ &= E[c(A, V, \beta_{min})|A = a_0] \\ &\leqslant Ec(A, V, \beta_{min}), \end{aligned}$$

where the second line follows from the above property (A.20) of $\bar{\beta}$ and the third line follows from our choice of a_0 satisfying (A.19). Thus $\bar{\beta}$ minimizes $E(c(A, V, \beta))$, and by our assumption of a unique minimizer of this quantity, $\bar{\beta} = \beta_{min}$. Since $\bar{\beta}$ by construction assigns 0 to the coefficients of all terms containing the variable A, the lemma follows.

We can apply Lemma 2 to $c(A, V, \beta) =$

$$E(E(Y|V)\theta - b(\theta)) = E\left\{E(Y|V)\dot{b}^{-1}(g^{-1}(\eta(A, V|\beta))) - b\left(\dot{b}^{-1}(g^{-1}(\eta(A, V|\beta)))\right)\right\},$$

since we had restricted to the case in which the expected log-likelihood has a unique, finite maximizer. This implies that the unique maximizer β^* of the expected log-likelihood $E(l(\beta; V, A, Y))$ has $\beta^{*(0)} = \mathbf{0}$. This completes the argument above that the maximum likelihood estimator for β converges to β^* with $\beta^{*(0)} = \mathbf{0}$. The Theorem (A.10) above then follows for the generalized linear models given in Section 5.2 since each of the robust variance estimators in Web Appendix B is asymptotically consistent, regardless of whether the model

is correctly specified, for this class of models.

Q.E.D.

We now give examples of a models that do not have the robustness property (3). In general, median regression models, that is, working models $m(A, V|\beta)$ for the median of Y given A, V, when fit with maximum likelihood estimation do not have the robustness property (3). This is not surprising, since the null hypothesis we consider is in terms of the mean (not the median) of Y given A, V. We note that even under the null hypothesis that the conditional median of Y given A, V, does not depend on A, the robustness property (3) still does not hold for median regression.

Appendix E: R Code for Data Example from Section 7

We give R Code for the hypothesis test (*) from the data example in Section 7. This hypothesis test used the following logistic regression model for the probability of HIV infection by the end of the trial, given treatment arm A and baseline variables V_1, V_2, V_3, V_4, V_5:

$$m(A, V|\beta) = \text{logit}^{-1}(\beta_0 + \beta_1 A + \beta_2 V_1 + \beta_3 V_2 + \beta_4 V_3 + \beta_5 V_4 + \beta_6 V_5 + \beta_7 A V_1 + \beta_8 A V_2 + \beta_9 A V_3 + \beta_{10} A V_4 + \beta_{11} A V_5). \quad \text{(A.21)}$$

The following R code executes hypothesis test (*):

```
rct_data <- data.frame(Y = HIV_STATUS, A = ARM, V1 = CONDOM, V2 = AGE,
V3 = HSV, V4 = SUBJECTRISK, V5 = PARTNERRISK)
logisticmodel <- glm(Y~1+A+V1+V2+V3+V4+V5+A:(V1+V2+V3+V4+V5),
family=binomial,data=rct_data)
# Get coefficients corresponding to terms containing A (treatment):
A_coefficient_estimate <- logisticmodel$coefficient[c(2,8:12)]
```

```
V_cov_matrix <- vcovHC(logisticmodel)[c(2,8:12),c(2,8:12)]
wald_statistic <- t(A_coefficient_estimate) %*%
solve(V_cov_matrix) %*% A_coefficient_estimate
pvalue <- 1-pchisq(wald_statistic,df=6)
```

We also give the estimated coefficients and robust standard errors for each coefficient in Table 7 of this Web Appendix

[Table 7 about here.]

