Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Bliss, J Biomet Biostat 2012, 3:4
DOI: 10.4172/2155-6180.1000142

Research Article Open Access

Determining the Probability Distribution and Evaluating Sensitivity and False Positive Rate of a Confounder Detection Method Applied To Logistic Regression

Robin Bliss[1,2], Janice Weinberg[3], Thomas Webster[1] and Veronica Vieira[1,4*]

[1]Department of Environmental Health, Boston University School of Public Health, Boston, MA, USA
[2]Orthopedic and Arthritis Center for Outcomes Research, Brigham and Women's Hospital/Harvard Medical School, Boston, MA, USA
[3]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[4]Department of Population Health and Disease Prevention, Program in Public Health, University of California Irvine, USA

Abstract

Background: In epidemiologic studies researchers are often interested in detecting confounding (when a third variable is both associated with and affects associations between the outcome and predictors). Confounder detection methods often compare regression coefficients obtained from "crude" models that exclude the possible confounder(s) and "adjusted" models that include the variable(s). One such method compares the relative difference in effect estimates to a cutoff of 10% with differences of at least 10% providing evidence of confounding.

Methods: In this study we derive the asymptotic distribution of the relative change in effect statistic applied to logistic regression and evaluate the sensitivity and false positive rate of the 10% cutoff method using the asymptotic distribution. We then verify the results using simulated data.

Results: When applied to a logistic regression models with a dichotomous outcome, exposure, and possible confounder, we found the 10% cutoff method to have an asymptotic lognormal distribution. For sample sizes of at least 300 the authors found that when confounding existed, over 80% of models had >10% changes in odds ratios. When the confounder was not associated with the outcome, the false positive rate increased as the strength of the association between the predictor and confounder increased. When the confounder and predictor were independent of one another, false positives were rare (most < 10%).

Conclusions: Researchers must be aware of high false positive rates when applying change in estimate confounder detection methods to data where the exposure is associated with possible confounder variables.

Keywords: 10% Rule, Variable Selection, Model Building, Sensitivity, False Positive Rate

Introduction

Detecting confounding is an important part of research. Uncontrolled confounding can cause spurious, magnified, or minimized results and can produce effect sizes comparable to those observed in epidemiologic studies. Confounding is defined as the mixing of effects between an exposure, an outcome, and a third extraneous variable, known as a confounder [1-3]. Three standards are often used to describe confounder variables: (1) Confounders must be associated with the exposure of interest, (2) confounders must be associated with the outcome of interest, and (3) confounders must not lie on the causal pathway between the exposure and outcome [1,2]. In other words, a confounder must be a separate and distinct trait that is associated with, but not affected by, the exposure and the outcome of interest.

Causal pathway models and Directed Acyclic Graphs (DAG) are graphical displays that can be used to describe and illustrate relationships between variables, including confounding [4], as shown in Figure 1. Arrows indicate associations between variables with the direction illustrating the causal direction. As there is no path that can be traced from exposure to outcome through, the confounder does not fall on the pathway between E and Y. DAGs are useful tools to describe associations between variables without the need for regression assumptions. When qualitative associations between variables are unknown, DAGs are limited in requiring associations to be stated a priori and are therefore not appropriate for model building scenarios [4,5].

In practice, it is difficult to determine whether a variable will be a confounder in the population of interest. Data driven confounder detection methods often examine "crude" and "adjusted" estimates of effect size in samples drawn from the population [1,6-8]. One popular method is to evaluate the magnitude of confounding by examining the relative change in estimated effect size before and after adjusting for the possible confounder. The difference in effect size is divided by the crude effect estimate, providing the magnitude of the difference, relative to the unadjusted effect size. If the relative difference is "large" (i.e., > 10%) [2,6,8], investigators conclude that there is confounding [1,6-8]. When researchers detect confounding of the association between the exposure and outcome using this "change in estimate" approach, the model adjusted for the confounder is generally preferred [1,6-8].

Maldonado and Greenland (1993) applied crude and adjusted regression models to simulated data. They selected the preferable model based on the results of several confounder detection methods by evaluating the estimated effect size, the average relative bias, the

***Corresponding author:** Veronica Vieira, University of California Irvine, Irvine, CA, United States, Email:vvieira@uci.edu

Received May 02, 2012; **Accepted** May 21, 2012; **Published** May 23, 2012

Citation: Bliss R, Weinberg J, Webster T, Vieira V (2012) Determining the Probability Distribution and Evaluating Sensitivity and False Positive Rate of a Confounder Detection Method Applied To Logistic Regression. J Biomet Biostat 3:142. doi:10.4172/2155-6180.1000142

the estimate obtained from the selected model compared to the model underlying the simulated data. The authors focused their evaluation on the precision of estimates produced from single realizations of data. While their results reflect the precision of estimates as the confounder detection methods are applied in practice, Maldonado and Greenland did not evaluate whether the correct model was selected given the simulation parameters [6]. It is important to evaluate the sensitivity and false positive rate (1-specificity) of confounding detection methods to better understand the method reliability in providing accurate information under a variety of scenarios.

