EERI

Economics and Econometrics Research Institute

Analysis of the effects of adjusting for binary non-confounders in a logistic regression model after all true confounders have been accounted for: A simulation study

Ravan Moret and Andrew G. Chapple

EERI Research Paper Series No 05/2022

ISSN: 2031-4892

EERI
Economics and Econometrics Research Institute
Avenue Louise
1050 Brussels
Belgium

Tel: +32 2271 9482
Fax: +32 2271 9480
www.eeri.eu

Analysis of the effects of adjusting for binary non-confounders in a logistic regression model after all true confounders have been accounted for: A simulation study

Ravan Moret and Andrew G. Chapple

Biostatistics Program, School of Public Health, LSU Health Sciences Center, New Orleans, LA
Corresponding Author: Andrew Chapple - achapp@lsuhsc.edu
Both authors contributed equally to the work

Abstract

In observational studies, confounding variables that affect both the exposure and an outcome of interest are a general concern. It is well known that failure to control for confounding variables adequately can worsen inference on an exposure's effect on outcome. In this paper, we explore how exposure effect inference changes when non-confounding covariates are added to the assumed logistic regression model, after the set of all true confounders are included. This is done via an exhaustive simulation study with thousands of randomly generated scenarios to make general statements about over-adjusting in logistic regression. Our results show that in general, adding non-confounders to the regression model decreases the mean squared error for non-null exposure effects. The probability of both type I and type II errors also decrease with addition of more covariates given that all true confounders are controlled for.

1 Introduction

In July 2020, Williamson et. al published a paper analyzing factors associated with death from COVID-19 [18]. The study consisted of over 17 million patients' information from the OpenSAFELY platform. From the study they found COVID-19 related deaths to be associated with being male, older age, severe asthma, diabetes, and a series of other medical conditions. To that date, the study had been the largest of its kind in relation to COVID-19 information. Aside from patient age and gender, which had previously been determined as strong factors related to COVID-19 related deaths, the study included 21 additional factors in its analysis including: age, sex, BMI, smoking status, ethnicity, IMD quintile, blood pressure, respiratory disease (not including asthma), asthma, chronic heart disease, diabetes, cancer (non-hematological), hematological malignancy, reduced kidney function, liver disease, stroke/dementia, other neurological disease, organ transplant, asplenia, rheumatoid/lupus/psoriasis, and other immunosuppressive condition.

Though the paper was praised for its large study size, it gained a lot of criticism as well for "over-adjustment". The adjusted effects of smoking led to a lot of confusion, finding that smoking actually reduced the hazard of death compared to non-smokers (HR= .89, 95% CI = $.82 - .97$). This interpretation sparked much criticism and debate on the methods for modeling and the unexpected results were believed to be caused by over-adjusting in the model. Specifically, by including a large number of covariates the authors failed to account for the presence of confounders in the model and that by over-adjusting for the covariates they masked true effects of certain covariates.

The term over-adjustment has been defined in many ways, from Breslow (1982) who said, "Statistical adjustment by an excessive number of variables or parameters, uniformed by substantive knowledge. It can obscure a true effect or create an apparent effect when none exists" [2]. Or the later definition by Greenland, Pearl, and Robins (1998), "intermediate variables, if controlled in an analysis, would usually bias results towards the null, such control of an intermediate may be viewed as a form of over-adjustment" [6]. Ultimately over-adjustment can be viewed as any regression adjustment that results in either an increase in net bias or a decrease in precision. [14]

When planning an observational study it is important to account for any factor that may have an influence on the outcome or exposure. A confounder or confounding factor can be identified by three criteria: 1) the exposure of the study is determined to be associated with the confounder, 2) the outcome remains associated with the confounder even among the unexposed, and 3) the confounder itself is not an intermediate factor [8]. When deciding on a study design it is important to address any known confounders and decide which to control for. However, often it is not possible to accurately choose the true set of confounding variables to adjust for [17].

When there are many possible covariates present, total knowledge of all causal paths are not available, and other methods are used to determine which covariates to adjust for, such as the ”common cause method”. With this approach the researcher adjusts for all pre-exposure covariates with known common causes to the exposure and outcome of interest. However, if the set of confounding variables is not known, the researcher is likely to control for an incorrect set of confounders, which may introduce bias [3] [17].

Prior knowledge, or causal inference, allows the researcher to define exactly what effects they seek prior to the start of analysis, often accomplished with the use of directed acyclic graphs. Directed acyclic graphs (DAGs) are often used to express directional causal effects of one variable to another. DAGs have become increasingly popular in applied health research as they help researchers identify factors such as confounders and colliders that may introduce bias into a study. While several statistical methods exist to leverage DAGS, the most common adjustment for confounders is including them in multivariable regression models. The effects of omitting and adjusting for true confounders has been intensely studied, however the effect of adding non-confounders after all confounders have been included in a regression model has yet to be studied. Specifically, we investigate over adjustment, characterized by including more covariates than necessary to explain an exposure-outcome relationship [7], rather than overfitting which is characterized by poor prediction and misbehaved standard error [11] [5] [16].

In a study by Robinson and Jewell (1991), they explored the potential effects of adjusting for non-confounders, or over-adjusting, by the addition of a single non-confounder to a model containing only the outcome and exposure [13]. In the case of a classic linear regression model, Robinson and Jewell (1991) showed that the adjustment of non-confounders increased the precision of the model. However, given a logistic regression model, they showed that adjusting for non-confounders lead to either a decrease or no change in precision. We will extend the exploration of Robinson and Jewell (1991) by controlling for the set of true confounders and different numbers of additional non-confounders.

The remainder of the paper is outlined as follows. In section 2 we will discuss a motivating example and introduce equations and notations to explain model fitting. In section 3 we will explain the simulation design of our study including how random simulation scenarios were obtained, and how random datasets were generated from these scenarios. 1,000 random scenarios were generated under differing numbers of true confounding variables to explore general trends of over-adjustment, rather than picking a few scenarios and generalizing conclusions based on those. In section 4 we will examine the results of our simulations in terms of how adding unneeded covariates to logistic regression models affects exposure effect mean squared error and hypothesis testing. A discussion of our findings will be presented in section 5.

2 Motivating Examples and Methods

In this section we are going to describe over-adjustment via an example. We will then discuss over-adjustment in general and what we will explore in this manuscript.

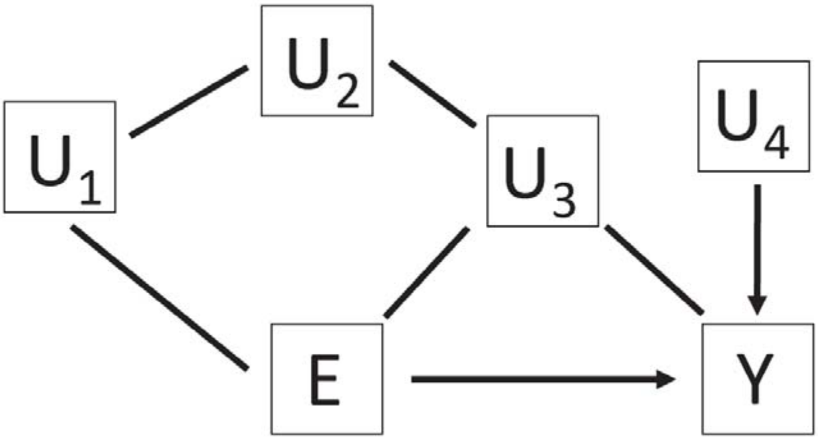

Figure 1: Causal Diagram: Here E is the exposure of interest, Y in the binary outcome, and U_1, U_2, U_3, U_4 are potential binary confounding variables for E. Only U_3 is a true confounder in this setting. U_4 is related to outcome, but not exposure. U_1 is related to exposure, but not outcome. U_2 is not directly related to either exposure or outcome.