The permutation-based test M3, applied to this data set, was computed using the following R code:

```
logisticmodel2 <- glm(Y~(1+V1+V2+V3+V4+V5)^2,family=binomial,data=rct_data)
# Get Pearson Residuals:
logisticmodel2predicted <- predict.glm(logisticmodel2,type="response")
PearsonResid <- (Y - logisticmodel2predicted)/sqrt(logisticmodel2predicted*
(1-logisticmodel2predicted))
# Do Wilcoxon Rank Sum Test on Residuals for A=1 vs. A=0
wilcoxontest <- wilcox.test(x=PearsonResid[A==1],y=PearsonResid[A==0])
return(wilcoxontest$p.value)
```

Appendix F: Comparison of Superpopulation Inference to the Randomization Inference Approach of Rosenbaum (2002)

We compare the framework of superpopulation inference, used in our paper, to the randomization inference framework and hypothesis tests of Rosenbaum (2002). For a more detailed comparison of these frameworks, see (Lehmann, 1986, Section 5.10), Rosenbaum (2002), and Robins (2002). These frameworks differ in the assumptions they make about how data is

generated and in the hypotheses tested. Below, we discuss these differences and how they relate to the hypothesis tests considered in our paper.

We first consider the ways superpopulation inference and randomization inference differ regarding the assumptions they make about how data is generated. The main difference lies in which characteristics of subjects are assumed to be fixed (but possibly unknown) and which are assumed to be random. To emphasize the contrast between superpopulation inference and randomization inference, we use uppercase letters to denote random quantities, and lowercase letters to denote fixed (non-random) quantities. The framework of superpopulation inference we use in our paper assumes that all data on each subject are random, and can be represented as a simple random sample from a hypothetical, infinite "superpopulation." That is, we assume the set of variables observed for each subject i, denoted by (V_i, A_i, Y_i), is an i.i.d. draw from an unknown probability distribution P. Examples of work cited in our paper in which superpopulation inference is used include (Robins, 2004; Tsiatis et al., 2007; Zhang et al., 2007; Moore and van der Laan, 2007).

In contrast to superpopulation inference, the randomization inference framework used by Rosenbaum (2002) (which can be traced back to Neyman (1923)) assumes that subjects' baseline characteristics and potential outcomes (defined below) are not random, but are fixed quantities for each subject. For example, in a randomized trial setting in which baseline variables, treatment, and outcome are observed, it would be assumed that for each subject i, there is a fixed set of baseline variables v_i, and a fixed set of potential outcomes $y_i(a)$ for each possible value a of the treatment; $y_i(a)$ represents the outcome that subject i would have, if he/she would be assigned to treatment arm a. In randomization inference, the only random quantity is the treatment assignment A_i for each subject. For each subject i, the only potential outcome that is observed is the one corresponding to the treatment assignment that is actually received: $y_i(A_i)$. It is assumed that the set of observed variables for each subject i is

$(v_i, A_i, y_i(A_i))$, denoting the baseline variables, treatment assignment, and observed outcome, respectively. Note that we put v_i and $y_i(a)$ in lowercase to emphasize that these are assumed to be fixed (non-random) in the randomization inference framework. Examples of work cited in our paper in which randomization inference is used include (Freedman, 2007a,b,c).

Because of the differences in assumptions between superpopulation inference and randomization inference, the hypotheses tested differ as well. In superpopulation inference, the hypotheses refer to parameters of the underlying data generating distribution P; inferences are made about what the effects of treatment would be for the hypothetical "superpopulation" from which the subjects are assumed to have been drawn. In contrast, in randomization inference, hypotheses refer only to the fixed quantities (such as baseline variables and potential outcomes) of the subjects actually in the study. We consider examples of hypotheses tested in each of these frameworks next.