As data driven methods generally rely on some change statistic, it is reasonable to assume that the sensitivity and false positive rate of the detection method rely on the probability distribution of the statistic. A better understanding of this distribution may provide inference on the strengths and limitations of such confounder detection methods. To our knowledge, no studies have examined the probability distribution of the relative change in estimate statistic, how frequently the adjusted model is selected when confounding is present (sensitivity) or how often it is falsely detected when confounding is absent (false positive rate; 1-specificity). In this study, we derive the probability distribution of a 10% change in odds ratios rule for logistic regression results and use synthetic data to compare simulated results to derived results. Under a variety of scenarios we determine how often we expect to correctly detect confounding and how often we misclassify a change in estimated effect size observed by chance as confounding.

Derivation of Probability Distribution of the Change in Estimate Statistic

Suppose that we have an observable, dichotomous outcome, Y, such that $Y \sim Binomial(p)$. If Y depends on two observable variables, E and C, the true association between Y, E, and C can be expressed as a function of the parameter p:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 E + \beta_2 C \tag{1}$$

We can represent the effect estimates of E and C as odds ratios, computed as $exp(\beta_1)$ and $exp(\beta_2)$. After observing Y, E, and C we may be interested in finding estimates for coefficients β_0, β_1, and β_2 to better understand the associations between the variables. While no closed-form solution exists in general, maximum likelihood estimates for coefficients β_0, β_1, and β_2 can be found using the Newton-Raphson Algorithm. As a result, the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ have asymptotic normal distributions.

In practice, we may wish to determine whether the variable C is a confounder. To do this, we can compare two nested logistic regression models: equation 2 including variable E as the sole predictor variable and equation 3 including both E and C

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\theta}_0 + \hat{\theta}_1 E\,. \tag{2}$$

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = \hat{\gamma}_0 + \hat{\gamma}_1 E + \hat{\gamma}_2 C \tag{3}$$

Note that if C is a confounder of the association between Y and E then we expect $\theta_1 \neq \gamma_1$ and can look for evidence of confounding by evaluating the relative difference between effect size estimates. When using logistic regression, we typically compare the relative difference between odds ratios:

$$R = \frac{\exp\left(\hat{\theta}_1\right) - \exp\left(\hat{\gamma}_1\right)}{\exp\left(\hat{\theta}_1\right)} = 1 - \exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right) \tag{4}$$

We compare R to some predetermined cutoff, such as +/-0.10, to evaluate the likelihood of confounding.

We can find the probability density function of R to determine the probability of having evidence of confounding in a given scenario, i.e. $P\left(|R| > c\right)$ where C is some predetermined cutoff. R is a function of estimates $\hat{\theta}_1$ and $\hat{\gamma}_1$, computed using the observed values of Y, E, and C. Though the parameters they estimate may differ $(\theta_1 \neq \gamma_1)$, the statistics are not independent and their correlation is nonzero. As $\hat{\theta}_1$ and $\hat{\gamma}_1$ are asymptotically normal, it is straightforward to show that if $\hat{\gamma}_1 \sim N\left(\gamma_1, \sigma^2\right)$, $\hat{\theta}_1 \sim N\left(\theta_1, \tau^2\right)$, and $\text{corr}\left(\hat{\gamma}_1, \hat{\theta}_1\right) = \rho$, then $\hat{\gamma}_1 - \hat{\theta}_1 \sim N\left(\gamma_1 - \theta_1, \sigma^2 + \tau^2 - 2\rho\sigma\tau\right)$. As a result, $\exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right) \sim LN\left(\gamma_1 - \theta_1, \sigma^2 + \tau^2 - 2\rho\sigma\tau\right)$ [9], where LN is the lognormal distribution. We are interested in $P\left(|R| > c\right)$ which can be re-expressed as follows:

$$\begin{aligned}
&P\left(|R| > c\right) \\
&= P\left(\left|1 - \exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right)\right| > c\right) \\
&= P\left(1 - \exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right) > c\right) + P\left(1 - \exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right) < -c\right) \\
&= 1 - P\left(-c < 1 - \exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right) < c\right) \\
&= 1 - P\left(-c - 1 < -\exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right) < c - 1\right) \\
&= 1 - P\left(1 - c < \exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right) < 1 + c\right)
\end{aligned} \tag{5}$$