Figure 1 displays an example of a true DAG in the relationship between several binary covariates, a binary exposure, and a binary outcome variable. Arrows indicate a causal effect for each pair of variables while straight lines indicate correlation between the variable pairs. Here U_1, U_2, U_3, and U_4 are measured potential binary confounders where U_3 is the only true confounder, affecting both the exposure (E) and outcome (Y). Y is determined from some true logistic regression model between E and U_3. Figure 1 could represent the logistic regression on outcome of:

$$logit(P[Y_i = 1|E_i, U_{3i}]) = \beta_0^{\text{true}} + \beta_E^{\text{true}} E_i + \beta_3^{\text{true}} U_{3i} + \beta_4^{\text{true}} U_{4i}$$

U_1 and U_2 do not have a direct effect on Y. U_1 is not a confounder but does affect E, and U_4 is also not a confounder but does affect Y, although they are correlated with U_3. U_2 is simply correlated to U_1 and U_3 but has no direct effect on E or Y. Returning to the motivating example involving COVID-19 [18], E would denote smoking status, and U_1, U_2, U_3, U_4 could represent some of the other binary covariates that were adjusted for (i.e. sex, ethnicity, respiratory disease (not including asthma), asthma, chronic heart disease, diabetes, cancer (non-hematological), hematological malignancy, reduced kidney function, liver disease, stroke/dementia, other neurological disease, organ transplant, asplenia, rheumatoid/lupus/psoriasis, and other immunosuppressive condition). Our primary concern of this paper is to analyze what happens when non-confounders (i.e. U_1, U_2, U_4) are added to the true logistic regression model in addition to U_3, in terms of estimating β_E^{true}. For example, if we fit a logistic regression model with U_{1i}, U_{2i}, U_{3i}, U_{4i}, E_i:

$$logit(P[Y_i = 1|U_{1i}, U_{2i}, U_{3i}, U_{4i}, E_i, \boldsymbol{\beta}]) = \beta_0 + \beta_E E_i + \beta_1 U_{1i} + \beta_2 U_{2i} + \beta_3 U_{3i} + \beta_4 U_{4i},$$

the addition of the variables U_1, U_2, U_4 would be considered over-adjustment and possibly may hurt estimation of β_E^{true}. Here U_3 was correctly adjusted for as it is a confounding variable in figure 1. This would explore whether criticisms of over-adjustment as it related to the result showing smoking being protective of COVID-19 death were valid [18]. In this paper we investigate if the effects of over-adjustment improve if one or two of U_1, U_2, U_4 are added to the model, after U_3 is appropriately controlled for. The goal of this paper is to investigate whether adding unnecessary confounders hurts the mean squared error and hypothesis testing for E → Y relationship, i.e. the estimation of β_E^{true} using $\hat{\beta}_E$ from the following models:

1. Where only E_i, U_{3i} are included in the assumed logistic regression model. This is the model where only confounders are included.

2. Where E_i, U_{3i} are included and one of U_1, U_2, U_4 in the assumed logistic regression model. The inclusion of one of these covariates would be over-adjustment as they do not effect both the exposure of interest and outcome.

3. Where E_i, U_{3i} are included and two of U_1, U_2, U_4 in the assumed logistic regression model.

4. Where E_i, U_{3i} are included and U_1, U_2, U_4 are included as well in the assumed logistic regression model.

In our simulation study we fit logistic regression models analogous to models 1-4 with $p = 14$ binary covariates under consideration and 3 and 6 true confounding variables for the $E \to Y$ effect.

We investigated whether the change in estimation and hypothesis testing of $H_0 : \beta_E = 0$ worsens for the enumerated examples above. In general, if there are p potential binary covariates to adjust for, we begin by including all true confounders in the model then sequentially adding non-confounders to our multivariable logistic regression model. Without loss of generality, assume that the p_c true confounders are denoted by $U_1, ..., U_{p_c}$. An assumed logistic regression model with no additional non-confounders included can be written in general as

$$logit(P[Y = 1|E_i, \boldsymbol{U}_i, \boldsymbol{\beta}]) = \beta_0 + \beta_E E_i + \sum_{k=1}^{p_c} \beta_k U_{ki}. \tag{1}$$

In this equation we consider all the true confounders in the model. In each simulation replication, p_{ex} extra non-confounders are added to the regression model (1). We investigate the effect of varying the number of non-confounders in the model (i.e. $p_{ex} = 1, ..., p - p_c$) via the following logistic regression model

$$logit(P[Y = 1|E_i, \boldsymbol{U}_i, \boldsymbol{\beta}]) = \beta_0 + \beta_E E_i + \sum_{k=1}^{p_c} \beta_k U_{ki} + \sum_{k=p_c+1}^{p_c+p_{ex}} \beta_k U_{ki} \tag{2}$$

Our goal is to determine if inference on β_E^{true} worsens as p_{ex} increases in (2).

3 Simulation Study Design

To analyze the effects of adding non-confounding variables to a logistic regression model we performed four sets of simulations across 1,000 randomly generated simulation scenarios. For each of these scenarios, 1,000 randomly generated data sets were obtained and operating characteristics were explored for various over-adjusted models. For each simulated dataset, we sequentially added each additional non-confounding covariate to the model and analyzed the affects of over-adjustment. We explored this in the setting of $p = 14$ binary covariates and one exposure of interest. We investigated settings where there were $p_c = 3$ and $p_c = 6$ true confounding variables for that exposure.

We will begin by introducing our true model and its components, which was used to generate dataset replications for a given simulation truth. Next we will discuss the steps taken to generate the parameters for a given simulation scenario. Additional steps were taken to correct for marginal separation and lack of variability issues, which is also described in this section. Summary statistics related to the randomly generated scenarios will be displayed graphically to give an idea of the simulation truths explored. For all simulation scenarios, the true logistic regression model can be written as:

$$logit(P[Y_i = 1|U_{1i}, ..., U_{pi}, E_i, \boldsymbol{\beta}^{\textbf{true}}]) = \beta_0^{\text{true}} + \beta_E^{\text{true}} E_i + \sum_{k=1}^{p} \beta_k^{\text{true}} U_{ki}. \tag{3}$$

Here β_0^{true} is our linear intercept, and β_E^{true} is our regression coefficient for the effect of our exposure variable E_i on outcome Y. β_k^{true} is the effect of a potential confounder U_k on Y. During scenario generation, some β_k^{true} were set to 0, indicating the corresponding U_k does not have an effect on our coutcome Y. Similarly, some correlations between E and U_k were set to 0 indicating that U_k is not related to exposure. We will examine the effects of adding additional non-confounders where $\beta_k^{\text{true}} = 0$ or $cor(E, U_k) = 0$, i.e. U_k is not a confounder, by sequentially adding non-confounders and examining regression estimates $\hat{\beta}_E$ under each assumed logistic regression model.