In superpopulation inference, in a two-armed randomized trial, one may test whether the mean of outcomes are affected by treatment assignment, as done by Tsiatis et al. (2007); Zhang et al. (2007); Moore and van der Laan (2007); this corresponds to testing the null hypothesis that $E(Y|A = 0) = E(Y|A = 1)$, where the expectations are taken with respect to the unknown data generating distribution P. In our paper, we test the null hypothesis (2) of no mean treatment effect within strata of baseline variables V; for a two-armed randomized trial this corresponds to $E(Y|A = 0, V) = E(Y|A = 1, V)$. Note that these null hypotheses refer to parameters of the underlying distribution P. In contrast, the hypotheses tested in the randomization inference framework of Rosenbaum (2002) are hypotheses about the potential outcomes of the subjects in the trial. For example, in a trial with two arms, the null hypothesis of no treatment effect at all would correspond to equality of the potential outcomes $y_i(0) = y_i(1)$ for each subject i; this null hypothesis is that every subject in the trial would have exactly the same outcome regardless of which treatment arm they were

assigned to. A null hypothesis considered by Rosenbaum (2002) is that of additive treatment effects; that is, for some τ_0, $y_i(1) = y_i(0) + \tau_0$ for all subjects i. This means that for every subject in the study, the impact of having received the treatment ($A = 1$) compared to the control ($A = 0$) would have been to increase the value of their outcome by exactly τ_0.

Because of the differences in assumptions and hypotheses tested in the superpopulation inference framework of our paper and the randomization inference framework of Rosenbaum (2002), the methods used for each situation are generally different. Due to these differences, the methods of Rosenbaum (2002) are in general not appropriate for testing the hypotheses considered in our paper. (The one exception is when the outcome is binary, as we further discuss below.) More precisely, the methods of Rosenbaum (2002) will not have asymptotically correct Type I error for testing the hypotheses considered in our paper in many situations. Before showing this, we explain the intuition for why this is the case. The null hypotheses of Rosenbaum (2002) imply that certain distributions are invariant under permutations of the treatment variable. For example, the null hypothesis of no treatment effect at all in randomization inference (defined above) implies that rearranging which subjects got the active treatment vs. control will have no effect on the outcomes. Therefore, under this null hypothesis, permutation-based tests (such as the Wilcoxon rank-sum test described in method M3 above) are appropriate. This is in stark contrast with the implications of the null hypotheses (2) considered in our paper, which only imply that conditional means of the outcome given baseline variables are unaffected by treatment. The distribution of the data, under our framework and null hypothesis (2), is not invariant to permutations of the treatment variable, and so permutation tests will not in general work, except in the special case considered next.

In the special case in which the outcome Y of a randomized trial is binary (taking values 0 or 1), our null hypothesis (2) is that $E(Y|A = 1, V) = E(Y|A = 0, V)$. Assuming

treatment assignment A is independent of the baseline variables V (due to randomization), this null hypothesis is equivalent to the outcome Y and baseline variables V being mutually independent of the treatment A. This means that under our null hypothesis (2), the treatment has no effect at all on the distribution of the outcome, and so permutation tests makes sense; in fact, they have exactly correct Type I error (not just asymptotically correct Type I error) under the null hypothesis (2) in this special case. Thus, in this case, the permutation tests of Rosenbaum (2002) have the robustness property (3), even under the superpopulation framework of this paper defined in Section 3. In our simulations in Section 6 comparing the power of various methods for a binary outcome, the permutation-based method (M3) generally performed well in terms of power; compared to the regression-based method (M0) of this paper, the permutation-based method sometimes had more power and sometimes had less power.

We now give an example to show that the permutation-based methods of Rosenbaum (2002) are not guaranteed to have the robustness property (3) when testing the null hypothesis (2) (except in the case of binary outcomes, as discussed above). We show that the basic permutation-based method of Rosenbaum (2002) does not have asymptotically correct Type I error under the null hypothesis (2), under the framework of our paper (given in Section 3) for the following data generating distribution: Let X be a random variable with the following skewed distribution: $X = 2$ with probability $1/3$ and equals -1 with probability $2/3$; thus, the mean of X is 0. Define baseline variable V to be a random variable independent of X such as a standard normal random variable. Let the treatment A take values 0 and 1 each with probability $1/2$, mutually independent of $\{V, X\}$. Let the outcome $Y = (2A-1)X+V$. Then the null hypothesis (2) is true since $E(Y|A, V) = (2A-1)EX+V = V$ and so is a function of V only (and not a function of A). This also implies that $E(Y|V) = V$. Assume, for example, a working model $m(V|\beta) = \beta_0 + \beta_1 V + \beta_2 V^2$ for $E(Y|V)$ is used by the permutation-based