Using the cumulative density function of the lognormal distribution, this can be expressed as:

$$P\left(|R| > c\right) = 1 - \int_{1-c}^{1+c} \frac{1}{\sqrt{2\pi\left(\sigma^2 + \tau^2 - 2\rho\sigma\tau\right)}} \frac{\exp\left(-\left(\hat{\gamma}_1 - \hat{\theta}_1 - \left(\gamma_1 - \theta_1\right)\right)^2 \Big/ 2\left(\sigma^2 + \tau^2 - 2\rho\sigma\tau\right)\right)}{\exp\left(\hat{\gamma}_1 - \hat{\theta}_1\right)} d\left(\hat{\gamma}_1 - \hat{\theta}_1\right) \tag{6}$$

For large sample sizes, we can use simulated data to obtain estimated (empirical) values for θ, γ, τ, σ, and ρ, and compute the expected probability of detecting confounding.

Note that the derivation of the asymptotic distribution of is independent of the distributions of the exposure and confounder variables and the derivation is true for continuous as well as categorical predictor variables.

Simulated Data

Simulated data had a dichotomous outcome (Y), a dichotomous exposure (E), and a dichotomous third variable and possible confounder (C). Probability of exposure was range of effect sizes commonly observed in epidemiologic research. When the confounder C was present, the probability of the outcome was decreased by half, stayed the same, or was doubled:

$$P(Y \mid C=1) = kP(Y) \quad (7) \quad \text{for} \quad k = 0.5, 1.0, 2.0$$

In scenarios where did not affect the probability of disease ($k = 1$) no confounding was present.

Associations between C and E were defined by a conditional probability.

$$P(C \mid E=1) = k$$
$$P(C \mid E=0) = 1-k, k = 0.25, 0.50, 0.67, 0.75 \quad (8)$$

For example, when $k = 0.25$, one in four exposed observations had confounder trait C, while three of four unexposed observations had trait C. For $k = 0.50$, one in two observations had trait C and there was no association between C and the exposure variables. No confounding was present.

When there is no association between E and Y (odds ratio = 1.0), there cannot be confounding. While a change in the estimate after adjustment could be misclassified as confounding, this scenario is not evaluated in this study.

The underlying probability of Y for unexposed subjects without C was 10%. Sample sizes were 300, 1000 and 3000. For each set of parameters, 1000 datasets were simulated.

Logistic regression models were applied to simulated data and estimated odds ratios and logodds were recorded for the crude model, predicting Y by E (equation 2), and for the adjusted model, which included the exposure E and the possible confounder C as independent variables (equation 3). All simulations and statistical analyses were performed using R version 2.11.1[10].

Detecting Confounding using Probability Distributions

For each set of simulation parameters, the mean and standard deviation of the logodds for the effect of exposure E were computed for the crude and adjusted models across the 1000 simulated datasets. We also computed the correlation between the crude and adjusted coefficients. Using the empirical estimates for θ and τ in the crude model, for γ and σ in the adjusted model, and for the correlation ρ, we computed the probability of $|R|$ having a magnitude greater than 10% based on the probability distribution derived above. We then determined the sensitivity for simulation scenarios where confounding

Confounder		N = 300				N = 1000				N = 3000			
		P(Confounder|Exposure)†				P(Confounder|Exposure)†				P(Confounder|Exposure)†			
OR*	Effect^	0.25	0.50	0.67	0.75	0.25	0.50	0.67	0.75	0.25	0.50	0.67	0.75
		S	FPR	S	S	S	FPR	S	S	S	FPR	S	S
0.5	P(Y)x0.5	0.8711	0.0447	0.8205	0.8893	0.9610	<0.0001	0.9253	0.9590	0.9985	<0.0001	0.9970	0.9983
0.5	P(Y)x1‡	0.6817	0.0002	0.5124	0.6889	0.4468	<0.0001	0.2303	0.4774	0.2109	<0.0001	0.0354	0.2155
0.5	P(Y)x2	0.9264	0.0488	0.8843	0.9157	0.9947	0.0002	0.9828	0.9944	>0.9999	<0.0001	0.9999	>0.9999
2	P(Y)x0.5	0.9096	0.0611	0.8798	0.9143	0.9887	0.0002	0.9840	0.9921	>0.9999	<0.0001	0.9999	>0.9999
2	P(Y)x1‡	0.6196	<0.0001	0.4035	0.6100	0.3672	<0.0001	0.1299	0.3394	0.0983	<0.0001	0.0067	0.1045
2	P(Y)x2	0.9779	0.1160	0.9375	0.9698	0.9999	0.0090	0.9987	0.9998	>0.9999	<0.0001	>0.9999	>0.9999
3	P(Y)x0.5	0.9257	0.0797	0.9075	0.9468	0.9953	0.0028	0.9952	0.9966	>0.9999	<0.0001	>0.9999	>0.9999
3	P(Y)x1‡	0.6049	<0.0001	0.4039	0.5845	0.2984	<0.0001	0.1085	0.3257	0.0745	<0.0001	0.0040	0.0930
3	P(Y)x2	0.9886	0.3127	0.9681	0.9915	>0.9999	0.1518	0.9999	>0.9999	>0.9999	0.0223	>0.9999	>0.9999