The multivariate relationship between $(E_i, U_{1i}, ...U_{ki})$ was characterized by a continuous $(p + 1)$ continuous latent vector $\boldsymbol{Z_i}$ similar to an approach first discussed by Albert and Chib for Bayesian analyses, and closely followed a multivariate method described by Papathomas [1] [10]. We assumed that $\boldsymbol{Z_i}$ comes from a multivariate normal distribution with mean $\Phi(\boldsymbol{\pi}^{\text{true}})$ and covariance matrix Σ^{true}. Here $\boldsymbol{\pi}^{\text{true}}$ is defined as a vector of the marginal probability of each binary covariate and the exposure. In our generation, some of the entries of the covariance matrix Σ^{true} will be set to zero, indicating Z_{ji} and Z_{ki} are not correlated. For each simulated observation, we generated $\boldsymbol{Z}_i \sim MVN(\Phi(\boldsymbol{\pi}^{\text{true}}), \Sigma^{\text{true}})$ and set $U_{ki} = 1$ if $Z_{ki} > 0$, (i.e. $U_{ki} = I[Z_{ki} > 0]$) which characterized the multivariate relationship of E_i and $\boldsymbol{U}_i$ through $\boldsymbol{\pi}^{\text{true}}$ and Σ^{true}. After obtaining $\boldsymbol{U}_i$ and E_i for each simulated observation, we generated Y_i from a Bernoulli distribution with probabilities derived from (4). This concludes how we generated data for each observation within a given scenario, which was done 1,000 times in all simulation scenarios for various sample sizes.

To analyze the effects of adding non-confounders to the logistic regression model, we took a general approach by randomly generating parameter settings, rather than investigating results over a few chosen simulation settings. We explored settings with $p = 14$ binary covariates and one exposure variable of interest. We performed two sets of simulations, one with $p_c = 3$ true confounding variables, and one with $p_c = 6$ true confounding variables. This allowed addition of up to 11 and 8 non-confounding variables, respectively. The goal of this study is to determine whether increasing the number of non-confounding variables adjusted for increases mean square error (MSE) or the average z-statistic for testing no exposure effect (i.e. increases type I error rates or power), as these were contentions in criticisms of Williamson et al (2020) [18]. For $p_c = 3, p_c = 6$, we randomly generated 1,000 random values of $\boldsymbol{\pi}^{\text{true}}$, Σ^{true}, and $\boldsymbol{\beta}^{\text{true}}$. These constituted 1,000 random simulation truths, which were used to generate 1,000 random datasets for each simulation truth.

We first generated a random vector $\boldsymbol{\pi}^{\text{true}}$ (the marginal probability of each covariate) from a uniform distribution with minimum and maximum values of 0.1 and 0.9, respectively. We chose these boundaries to avoid binary covariates that marginally occurred very frequently or rarely. We generated Σ^{true} from a Wishart distribution, $\Sigma^{\text{true}} \sim W_p(\Sigma_0)$, since it gives symmetric positive-definite matrices which are required for the multivariate normal distribution on the latent vector $\boldsymbol{Z_i}$ [19]. Σ_0 was constructed as a diagonal matrix with entries drawn from a uniform (0, .05) distribution. We instituted regulatory steps where low matrix values in the first row and first column, $(|\Sigma_{Ek}^{\text{true}}| < 0.1$ and $|\Sigma_{Ek}^{\text{true}}| < 0.1)$, which we considered to be low or negligible correlations, were set to 0. The resulting Σ^{true} matrix, under these restrictions, was tested for positive definiteness, since adding sparsity may violate positive definiteness. We repeatedly generated the matrix Σ^{true} repeatedly until positive definiteness was achieved.

We generated all $\boldsymbol{\beta}^{\textbf{true}}$ values, $(\beta_0^{\text{true}}, \beta_E^{\text{true}}, \beta_1^{\text{true}}, ..., \beta_p^{\text{true}})$, from a standard normal distribution. We then determined how many values of $\boldsymbol{\beta}^{\text{true}}$ should be set to 0, indicating no relationship between those covariates and outcome, by first randomly drawing a value δ from a discrete uniform distribution on the set $p_c, ..., p$. Afterwards, δ randomly chosen values of $\beta_1^{\text{true}}, ..., \beta_p^{\text{true}}$ were set to 0. If both β_k^{true} and $\Sigma_{Ek}^{\text{true}}$ values are not equal to zero, this indicates the covariate U_k us is a true confounder for our exposure E. Our number of true confounders, p_c, is defined as

$$p_c = \sum_{k=1}^{p} I[\beta_k^{\text{true}} \neq 0 \ \ \& \ \ \Sigma_{Ek} \neq 0].$$

We repeated this process until we had $p_c = 3$ and $p_c = 6$, respectively. We performed an additional check to address issues with non-variability of binary random variables and whether marginal separation was present. Marginal separation is defined by the existence of some cutoff c where all $Y_i = 1$ when a covariate is greater than c and all $Y_i = 0$ otherwise (or vice versa) [9]. In general, separation can occur when any linear combination of covariates can lead to a situation where some cutoff value c discriminates between $Y_i = 1$ and $Y_i = 0$ using that linear combination of covariates. Here, we only checked for marginal separation, i.e. whether each covariate individually leads to separation - as checking for separation caused by linear combination was not possible. This allowed us to study cases where apparent marginal separation was not present and there was no need for Firth's correction [4].

A checking system was put in place to remove any generated parameter settings that caused separation and no variability. Scenarios that resulted in no variability in any entry of $(E, U_1, ..., U_{14}, Y)$ were removed. For 1,000 randomly generated datasets with a given simulation truth, we created a contingency table for the generated Y responses and each column of the data matrix $(E, U_1, ..., U_p)$ entry. If any cell of the 2x2 table contained a 0, indicating marginal separation, that simulation scenario was discarded. We did this because it's known that separation leads to unreasonable effect estimates [9]. This separation check was conducted under $n = 1,000$ observations. It was not possible to generate non-pathological scenarios under smaller sample sizes (i.e. $n = 200$) with this many covariates, so we restrict this paper to exploration of over-adjustment effects in larger populations of $n = 1,000$ and $n = 10,000$.