method, and is fit with ordinary least squares regression. Asymptotically, the estimates of the coefficient vector β will tend to $(0, 1, 0)$. So for large sample sizes, the residuals $\{\epsilon_i\}$ corresponding to this model fit will be, approximately, $\epsilon_i = Y_i - V_i = (2A_i - 1)X_i$, where the subscript i denotes the value corresponding to the ith subject. Note, as mentioned above, we are assuming the framework of Section 3 in which each observation is i.i.d. Then for large sample sizes the empirical distribution of the residuals ϵ_i corresponding to treatment $(A_i = 1)$ will be right-skewed (having roughly the same distribution as X), while the empirical distribution of the residuals ϵ_i corresponding to the control arm $(A_i = 0)$ will be left-skewed (having roughly the same distribution as $-X$). Thus, the Wilcoxon rank sum test on the set of residuals ϵ_i will reject the null with probability tending to 1 as sample size goes to infinity. This implies Type I error for testing the null hypothesis (*) will tend to 1. We stress this occurs because the methods of Rosenbaum (2002) are designed for a different framework and different type of null hypothesis than considered here.

References

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.

Freedman, D. A. (2007a). On regression adjustments to experimental data. *Advances in Applied Mathematics (To Appear)* .

Freedman, D. A. (2007b). On regression adjustments to experiments with several treatments. *Annals of Applied Statistics (To Appear)* .

Freedman, D. A. (2007c). Randomization does not justify logistic regression. *Technical Report: http://www.stat.berkeley.edu/ ˜census/neylogit.pdf* .

Hardin, J. W. and Hilbe, J. M. (2007). *Generalized Linear Models and Extensions, 2nd Edition.* Stata Press, College Station.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. WileyInterscience, New York.

Huber, P. J. (1967). The behavior of of the maximum likelihood estimator under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist., Prob. 1* pages 221–233.

Jewell, N. P. (2004). *Statistics for Epidemiology*. Chapman and Hall/CRC, Boca Raton.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Wiley, New York, 2 edition.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22,** 719–748.

McCullagh, P. and Nelder, J. A. (1998). *Generalized Linear Models*. Chapman and Hall/CRC, Monographs on Statistics and Applied Probability 37, Boca Raton, Florida, 2nd edition.

Moore, K. L. and van der Laan, M. J. (2007). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 215. http://www.bepress.com/ucbbiostat/paper215* .

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych (in Polish)* **10,** 1–51.

R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, ISBN 3-900051-00-3, URL http://www.R-project.org/., Vienna, Austria.

Robins, J. M. (2002). Comment: Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17,** 309–321.

Robins, J. M. (2004). Optimal structural nested models of optimal sequential decisions. *Proceedings of the Second Seattle Symposium on Biostatistics, D. Y. Lin and P. Heagerty (Eds.)* pages 6–11.

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

Rosenblum, M. and van der Laan, M. (2007). Using regression to analyze randomized trials: Valid hypothesis tests despite incorrectly specified models. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 219. http://www.bepress.com/ucbbiostat/paper219* .

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2007). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine (To Appear)* .

van der Laan, M. J. and Hubbard, A. E. (2005). Quantile-function based null distribution in resampling based multiple testing. *Statistical Applications in Genetics and Molecular Biology (Technical Report available at http://www.bepress.com/ucbbiostat/paper198)* **5,**.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 213. http://www.bepress.com/ucbbiostat/paper213* .

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley and Sons.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge, UK.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2007). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics (To Appear)* .