*OR is the odds ratio between exposure and outcome.
^Confounder Effect is the influence of the presence of the confounder variable on the outcome. The confounder halves the probability of the outcome (P(Y)x0.5), has no effect on the outcome (P(Y)x1), or doubles the probability of the outcome (P(Y)x2).
†P(Confounder|Exposure) is the association between the possible confounder and exposure variable. The probability of the confounder among exposed subjects is 0.25, 0.50, 0.67, or 0.75. The probability of confounder among unexposed subjects is 0.75, 0.50, 0.33, and 0.25, respectively.
‡Row displays False Positive Rate as there is no association between possible confounder and outcome, Y, and therefore no confounding.

Table 1: Sensitivity and False Positive Rate of 10% Change in Odds Ratio Rule Using Lognormal Probability Distribution (P(Exposure)=0.50)

Confounder		N = 300				N = 1000				N = 3000			
		P(Confounder|Exposure)†				P(Confounder|Exposure)†				P(Confounder|Exposure)†			
OR*	Effect^	0.25	0.50	0.67	0.75	0.25	0.50	0.67	0.75	0.25	0.50	0.67	0.75
		S	FPR	S	S	S	FPR	S	S	S	FPR	S	S
0.5	P(Y)x0.5	0.869	0.061	0.821	0.899	0.961	<0.001	0.926	0.955	0.999	<0.001	0.999	0.999
0.5	P(Y)x1‡	0.686	0.010	0.487	0.690	0.452	<0.001	0.224	0.479	0.219	<0.001	0.034	0.218
0.5	P(Y)x2	0.938	0.057	0.901	0.927	0.994	0.002	0.989	0.995	>0.999	<0.001	>0.999	>0.999
2	P(Y)x0.5	0.922	0.075	0.882	0.914	0.993	0.003	0.995	0.991	>0.999	<0.001	>0.999	>0.999
2	P(Y)x1‡	0.630	0.004	0.404	0.590	0.350	<0.001	0.123	0.324	0.101	<0.001	0.007	0.095
2	P(Y)x2	0.985	0.113	0.955	0.973	>0.999	0.021	0.998	>0.999	>0.999	<0.001	>0.999	>0.999
3	P(Y)x0.5	0.936	0.084	0.914	0.951	0.997	0.006	0.999	0.998	>0.999	<0.001	>0.999	>0.999
3	P(Y)x1‡	0.590	0.003	0.380	0.580	0.288	<0.001	0.114	0.324	0.074	<0.001	0.004	0.090
3	P(Y)x2	0.996	0.267	0.978	0.994	>0.999	0.150	>0.999	>0.999	>0.999	0.030	>0.999	>0.999

*OR is the odds ratio between exposure and outcome.
^Confounder Effect is the influence of the presence of the confounder variable on the outcome. The confounder halves the probability of the outcome (P(Y)x0.5), has no effect on the outcome (P(Y)x1), or doubles the probability of the outcome (P(Y)x2).
†P(Confounder|Exposure) is the association between the possible confounder and exposure variable. The probability of the confounder among exposed subjects is 0.25, 0.50, 0.67, or 0.75. The probability of confounder among unexposed subjects is 0.75, 0.50, 0.33, and 0.25, respectively.
‡Row displays False Positive Rate as there is no association between possible confounder and outcome, Y, and therefore no confounding.

Table 2: Sensitivity and False Positive Rate of 10% Change in Odds Ratio Rule From Simulated Data (P(Exposure)=0.50).

truly existed and the false positive rates for scenarios where confounding did not exist.