Collectively, this process resulted in $1,000$ randomly generated scenarios for $p_c = 3$ and $p_c = 6$ true confounding variables. This resulted in 1,000 random Σ^{true} and $\boldsymbol{\beta}^{\text{true}}$ setting. These parameter settings were used for both sample sizes of one and ten thousand. Figure 2 displays several empirical densities of parameters related to each scenario, including the average proportion of events for $n = 1,000$, the average magnitude of the β_k^{true} values for confounding variables, the average magnitude of the $\Sigma_{Ek}^{\text{true}}$ values for confounding variables, and a histogram displaying the randomly generated number of non-confounders of each type for $p_c = 3$ true confounders. These three types of non-confounders are those related to outcome only (i.e. U_4 in Figure 1), those related to exposure only (i.e. U_1 in Figure 1), and those related to neither (i.e. U_2 in Figure 1).

The top left of figure 2 displays the density of the randomly generated $E(Y|E, \boldsymbol{U})$ across the 1,000 randomly generated scenarios for $p_c = 3, 6$ true confounders. Most of the density is concentrated around .5, which is an artifact of $\beta_0^{\text{true}} \sim N(0, 1)$ for each random scenario. A summary of $E(Y|E, \boldsymbol{U})$ is (min, .25 quantile, mean, .75 quantile, max) = (.07, .35, .50, .64. .94) for $p_c = 3$ and $(.06, .36, .51, .66, .94)$ for $p_c = 6$. The top right of figure 2 displays the densities of the average $|\beta_k^{\text{true}}|$ values for confounding variables across the 1,000 randomly generated scenarios. There is a greater discrepancy between $p_c = 3$ and $p_c = 6$ for these average magnitudes than for $E(Y|E, \boldsymbol{U})$ with $p_c = 6$ having a higher density around .8. The summaries of the values of $|\beta_k^{\text{true}}|$ for the confounding variables was $(.13, .62, .77.91, 1.81)$ for $p_c = 3$ and $(.24, .65, .77, .89, 1.50)$ for $p_c = 6$.

The densities of the 1,000 randomly generated average $|\Sigma_{Ek}^{\text{true}}|$ for true confounders is shown in the bottom left of figure 2. There is a clear rightward shift of the density for $p_c = 6$ confounders. The summaries for these true values of $|\Sigma_{Ek}^{\text{true}}|$ were rather close being $(.11, .14, .17, .18, .32)$ for $p_c = 3$ and $(.12, .16, .18, .20, .31)$ for $p_c = 6$. The reader should note that all 1,000 randomly generated $|\Sigma_{Ek}^{\text{true}}|$ for confounding variables are above .10 which reflects the way that random scenarios were generated, i.e. setting all $|\Sigma_{Ek}^{\text{true}}| < .10$ to 0 - making it a non-confounder. Lastly, the distribution on the number of each type of non-confounding variable is shown in the bottom right of figure 2 for $p_c = 3$. This figure shows that about 38 % of randomly generated simulation scenarios have ≥ 7 non-confounders which are only related to outcome, otherwise this distribution is even across 1-6 outcome-only non-confounders. About 30% of the scenarios had 1 exposure-only non-confounder, while 27% had 2 exposure-only non-confounding variables. The number of non-confounding variables that weren't related to exposure or outcome was $9 - 18\%$ across all configurations.

4 Simulation Results

In this section we will discuss the results from the simulations study and examine the effects of adding true non-confounders to the logistic regression model. Since there are $p = 14$ covariates excluding the exposure E, there are $14 - 3 = 11$ possible non-confounders that can be added when $p_c = 3$ and $14 - 6 = 8$ possible non-confounders that can be added when $p_c = 3$. We investigate operating characteristics related to estimation accuracy and hypothesis testing for different numbers of non-confounders added to the model, i.e. $p_{ex} = 0, ..., 14 - p_c$. For each of the $1,000$ randomly generated simulation scenarios, we randomly generate $1,000$ datasets and compute $\hat{\beta}_E$ and $\hat{SE}\left(\hat{\beta}_E\right)$ for each choice of p_{ex} and each random dataset. For the bth randomly generated dataset, we compute the squared error as

$$MSE^b = \left(\hat{\beta}_E - \beta_E^{\text{true}}\right)^2,$$

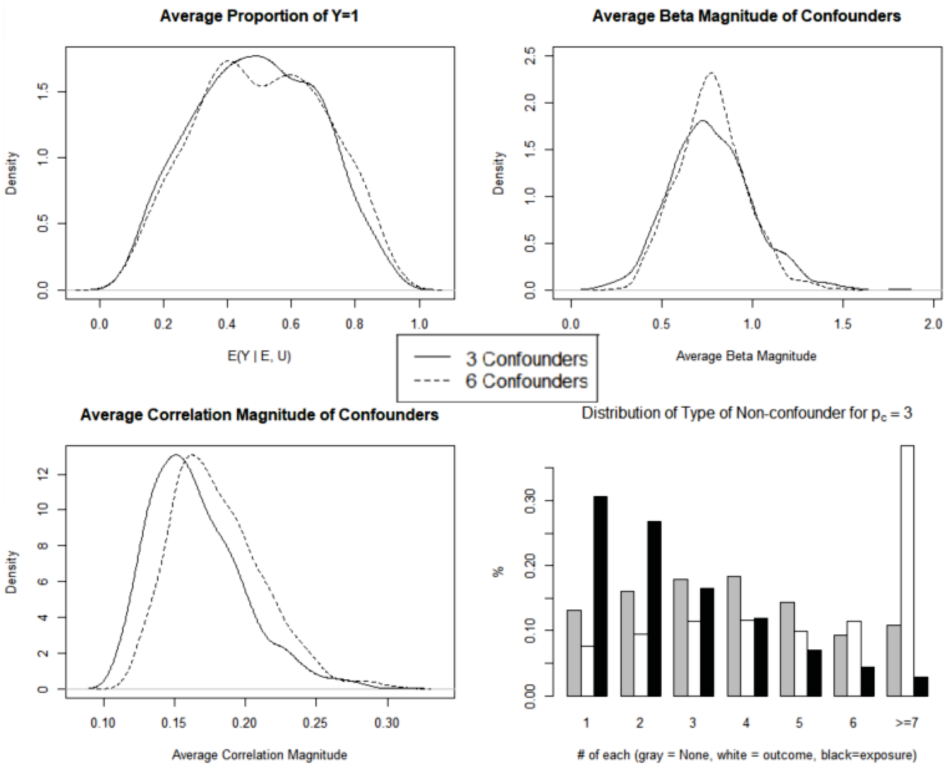

Figure 2: Summary of randomly generated simulation scenarios. (top left) The density of the average proportion of $Y = 1$ across the random scenarios, (top right) Density of average magnitude of β_k^{true} for confounding variables for a given simulation scenario, (bottom left) Density of average magnitude of $\Sigma_{Ek}^{\text{true}}$ for confounding variables for a given simulation scenario, (bottom right) Distribution of the different types of non-confounders across the randomly generated scenarios for $p_c = 3$.

for each choice of p_{ex}. Then we average over 1,000 random datasets to obtain the mean squared error (MSE) for a given scenario. According to the theory posited by Robinson, with very large sample sizes the bias will go down with over-adjustment while the standard error will go up [13] - but ultimately, they believe the MSE will go down because the bias will go down faster than standard error will go up.