Table 1 of this Web Appendix

Type I error for methods M0-M5, at various data generating distributions. Working Model used by methods M0, M4, M5 below is $logit(P(Y=1|A,V)) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV$. *Below,* W *is a standard normal random variable independent of* V, A, Y. *Involving* W *in the data generating distribution represents the situation, which will almost always occur in practice, where unmeasured variables affect the mean outcome. We consider sample sizes 400 and 200, respectively.*

	Type I Error When Data Generated by Logistic Regression Model logit$(P(Y\|A,V)) =$				
	V	$V^2/2$	$e^V/1.63$	$V+W$	$e^V/1.63+W$
Hypothesis Testing Methods					
			At Sample Size 400		
M0: Regression Based	0.05	0.05	0.05	0.05	0.05
M1: Intention-to-Treat	0.05	0.05	0.05	0.05	0.05
M2: C-M-H Test	0.04	0.04	0.04	0.04	0.04
M3: Permutation Based	0.05	0.05	0.05	0.05	0.05
M4: Targeted MLE	0.05	0.05	0.05	0.05	0.05
M5: Aug. Estimating Fn.	0.05	0.05	0.05	0.05	0.05
			At Sample Size 200		
M0: Regression Based	0.04	0.04	0.04	0.05	0.05
M1: Intention-to-Treat	0.05	0.05	0.05	0.05	0.05
M2: C-M-H Test	0.04	0.04	0.04	0.04	0.04
M3: Permutation Based	0.05	0.05	0.05	0.05	0.05
M4: Targeted MLE	0.05	0.05	0.05	0.05	0.05
M5: Aug. Estimating Fn.	0.05	0.05	0.05	0.05	0.05

Table 2 of this Web Appendix

Type I error for methods M0, M3, M4, M5 at a single data generating distribution, for various working models containing different numbers of baseline variables. The working model used by methods M0, M4, M5 is $m(A,V|\beta) = logit^{-1}(\beta_0 + \beta_0' A + \beta_1 V_1 + \ldots + \beta_j V_j)$, *for each* $j \in \{2,4,6,8,10\}$. *The working model used by method M3 is* $logit^{-1}(\beta_0 + \beta_1 V_1 + \ldots + \beta_j V_j)$, *for each* $j \in \{2,4,6,8,10\}$. *We consider sample sizes 400 and 200 respectively.*

	Type I Error When Working Model Used Contains j Baseline Variables:				
	$j=2$	$j=4$	$j=6$	$j=8$	$j=10$
Hypothesis Testing Methods					
			At Sample Size 400		
M0: Regression Based	0.05	0.05	0.05	0.05	0.05
M3: Permutation Based	0.05	0.05	0.05	0.05	0.05
M4: Targeted MLE	0.05	0.05	0.05	0.06	0.06
M5: Aug. Estimating Fn.	0.05	0.05	0.05	0.06	0.06
			At Sample Size 200		
M0: Regression Based	0.05	0.05	0.05	0.05	0.04
M3: Permutation Based	0.05	0.05	0.05	0.05	0.05
M4: Targeted MLE	0.05	0.06	0.06	0.06	0.06
M5: Aug. Estimating Fn.	0.05	0.06	0.06	0.06	0.06

Table 3 of this Web Appendix

Power when Working Model is Incorrectly Specified. Sample size is 200. Working Models used are defined above. The data generating distributions corresponding to each column are those described in Section 6–exactly the same as in Table 1 of the paper.