Detecting Confounding using Simulated Data

The relative change in estimate statistic *R* was computed using the odds ratios for crude and adjusted models applied to each simulated dataset. Confounding was defined as a change from the crude to the adjusted odds ratio with magnitude exceeding +/− 10%. We determined the observed sensitivity and false positive rates for the simulation scenarios.

Results

Table 1 displays the derived sensitivity and false positive rates for a probability of exposure of 50%. Table 2 contains corresponding sensitivity and false positive rate estimates from simulated data. Similarly, Tables 3 and 4 display the computed and simulated sensitivity and false positive rates, respectively, for a 5% probability of exposure.

Across odds ratios between *Y* and *E* with 50% exposure, stronger associations between *E* and *C* corresponded to increased sensitivity. For example, for an odds ratio of 3.0 and a sample size of 300, when *C* doubled the probability of disease and the probability of *C* was 0.67 among exposed subjects, the sensitivity computed using the lognormal distribution (equation 6) was 0.9681. This is compared to a sensitivity of 0.9915 when the probability of *C* was 0.75 among exposed subjects. For the same two scenarios, 978 and 994 of the 1000 simulated datasets respectively, had relative differences of the crude and adjusted effect sizes that exceeded 10% (sensitivity of 0.978 and 0.994).

Sensitivity was higher when the probability of *Y* was doubled in the presence of the confounder than when the probability of *Y* was reduced by half. The sensitivity for detecting confounding was roughly symmetric for inverse associations between *E* and *C* (i.e. the probability of *C* was 0.75 vs. 0.25 among exposed subjects; Table 1, 2).

When *C* did not impact the probability of disease but was associated with the exposure variable, the false positive rates were substantially larger for reduced sample sizes (Figure 2). False positive rates were at least 40% for a sample size of 300, at least 10% for a sample size of 1000, and at least 0.4% for a sample size of 3000 (Table 1). In all scenarios, the false positive rate when *C* was not associated with exposure *E*

Confounder		N = 300				N = 1000				N = 3000			
		P(Confounder|Exposure)†				P(Confounder|Exposure)†				P(Confounder|Exposure)†			
OR*	Effect^	0.25	0.50	0.67	0.75	0.25	0.50	0.67	0.75	0.25	0.50	0.67	0.75
		S	FPR	S	S	S	FPR	S	S	S	FPR	S	S
0.5	P(Y)x0.5	0.8590	0.4761	0.8700	0.9003	0.9690	0.0964	0.9273	0.9653	0.9998	0.0036	0.9928	0.9992
0.5	P(Y)x1‡	0.7003	0.0593	0.5154	0.7009	0.4074	<0.0001	0.1945	0.4278	0.1363	<0.0001	0.0206	0.1429
0.5	P(Y)x2	0.9220	0.4261	0.8329	0.9283	0.9922	0.1388	0.9712	0.9977	>0.9999	0.0190	0.9995	>0.9999
2	P(Y)x0.5	0.8606	0.4779	0.8392	0.9177	0.9784	0.1022	0.93	0.9661	0.9998	0.0055	0.9951	0.9991
2	P(Y)x1‡	0.7059	0.0724	0.5096	0.6681	0.4099	<0.0001	0.2005	0.4069	0.1363	<0.0001	0.0136	0.1498
2	P(Y)x2	0.9370	0.4857	0.8189	0.9235	0.9946	0.2147	0.9585	0.9978	>0.9999	0.0694	0.9991	>0.9999
3	P(Y)x0.5	0.8705	0.6537	0.8700	0.9048	0.9569	0.0672	0.8843	0.9493	0.9997	0.0014	0.9921	0.9984
3	P(Y)x1‡	0.6974	0.2368	0.5431	0.6908	0.4399	<0.0001	0.1899	0.4173	0.1625	<0.0001	0.0237	0.1798
3	P(Y)x2	0.9093	0.5116	0.8327	0.9128	0.9923	0.1261	0.9660	0.9975	>0.9999	0.0038	0.9998	>0.9999

*OR is the odds ratio between exposure and outcome.
^Confounder Effect is the influence of the presence of the confounder variable on the outcome. The confounder halves the probability of the outcome (P(Y)x0.5), has no effect on the outcome (P(Y)x1), or doubles the probability of the outcome (P(Y)x2).
†P(Confounder|Exposure) is the association between the possible confounder and exposure variable. The probability of the confounder among exposed subjects is 0.25, 0.50, 0.67, or 0.75. The probability of confounder among unexposed subjects is 0.75, 0.50, 0.33, and 0.25, respectively.
‡Row displays False Positive Rate as there is no association between possible confounder and outcome, Y, and therefore no confounding.