We also look at two versions of the z-statistic for testing $H_0 : \beta_E = 0$, i.e. that E_i has no effect on Y_i. For this hypothesis test, we look at one z-statistic when the null is true ($\beta_E^{\text{true}} = 0$) and one under the alternative where the null is false ($|\beta_E^{\text{true}}| > 0.1$). To do this we performed an additional set of each sample size and p_c and set $\beta_E^{\text{true}} = 0$. Within a given simulation scenario, for the bth randomly generated dataset, we compute

$$Z - statistic^b = \frac{\left|\hat{\beta}_E\right|}{\hat{SE}(\hat{\beta}_E)}$$

then we average over the 1,000 datasets to obtain Z_{Null} (when $\beta_E^{\text{true}} = 0$) and Z_{Alt} (when $|\beta_E^{\text{true}}| > .1$). This process is done for each of the $1,000$ randomly generated scenarios for $p_c = 3, 6$ and both sample sizes. We display the average values of MSE under the null and alternative and the average values of $|Z_{\text{Null}}|, |Z_{\text{Alt}}|$ across the randomly generated scenarios in figure 3.

The top 4 figures display average trends in MSE and the magnitude of the z-statistic under cases where $\beta_E^{\text{true}} \neq 0$. For the z-statistic magnitude, this is only computed when $|\beta_E^{\text{true}}| > .1$ to avoid cases where $\beta_E^{\text{true}} \approx 0$. We see that for the top row, where $n = 1,000$ and $p_c = 3$, the MSE drops from .081 with $p_{ex} = 0$, i.e. no non-confounders added, to .071 for $p_{ex} = 6$ which then levels off. For p_6 there was a similar pattern, with average MSE dropping from .078 with $p_{ex} = 0$ to .071 for $p_{ex} = 4$, which is maintained. For each choice of p_{ex} from 1 to 8, the simulations with $p_c = 6$ true confounders had lower average MSE than for $p_c = 3$ true confounders. For $n = 10,000$, there is a monotone decreasing trend in average MSE as p_{ex} increases for both $p_c = 3, 6$. The decrease in MSE was bigger for $n = 10,000$ (.03) compared to $n = 1,000$ (.01) as p_{ex} increases, suggesting that the estimation benefit for over-adjustment is larger for larger sample sizes.. Similar to $n = 1,000$, the average MSE is lower for $p_c = 6$ than $p_c = 3$ when $n = 10,000$. In cases where $|\beta_E^{\text{true}}| > .1$ the average magnitude of

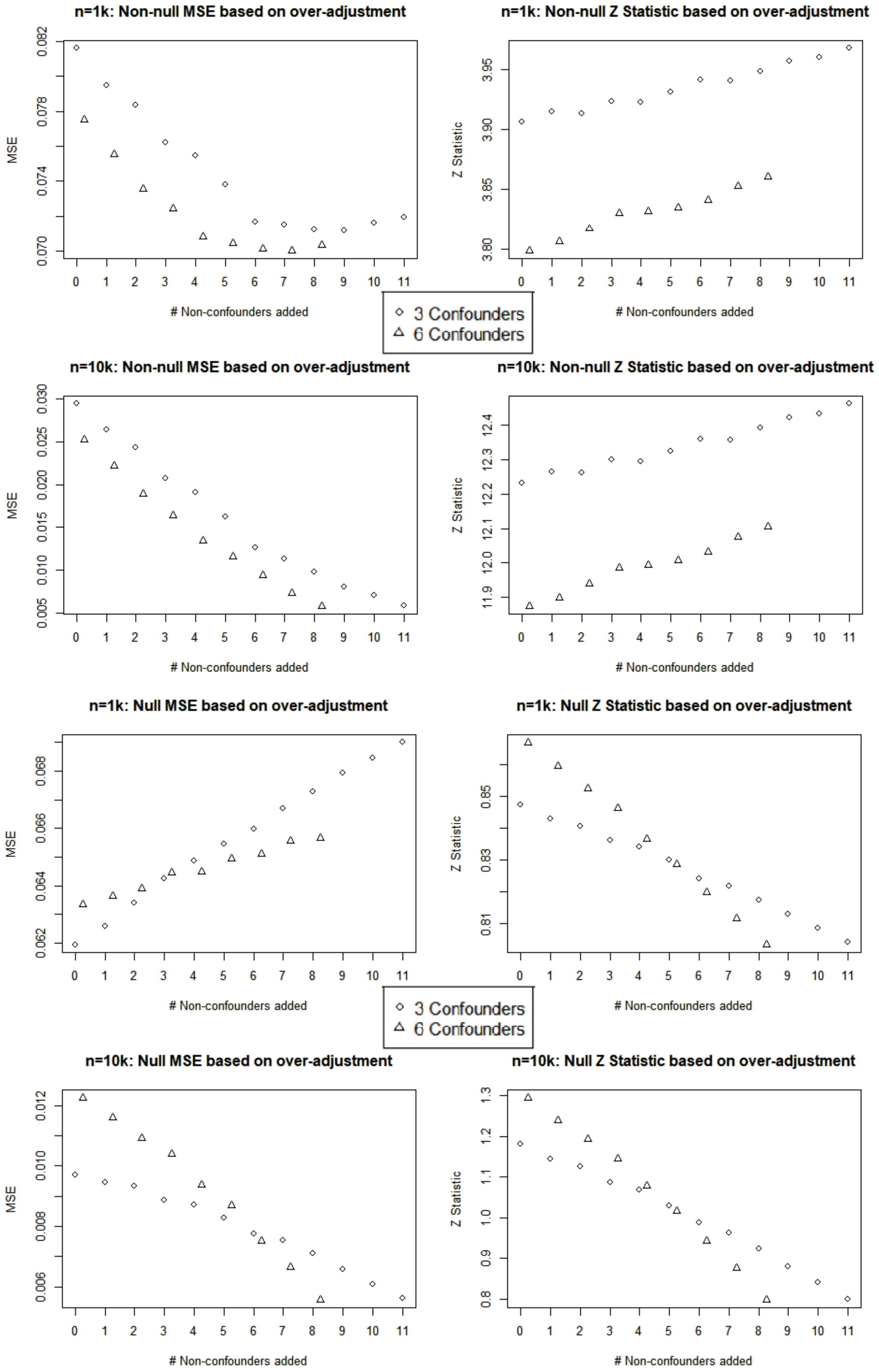

Figure 3: Simulation Results: Average MSE and magnitude of the Z-statistic testing $H_0 : \beta_E = 0$ across the 1,000 randomly generated scenarios. This is shown for $p_c = 3, 6,\ n = 1,000, n = 10,000$ and simulations were $\beta_E^{\text{true}} \neq 0$ (Non-null) and $\beta_E^{\text{true}} = 0$ (Null).

the z-statistic increases as p_{ex} increases for both $n = 1,000, 10k$. Interestingly, the $|Z_{\text{Alt}}|$ values are higher for $p_c = 3$ true confounders compared to $p_c = 6$ true confounders, but in general the average test statistics are above 3.8 for $n = 1,000$ and 11.8 for $n = 10,000$ suggesting a high likelihood of rejecting the null hypothesis in general based on a Wald test critical value of 1.96.