	Power When Data Generated by:		
	Logistic Regression Treatment Term Only: A	Logistic Regression Main Terms Only: A, V	Logistic Regression Main + Interaction Terms: A, V, AV
Hypothesis Testing Methods			
	Using Working Model 4 (*Misspecified Due to Wrong Link Function*)		
M0: Regression Based	0.85	0.73	0.93
M1: Intention-to-Treat	0.93	0.76	0.52
M2: C-M-H Test	0.91	0.79	0.49
M3: Permutation Based	0.92	0.79	0.65
M4: Targeted MLE	0.93	0.84	0.54
M5: Aug. Estimating Fn.	0.92	0.83	0.53
	Using Working Model 5 (*Misspecified Functional Form*)		
M0: Regression Based	0.85	0.65	0.56
M1: Intention-to-Treat	0.93	0.76	0.52
M2: C-M-H Test	0.91	0.73	0.48
M3: Permutation Based	0.81	0.67	0.48
M4: Targeted MLE	0.93	0.77	0.52
M5: Aug. Estimating Fn.	0.92	0.76	0.51
	Using Working Model 6 (*Misspecified Due to Measurement Error*)		
M0: Regression Based	0.85	0.63	0.48
M1: Intention-to-Treat	0.93	0.76	0.52
M2: C-M-H Test	0.91	0.73	0.48
M3: Permutation Based	0.81	0.64	0.44
M4: Targeted MLE	0.93	0.76	0.52
M5: Aug. Estimating Fn.	0.92	0.76	0.51

Table 4 of this Web Appendix
Power when Working Model is Incorrectly Specified. Sample size is 200. Working Model 3 is used, in which baseline variable V is replaced by a "noisy" version, to represent measurement error.

	Power When Data Generated Using: Logistic Regression Model for logit$(P(Y\|A,V))$ =:		
	$A + V^2/1.5$	$A + V^2/1.5 - AV^2/1.5$	A+sign(V) - A sign(V)
Hypothesis Testing Methods			
M0: Regression Based	0.75	0.16	0.73
M1: Intention-to-Treat	0.79	0.16	0.63
M2: C-M-H Test	0.75	0.14	0.60
M3: Permutation Based	0.52	0.12	0.66
M4: Targeted MLE	0.78	0.16	0.64
M5: Aug. Estimating Fn.	0.78	0.16	0.63

Table 5 of this Web Appendix
Power of Regression-Based Method M0, Based on Different Sets of Coefficients from the Working Model: $logit^{-1}(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV)$. Sample Size is 200.

	Power When Data Generated Using:		
	Logistic Regression Treatment Term Only: A	Logistic Regression Main Terms Only: A, V	Logistic Regression Main + Interaction Terms: A, V, AV
Hypothesis Testing Methods			
M0: Using β_1 only	0.86	0.80	0.83
M0: Using β_3 only	0.04	0.04	0.88
M0: Using β_1 and β_3	0.86	0.71	0.93

Table 6 of this Web Appendix
Power of Regression-Based Method M0, Based on Two Working Models. 1st Working Model: logit$^{-1}(\beta_0 + \beta_1 A + \beta_2 V)$; 2nd Working Model: logit$^{-1}(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV)$. Sample Size is 200.

	Power When Data Generated Using:		
	Logistic Regression Treatment Term Only: A	Logistic Regression Main Terms Only: A, V	Logistic Regression Main + Interaction Terms: A, V, AV
Hypothesis Testing Methods			
M0: 1st Working Model	0.92	0.82	0.50
M0: 2nd Working Model	0.86	0.71	0.93
M0: Both Working Models	0.87	0.74	0.88

Table 7 of this Web Appendix
Estimated Coefficients and Robust Standard Errors for Logistic Regression Model (A.21).

Coefficient:Term	Estimated Coefficient	Robust Standard Error
β_0 : (Intercept)	-1.97	0.30
$\beta_1 : A$	-0.22	0.59
$\beta_2 : V_1$	0.26	0.19
$\beta_3 : V_2$	-0.07	0.01
$\beta_4 : V_3$	0.76	0.18
$\beta_5 : V_4$	0.08	0.18
$\beta_6 : V_5$	0.65	0.21
$\beta_7 : AV_1$	-0.13	0.27
$\beta_8 : AV_2$	0.01	0.02
$\beta_9 : AV_3$	0.01	0.27
$\beta_{10} : AV_4$	0.08	0.25
$\beta_{11} : AV_5$	-0.16	0.30