Table 3: Sensitivity and False Positive Rate of 10% Change in Odds Ratio Rule Using Lognormal Probability Distribution (P(Exposure)=0.05).

Confounder		N = 300				N = 1000				N = 3000			
		P(Confounder|Exposure)†				P(Confounder|Exposure)†				P(Confounder|Exposure)†			
OR*	Effect^	0.25	0.50	0.67	0.75	0.25	0.50	0.67	0.75	0.25	0.50	0.67	0.75
		S	FPR	S	S	S	FPR	S	S	S	FPR	S	S
0.5	P(Y)x0.5	0.880	0.327	0.787	0.866	0.972	0.061	0.939	0.974	>0.999	0.003	0.998	>0.999
0.5	P(Y)x1‡	0.672	0.093	0.459	0.655	0.398	0.002	0.177	0.396	0.167	<0.001	0.029	0.183
0.5	P(Y)x2	0.948	0.327	0.858	0.916	>0.999	0.087	0.986	>0.999	>0.999	0.003	>0.999	>0.999
2	P(Y)x0.5	0.872	0.316	0.804	0.910	0.964	0.093	0.952	0.975	>0.999	0.008	0.999	>0.999
2	P(Y)x1‡	0.637	0.061	0.405	0.632	0.406	<0.001	0.184	0.410	0.129	<0.001	0.017	0.144
2	P(Y)x2	0.954	0.334	0.838	0.942	0.998	0.132	0.983	0.999	>0.999	0.023	>0.999	>0.999
3	P(Y)x0.5	0.872	0.312	0.826	0.866	0.983	0.109	0.962	0.986	0.999	0.014	0.999	>0.999
3	P(Y)x1‡	0.659	0.074	0.445	0.622	0.375	0.001	0.177	0.385	0.140	<0.001	0.013	0.163
3	P(Y)x2	0.949	0.400	0.815	0.939	>0.999	0.192	0.975	>0.999	>0.999	0.071	>0.999	>0.999

* OR is the odds ratio between exposure and outcome.
^ Confounder Effect is the influence of the presence of the confounder variable on the outcome. The confounder halves the probability of the outcome (P(Y)x0.5), has no effect on the outcome (P(Y)x1), or doubles the probability of the outcome (P(Y)x2).
† P(Confounder|Exposure) is the association between the possible confounder and exposure variable. The probability of the confounder among exposed subjects is 0.25, 0.50, 0.67, or 0.75. The probability of confounder among unexposed subjects is 0.75, 0.50, 0.33, and 0.25, respectively.
‡ Row displays False Positive Rate as there is no association between possible confounder and outcome, Y, and therefore no confounding.

Table 4: Sensitivity and False Positive Rate of 10% Change in Odds Ratio Rule From Simulated Data (P(Exposure)=0.05).

was similar to or smaller than the false positive rate for *C* having no association with the outcome. For large sample sizes, when *C* was not associated with exposure *E*, the false positive rates approached zero (Table 1, 2; Figure 2).

The sensitivity and false positive rate of the change in estimate method for detecting confounding derived from equation 6 were very similar to probabilities from simulated data. The largest difference observed had a magnitude of 0.05; however most differences were less than +/- 0.015 (Table 1-4). As derivations based on equation 6 are asymptotic in nature, increased sample sizes provided more similar sensitivity and false positive rates between the computed value from the lognormal distribution and the observed values from simulated data.

Similar results were observed for a 5% probability of *E* with reduced sensitivity and increased false positive rates, particularly for a sample size of 300 (Table 3, 4).

Discussion

Across the examined sample sizes and strengths of associations between confounders, exposures, and outcomes, the change in estimate confounding detection method, applied with a 10% cutoff, had high sensitivity. The method had low false positive rates when there was no association between the exposure and third extraneous variable. False positives were more common when the exposure and third variable were associated but there was no association between the third variable and outcome (Figure 2). The inflated false positive rate may be due to the strong association between the exposure and the third variable resulting in an apparent weak association between the third variable and the outcome. As a result, the third variable contributes similar information about the outcome as the exposure and the regression coefficient for the exposure variable may be affected. When the exposure and third variable are independent they contribute different information about the outcome and the crude and adjusted regression coefficients are similar.