These trends for the null cases, where $\beta_E^{\text{true}} = 0$, are shown in the bottom four graphs of figure 3. For $n = 1,000$, we see a reversal of the trend when $\beta_E^{\text{true}} \neq 0$ - that average MSE actually increases as p_{ex} increases. This increase is smaller for $p_c = 6$ when adding additional covariates.

For $n = 10,000$ we see a downward trend in MSE for both $p_c = 3, 6$ which decreases faster with $p_c = 6$. For smaller p_{ex} the average MSE was higher for $p_c = 6$ which is reversed when $p_{ex} \geq 6$. We see a similar crossing trend for $|Z_{\text{Null}}|$ for both sample sizes, with average $p_c = 6$ values being higher (lower) than $p_c = 3$ for $p_{ex} < 5$ ($p_{ex} \geq 5$). The average $|Z_{\text{Null}}|$ values decrease as more non-confounders (i.e. p_{ex}) are added and are below 1.3 for $n = 10,000$ and .9 for $n = 1,000$. All of the average $|Z_{\text{Null}}|$ values are lower than the Wald test critical value of 1.96, indicating that the null hypothesis will most likely be upheld.

These results collectively suggest that estimation of β_E improves as p_{ex} increases except with a moderate sample size and null exposure effect. Hypothesis testing becomes more conservative as p_{ex} increases when there is a null exposure effect, which counters over-adjustment criticisms about the paper by Williamson et al (2020) [18]. The $|Z_{\text{Alt}}|$ values also increase as p_{ex} increases, suggesting an improvement in the power as more unneeded non-confounders are added to the logistic regression.

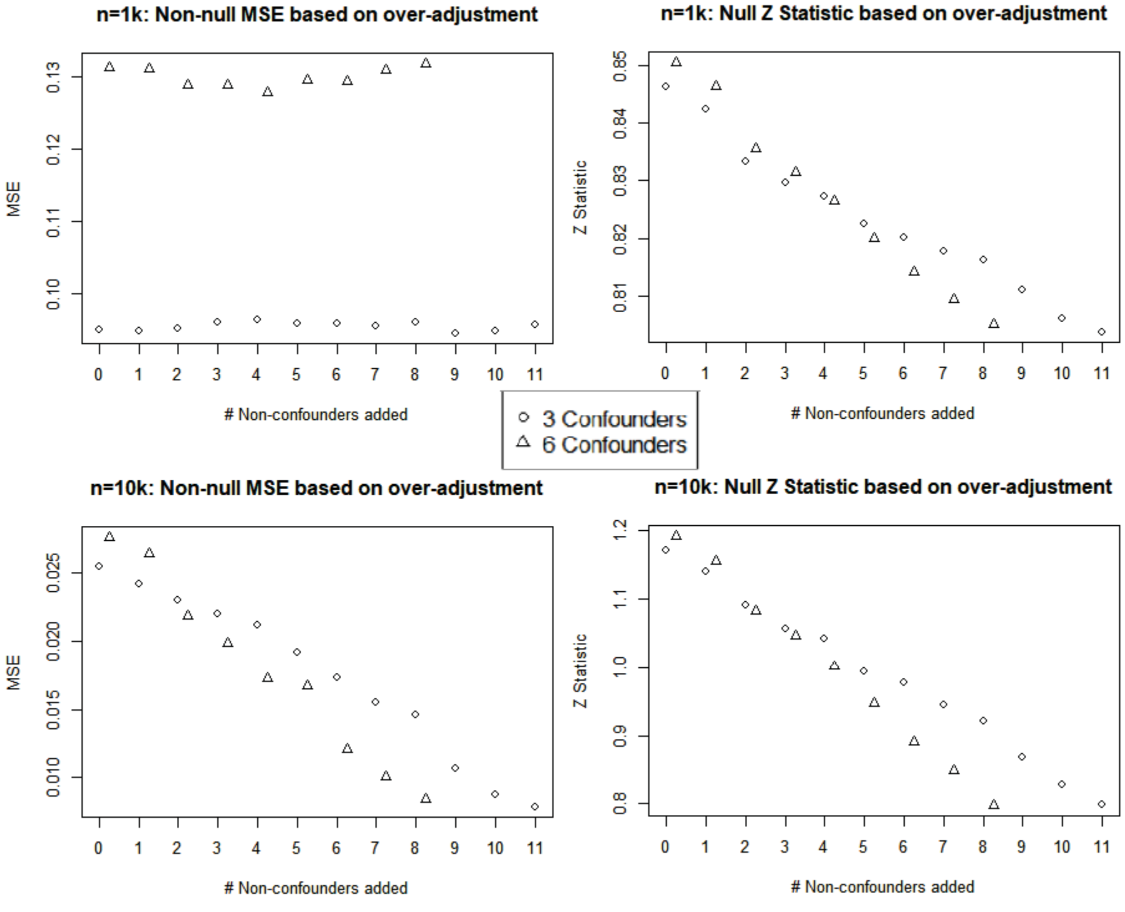

Figure 4: Operating characteristics in rare events, when $E(Y|E, \boldsymbol{U} < .2)$ or $E(Y|E, \boldsymbol{U} > .8)$.

Figure 4 displays the non-null MSE values and $|Z_{\text{Null}}|$ values for cases where $E(Y|E, \boldsymbol{U}) < .2$ or $E(Y|E, \boldsymbol{U}) > .8$, which are rare and common events. For $n = 1,000$, the non-null MSE remains mostly constant as p_{ex} increases for both $p_c = 3, 6$, whereas the MSE decreases with p_{ex} for $n = 10,000$. For smaller p_{ex}, MSE is actually higher for $p_c = 6$ which is then flipped.

Similar to figure 3, average $|Z_{\text{Null}}|$ decreases as p_{ex} increases - suggesting again that criticisms of over-adjustment in Williams et al (2020) may not be appropriate [18]. These values were below .9 for $n = 1,000$ and 1.2 for $n = 10,000$ which are below the 1.96 critical value. Though not shown here, there was a slight upward trend in $|Z_{\text{Alt}}|$ for $p_c = 6$ for both sample sizes, but not much of a change for $p_c = 3$. All of these values were above 9, indicating a high likelihood that the hypothesis test of no exposure effect would be rejected.