The results presented from this study are for a logistic regression with a single exposure and possible confounder variable. Despite this, the mathematical derivation of the lognormal distribution can be similarly applied to other generalized linear models and we would expect similar results. We have not examined the implications of these results in the context of time to event analysis, Cox proportional hazard models, or semi- and nonparametric analyses. In this study we considered a dichotomous exposure and confounder. Results do not rely on the distribution of either variable and similar results are expected for continuous, nominal, and ordinal predictor variables. We examined the implications of a 10% rule for confounder detection. Similar discussions can be made for a 15% or 20% cutoff which provide lower sensitivity and higher specificity than those observed here. Of note, in these scenarios we would expect the same patterns of increased false positive rates when the outcome is not associated with the third variable regardless of the cutoff percentage. While we have not examined all scenarios, we believe that the presence of additional covariates, confounders, and interactions in regression models would not affect the results of the confounder detection methods examined in this study.

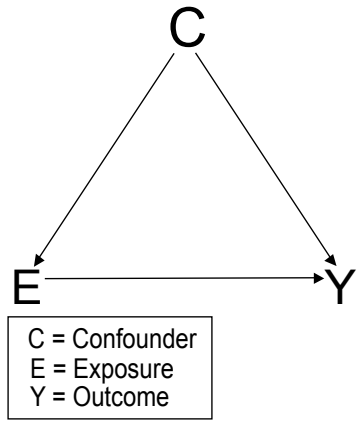

Figure 1: Causal Diagram of a Single Confounder
Figure 1 displays a causal diagram where there is one outcome (*Y*), one exposure variable (*E*), and one confounder (*C*).

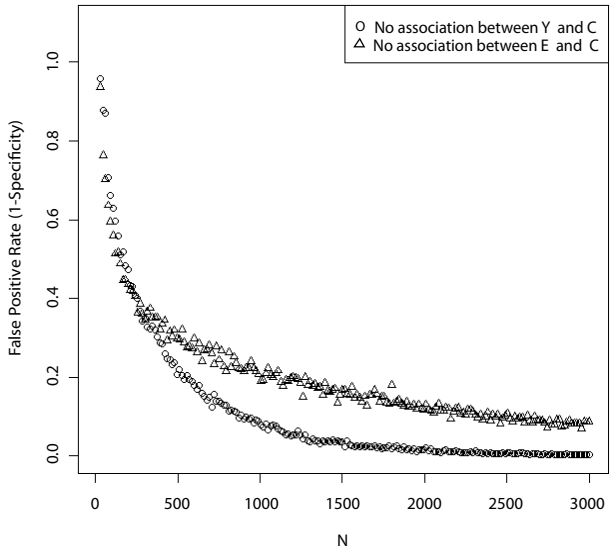

Figure 2: False Positive Rates of Contrasting Scenarios, Odds Ratio Between Exposure and Outcome=2.0, Derived From Lognormal Distribution
Figure 2 illustrates the false positive rates (1-specificity) of two contrasting scenarios, each with an odds ratio between exposure (*E*) and outcome (*Y*) of 2.0. In the first scenario there is no association between *Y* and third variable (*C*) but there is an association between *E* and *C*. Exposed subjects have twice the probability of trait *C* as unexposed subjects. (Line *CY* is missing in Figure 1.) In the second scenario there is no association between *E* and *C* but trait *C* doubles the probability of outcome *Y*. (Line *EC* is missing in Figure 1.) All probabilities are derived from the lognormal distribution using empirical estimates from simulated data for parameter values.

For linear regression, it is straightforward to show that the comparable statistic for the relative difference in parameter estimates can be expressed as the ratio of two correlated normally distributed random variables. Determining the sensitivity and false positive rates in this setting are left for future research. Further research is also required to understand the implications of false positive and false negative confounding results with survival analysis, semi-, and nonparametric regression methods.

While investigators often assume that a change in effect size in a single realization of data indicates confounding [3,11], crude and adjusted effect sizes may differ for other reasons. Noncollapsibility is when effect measures change upon stratification of the covariate [1-3]. It can occur when the third variable is affected by the exposure, and, therefore, does not meet the definition of a confounder. The result can

be large changes between crude and adjusted models that are not due to confounding [3].

The effect measure (i.e., risk ratios, odds ratios, logodds, etc.) compared will affect the distribution of the change in effect size statistic and the results of the comparison [7]. For example, simulated data with 1000 observations, a dichotomous predictor, an odds ratio of 2.0, a probability of confounder of 0.75 among exposed subjects, and a doubling of the probability of disease when the confounder is present had an unadjusted odds ratio, log odds and risk ratio of 1.641, 0.495, and 1.076, respectively. The adjusted model had an odds ratio, log odds and risk ratio of 1.994, 0.696 and 1.101, respectively corresponding to a 22.2%, 40.5% or 2.3% change. As a result, the conclusion of whether confounding exists would depend on the effect measure selected.