$n = 1,000$	Neither Outcome/exposure		Outcome Only		exposure Only	
p_{ex}	MSE_{Alt}	$\lvert Z_{\text{Null}}\rvert$	MSE_{Alt}	$\lvert Z_{\text{Null}}\rvert$	MSE_{Alt}	$\lvert Z_{\text{Null}}\rvert$
0	0.078 (0.131)	0.838 (0.090)	0.078 (0.132)	0.839 (0.091)	0.076 (0.129)	0.833 (0.082)
1	0.075 (0.128)	0.829 (0.076)	0.076 (0.129)	0.832 (0.081)	0.074 (0.126)	0.825 (0.070)
2	0.073 (0.126)	0.824 (0.068)	0.074 (0.127)	0.827 (0.073)	0.073 (0.125)	0.820 (0.061)
3	0.072 (0.125)	0.819 (0.058)	0.073 (0.126)	0.823 (0.067)	0.072 (0.124)	0.816 (0.054)
4	0.072 (0.124)	0.815 (0.050)	0.072 (0.125)	0.820 (0.060)	0.072 (0.124)	0.814 (0.048)
5	0.072 (0.124)	0.812 (0.044)	0.072 (0.124)	0.817 (0.054)	0.072 (0.124)	0.811 (0.043)
6	0.072 (0.124)	0.810 (0.040)	0.071 (0.124)	0.814 (0.048)	0.072 (0.124)	0.810 (0.039)
7	0.072 (0.124)	0.808 (0.035)	0.072 (0.124)	0.812 (0.043)	0.072 (0.124)	0.810 (0.039)
8	0.072 (0.124)	0.807 (0.031)	0.072 (0.124)	0.810 (0.040)	0.072 (0.124)	0.807 (0.032)
9	0.072 (0.124)	0.805 (0.027)	0.072 (0.124)	0.807 (0.033)	0.072 (0.124)	0.806 (0.027)
10	0.072 (0.124)	0.809 (0.038)	0.072 (0.125)	0.804 (0.022)	0.072 (0.125)	0.804 (0.020)
11	0.072 (0.125)	0.804 (0.020)	NA (NA)	NA (NA)	NA (NA)	NA (NA)
$n = 10,000$	Neither Outcome/exposure		Outcome Only		exposure Only	
p_{ex}	MSE_{Alt}	$\lvert Z_{\text{Null}}\rvert$	MSE_{Alt}	$\lvert Z_{\text{Null}}\rvert$	MSE_{Alt}	$\lvert Z_{\text{Null}}\rvert$
0	0.023 (0.061)	0.839 (0.094)	0.024 (0.062)	0.839 (0.092)	0.020 (0.053)	0.833 (0.085)
1	0.017 (0.047)	0.829 (0.078)	0.019 (0.051)	0.832 (0.082)	0.015 (0.04)	0.825 (0.070)
2	0.014 (0.038)	0.824 (0.068)	0.016 (0.043)	0.827 (0.074)	0.012 (0.032)	0.820 (0.061)
3	0.011 (0.028)	0.819 (0.058)	0.013 (0.036)	0.824 (0.068)	0.010 (0.024)	0.816 (0.054)
4	0.009 (0.020)	0.815 (0.050)	0.012 (0.030)	0.820 (0.064)	0.009 (0.019)	0.814 (0.048)
5	0.008 (0.015)	0.812 (0.044)	0.010 (0.023)	0.817 (0.056)	0.008 (0.015)	0.811 (0.043)
6	0.008 (0.012)	0.810 (0.040)	0.009 (0.017)	0.815 (0.052)	0.008 (0.012)	0.810 (0.039)
7	0.007 (0.010)	0.808 (0.035)	0.008 (0.014)	0.812 (0.045)	0.007 (0.011)	0.810 (0.039)
8	0.007 (0.008)	0.807 (0.031)	0.007 (0.011)	0.81 (0.041)	0.007 (0.008)	0.807 (0.032)
9	0.006 (0.007)	0.805 (0.027)	0.007 (0.008)	0.807 (0.033)	0.006 (0.007)	0.806 (0.027)
10	0.007 (0.009)	0.809 (0.038)	0.006 (0.005)	0.804 (0.022)	0.006 (0.005)	0.804 (0.020)
11	0.006 (0.005)	0.804 (0.020)	NA (NA)	NA (NA)	NA (NA)	NA (NA)

Table 1: Simulation results for models with $p_c = 3$ true confounders, $n = 1,000$ and $n = 10,000$, and various choices of p_{ex}. The mean operating characteristics and standard deviation were calculated for intervals of the number of non-confounders added (p_{ex}). The Z-statistic for simulation results with $\beta_E^{\text{true}} = 0$ (Z_{Null}) and MSE when $\beta_E^{\text{true}} \neq 0$ (MSE_{Alt}) were calculated.

We have shown in general that non-null MSE decreases in large sample sizes as p_{ex} increases and that over-adjustment concerns about type I errors are unsubstantiated. In table 1, we investigate these two trends in terms of how adding different types of non-confounders are over-adjusted for. These types of non-confounders are those related to outcome only, exposure only, and neither outcome nor exposure. In this table, we also report the standard deviation of these operating characteristics across the 1,000 randomly generated scenarios. It should be noted that when we look at models that add non-confounders of each type, we could potentially have already added non-confounders that are related to outcome or neither. For example, models that have 2 extra non-confounders that are only related to exposure may also have a non-confounder related to outcome only included as well.

For MSE_{Alt} and $n = 1,000$, we see a similar decreasing followed by leveling off trend as p_{ex} increases for each of the 3 types of non-confounders. We see a monotone decreasing trend for $n = 10,000$ and the three different types of non-confounders. For $n = 1,000$, the standard error on the MSE_{Alt} values is mostly constant, with a slight increase when no additional non-confounders are adjusted for. For $n = 10,000$, the standard error goes down as p_{ex} increases, suggesting that the bias of the estimates $\hat{\beta}_E$ and their variability decreases with p_{ex} - which is an appealing feature. This holds for all three non-confounding types. We see a monotone decreasing trend in average $|Z_{\text{Null}}|$ values and their standard deviation for each of the three non-confounding types and both $n = 1,000, n = 10,000$. This decrease is faster for adding exposure only non-confounders than the other two types for both $n = 1,000, 10k$.

Table 2 displays these same operating characteristics for $p_c = 6$ true confounding variables for exposure. For each of the three non-confounding types, MSE_{Alt} decreases as p_{ex} increases, which is more striking for $n = 10,000$. Similar to $p_c = 3$, the standard error on MSE_{Alt} across the 1,000 randomly generated simulation truths decreases with $n = 10,000$ but was mostly constant for $n = 1,000$. Similar to $p_c = 3$, adding non-confounders of each type reduced $|Z_{\text{Null}}|$ thereby making

$n = 1,000$	Neither Outcome/exposure		Outcome Only		Exposure Only	
p_{ex}	MSE_{Alt}	$\|Z_{\text{Null}}\|$	MSE_{Alt}	$\|Z_{\text{Null}}\|$	MSE_{Alt}	$\|Z_{\text{Null}}\|$
0	0.074 (0.115)	0.848 (0.102)	0.076 (0.116)	0.860 (0.117)	0.073 (0.115)	0.844 (0.099)
1	0.071 (0.115)	0.831 (0.078)	0.073 (0.115)	0.849 (0.102)	0.071 (0.115)	0.830 (0.077)
2	0.071 (0.115)	0.822 (0.063)	0.072 (0.115)	0.837 (0.086)	0.071 (0.115)	0.822 (0.063)
3	0.070 (0.116)	0.816 (0.052)	0.071 (0.115)	0.828 (0.071)	0.070 (0.116)	0.816 (0.050)
4	0.070 (0.116)	0.812 (0.043)	0.070 (0.115)	0.821 (0.058)	0.070 (0.116)	0.811 (0.041)
5	0.070 (0.116)	0.808 (0.034)	0.070 (0.116)	0.815 (0.047)	0.070 (0.116)	0.810 (0.038)
6	0.070 (0.116)	0.806 (0.027)	0.070 (0.116)	0.809 (0.035)	0.070 (0.116)	0.803 (0.020)
7	0.070 (0.117)	0.803 (0.020)	0.070 (0.116)	0.804 (0.023)	NA (NA)	NA (NA)
$n = 10,000$	Neither Outcome/exposure		Outcome Only		exposure Only	
p_{ex}	MSE_{Alt}	$\|Z_{\text{Null}}\|$	MSE_{Alt}	$\|Z_{\text{Null}}\|$	MSE_{Alt}	$\|Z_{\text{Null}}\|$
0	0.018 (0.035)	0.848 (0.106)	0.023 (0.042)	0.861 (0.120)	0.017 (0.033)	0.845 (0.099)
1	0.013 (0.024)	0.831 (0.078)	0.018 (0.033)	0.849 (0.103)	0.012 (0.023)	0.830 (0.082)
2	0.010 (0.018)	0.822 (0.063)	0.014 (0.026)	0.837 (0.088)	0.010 (0.018)	0.822 (0.066)
3	0.009 (0.015)	0.816 (0.052)	0.011 (0.021)	0.828 (0.074)	0.009 (0.014)	0.816 (0.050)
4	0.008 (0.011)	0.812 (0.043)	0.010 (0.017)	0.821 (0.059)	0.008 (0.011)	0.811 (0.041)
5	0.007 (0.008)	0.808 (0.034)	0.008 (0.013)	0.815 (0.048)	0.007 (0.009)	0.810 (0.038)
6	0.006 (0.006)	0.806 (0.027)	0.007 (0.008)	0.809 (0.035)	0.006 (0.004)	0.803 (0.020)
7	0.006 (0.004)	0.803 (0.020)	0.006 (0.005)	0.804 (0.023)	NA (NA)	NA (NA)

Table 2: Simulation results for models with $p_c = 6$ true confounders, $n = 1,000$ and $n = 10,000$, and various choices of p_{ex}. The mean operating characteristics and standard deviation were calculated for intervals of the number of non-confounders added (p_{ex}). The Z-statistic for simulation results with $\beta_E^{\text{true}} = 0$ (Z_{Null}) and MSE when $\beta_E^{\text{true}} \neq 0$ (MSE_{Alt}) were calculated.

the adjusted exposure-outcome hypothesis test more conservative. This decrease was much slower for outcome-only related non-confounders. Standard errors of $|Z_{\text{Null}}|$ across the 1,000 randomly generated scenarios also decreased as the number of non-confounders increased.

5 Discussion

Ideally, in a logistic regression all confounders will be accounted and adjusted for. However, in many cases where there are a large number of potential confounders or little information is known of the exposure of interest, the possibility of adjusting for all confounders is not likely. In such an instance, scientists may attempt to adjust for potential confounders, this however may often lead to adjusting for false or non-confounders, a concept known as over-adjustment. While the concept of omitting and adjusting for confounders has been intensely discussed in studies concerning logistic regression, in this paper we analyzed the effects of over-adjustment with binary covariates when all true confounders are already adjusted for. We did this across 1,000 randomly generated simulation truths, investigating mean squared error and the average Z statistic magnitude under cases where the exposure does and does not have an effect on outcome. In general, adding non-confounders improved estimation of the true exposure effect - other than with a moderate sample size and null exposure effect or in non-null exposure effects. This trend was not evident for rare outcome events.

When there truly was no exposure-outcome relationship, the average magnitude of the z statistic testing no effect decreased as more non-confounders were added. This indicates that concerns about over-adjustment in relation to the smoking effect in the Williamson et al paper may be unwarranted [18]. This is not to argue that smoking does protect against COVID-19 death, but to say that this finding was likely not related to inclusion of too many additional covariates in their assumed logistic regression model.

We saw that when there was a true exposure-outcome relationship (i.e. null is not true), the average z-statistic increased in general. This trend was not as evident in rare outcome events. We showed that these trends held in general when adding non-confounders related only to exposure, only to outcome, or neither.

We did not explore how addition of continuous non-confounding variables affected results in terms of overfitting for two reasons. The first reason is that many studies coming out today focus on binary predictors (disease related/whatnot). The second reason was practicality, If we were to consider continuous covariates, we'd need to consider (1) Different distributions that generated the continuous variables (gamma, normal, etc) and (2) Different functional relationships between those covariates, the exposure, and the outcome (linear, quadratic, piecewise, other). There could be more exploration in this area in future research.

In summary, this paper investigated whether including additional non-confounders negatively impacted exposure effect estimation and inference, after appropriately controlling for all true confounders. This was done via exhaustive simulation with varying numbers of true confounders and sample sizes of $1k$ and $10k$. One thousand random scenarios were generated for each sample size and number of true confounders, while pathological scenarios that resulted in separation or no variability were removed. Prior to removing these simulation scenarios, the mean results were greatly affected by scenarios where separation caused poor inference - and we were not interested in exploring Firth's correction in this paper.

Acknowledgements: This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science ([12], [15]).

References

[1] James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[2] N Breslow. Design and analysis of case-control studies. *Annual Review of Public Health*, 3(1):29–54, 1982. PMID: 6756431.

[3] Miguel A Hernan et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology*, 155(2):176–184, 2002.

[4] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38, 1993.

[5] Laurence S. Freedman and David Pee. Return to a note on screening regression equations. *The American Statistician*, 43(4):279–282, 1989.

[6] Sander Greenland, Judea Pearl, and James M. Robins. Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29 – 46, 1999.

[7] D Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[8] P Howards. An overview of confounding. part 2: how to identify it and special situations. *Acta obstetricia et gynecologica Scandinavica*, 97(4):400–406, 2018.

[9] MA Mansournia, A Geroldinger, S Greenland, and Heinze. G. Separation in logistic regression: Causes, consequences, and control. *Am J Epidemiology*, 187:864–870, 2018.

[10] M Papathomas. Correlated binary variables and multi-level probability assessments. *Scandinavian journal of statistics*, 35(1):169–185, 2008.

[11] Peduzzi, P., J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 12:1373–1379, December 1996.

[12] R. et al Pordes. The open science grid. *J. Phys. Conf. Ser. 78*, 2007.

[13] L.D. Robinson and N.P. Jewell. Some surprising results about covariate adjustment in logistic regression models. *International statistical review*, 1991.

[14] E Schisterman, S Cole, and R. Platt. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20:488–495, 2009.

[15] I. Sfiligoi, D. C. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, and F. Wurthwein. The pilot way to grid resources using glideinwms. *2009 WRI World Congress on Computer Science and Information Engineering*, 2:428–432, 2009.

[16] M. et al van Smeden. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol.*, 16(1), 2016.

[17] T.J. VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 34:211–219, 2019.

[18] E.J. Williamson, A.J. Walker, and K. et al. Bhaskaran. Factors associated with covid-19-related death using opensafely. *Nature*, 584:430–436, 2020.

[19] J Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928.