It is possible for a variable to satisfy all three criteria of being a confounder (see Figure 1) while not meeting the definition of a confounder. For example, a dichotomous variable may have opposing directions of association with the outcome and exposure variable, thereby canceling the effect. In this circumstance, the variable is associated with both the exposure and the outcome and does not lie in the causal pathway; however it may have no effect on the association between the exposure and outcome [7]. The third variable would not be classified as a confounder by definition or by a change in estimate method, though it does meet all three standards.

It is important to identify and control for confounder variables to better understand the underlying relationships between predictors and outcomes. Unmeasured confounding alone can cause effect sizes of magnitudes commonly observed in epidemiologic studies [12]; however the expansion of models to include additional variables is not without cost. In this study, we found that when the exposure and possible confounder were associated with one another but the possible confounder was not associated with the outcome, false positives where the change in estimate had magnitude of at least 10% were often observed. This rate was magnified for smaller sample sizes. A practical implication of this is when investigators performing randomized clinical trials compare the experimental and control treatment arms at baseline. The groups may differ on some underlying characteristic and it is standard practice to control for variables associated with treatment group [1]. The result is the misclassification of confounding, the selection of models with more predictor terms than needed, with fewer available degrees of freedom, and with inflated confidence intervals for the exposure of interest. Such misclassification is more problematic with small sample sizes where less precision and fewer degrees of freedom are available. For small or moderate sample sizes, in a single realization of data, investigators will have an increased probability of false positives and may include variables in regression models that are, in truth, unnecessary and may alter the association observed; conclusions made, and reduce the precision of estimated effects when compared to population parameters [2,6,7].

Conclusions

Comparing the relative change in estimated odds ratios between crude and adjusted logistic regression models provides a simple, readily available and easily applicable confounder detection method for statistical and epidemiological applications. Such methods have high sensitivity and low false positive rates when there are true associations between the outcome and exposure and the outcome and possible confounder, regardless of sample size or the association between the exposure and confounder. Higher false positive rates are expected when the outcome is not associated with the third variable, particularly for small or moderate sample sizes.

Researchers must be aware of possible false positives when applying change in estimate methods. When evaluating confounding in studies with fewer than 1000 subjects, we recommend evaluating the association between the outcome and possible confounder variable as well as the change in estimated odds ratios. While the examination of the magnitude of effect size change should not be used as the only source of information when selecting possible confounders and building models for analysis, when used in conjuncture with other information, change in estimate methods may provide useful information to identify confounders, build parsimonious models, and better understand associations between outcomes and exposures.

Acknowledgements

This research was supported by grant P42ES007381 from the National Institute of Environmental Health (NIEHS), NIH and grant T32AR055885 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIEHS, NIAMS, or NIH.

Competing interests

The authors declare that they have no competing interests.

Author Contributions

RLB performed mathematical derivations, simulated data, and drafted the manuscript. JW reviewed the mathematical derivations. RLB, JW, VMV, and TFW discussed the concept and design of the study, simulation scenarios to be included, and each participated in the reviewing and revising of the manuscript. All authors read and approved the final manuscript.

References

1. Aschengrau A, Seage GR (2003) Essentials of Epidemiology in Public Health. Boston, MA: Jones and Bartlett Publishers.
2. Rothman KJ, Greenland S, Lash TL (2008) Modern Epidemiology, 3rd edn. New York, NY: Lippincott Williams & Wilkins.
3. Greenland S, Robins JM (2009) Identifiability exchangeability and confounding revisited. Epidemiol Perspect Innov 6: 4.
4. Greenland S, Pearl J, Robins JM (1999) Causal Diagrams for Epidemiologic Research. Epidemiology 10: 37-48.
5. Weinberg CR (2007) Can DAGs Clarify Effect Modification? Epidemiology 18: 569-572.
6. Maldonado G, Greenland S (1993) Simulation Study of Confounder-Selection Strategies. Am J Epidemiol 138: 923-936.
7. McNamee R (2003) Confounding and Confounders. Occup Environ Med 60: 227-234.
8. Grayson DA (1987) Confounding confounding. Am J Epidemiol 126: 546-553.
9. Casella G, Berger RL: Statistical Inference 2nd Edition. 2nd edition edn. Pacific Grove, CA: Duxbury Thomson Learning; 2002.
10. R: v 2.11.1. The R Foundation for Statistical Computing; 2010.
11. Miettinen OS, Cook EF (1981) Confounding: Essence and Detection. Am J Epidemiol 114: 593-603.
12. Fewell Z, Smith GD, Sterne JAC (2007) The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol 166: 646-655.