Research Article

An Investigation of the Significance of Residual Confounding Effect

Wenbin Liang,1 Yuejen Zhao,2 and Andy H. Lee3

1 National Drug Research Institute, Curtin University, G.P.O. Box U 1987, Perth, WA 6845, Australia
2 Northern Territory Department of Health, Darwin, NT 0800, Australia
3 School of Public Health, Curtin University, Perth, WA 6845, Australia

Correspondence should be addressed to Wenbin Liang; w.liang@curtin.edu.au

Received 18 December 2013; Accepted 10 January 2014; Published 17 February 2014

Academic Editor: Tanya Chikritzhs

Copyright © 2014 Wenbin Liang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. Observational studies are commonly conducted in health research. However, due to their lack of randomization, the estimated associations between the outcome and the exposure can be affected by unmeasured confounding factors. It is important to determine how likely a significant association observed between an outcome variable and a noncausally related exposure may be introduced by residual confounding factors.

Methods. A simulation approach is developed based on the sufficient cause model to test the likelihood of significant associations observed between a noncausally related exposure and the outcome.

Results. Based on the estimates from all 500 replicates, the association between the exposure and the outcome is found to be significant in 386 (77%) replicates when all confounders (component causes) are controlled for in the model. However, when a subset of real component causes and some noncausal factors are controlled for in the model, the association between exposure and the outcome becomes significant in 487 (97%) replicates.

Conclusion. Even when all confounding factors are known and controlled for using conventional multivariate analysis, the observed association between exposure and outcome can still be dominated by residual confounding effects. Therefore, an observed significant association apparently provides limited evidence for a causal relationship.

1. Introduction

Ethical and budgetary constraints often limit the application of experimental study designs in health research, so that observational studies such as cohort or case-control studies have been widely undertaken as methodological alternatives [1–5]. However, due to the lack of randomization, the estimates so obtained can be influenced by uncontrolled or unmeasured confounders and typically, the confounders bias estimates from their true values [6–12]. According to the epidemiological literature, a confounder must meet the following conditions: (i) being a cause of the disease, or a proxy of cause(s), in unexposed people; (ii) being correlated with exposure in the study population; (iii) not being an intermediate step in the causal pathway between the exposure and the disease [1, 13–16]. To deal with confounding effects, known or suspected confounders are measured together with the exposure and outcome of interest. Multivariate analyses are then performed to measure the association between the exposure and the outcome while attempting to remove the effects of such known or suspected confounders [8, 13, 17–19].

Under the sufficient cause model, a sufficient cause means a complete causal mechanism, which can be defined as a combination of minimal conditions (necessary elements) and events that inevitably produce disease, while the necessary elements that constitute a sufficient cause are component causes [2]. It is common that component causes and compositions of sufficient causes are unknown, with simultaneous existence of measurement errors, misclassifications for exposures, confounders, and outcomes [8, 20–23]. Consequently, the estimated associations between the outcome and the exposure remain likely to be affected by unmeasured confounding factors. For example, even in well-designed studies, significant protective associations occurred between true nonprotective exposures and outcomes are actually caused by unmeasured confounding factors [24, 25]. It is thus important to investigate how likely a significant association observed between an outcome variable and a noncausally
related exposure may be introduced by residual confounding
to test the likelihood of observing significant associations between a
in this study, we develop a simulation approach to test a noncausally related exposure and the outcome variable based on
standard multivariate analysis, given that the compositions of sufficient causes are not recognized, but either all risk factors/component causes are known and controlled, or
only some of the risk factors/component causes are known and controlled. There are two objectives: (1) to investigate
the likelihood of false positive observations in observational studies, (2) to propose a simulation framework for assessing
epidemiologic methods which deal with confounding effects.

2. Methods

2.1. Overview of the Simulation. The simulation process
follows the sufficient cause model [2]. For an event to occur,
at least one sufficient cause has to occur. The components of a
sufficient cause are randomly chosen from a pool of low to
moderate correlated variables, which include the exposure of
interest and 99 other variables. The exposure of interest is set
to be noncausal for the outcome and therefore it will never
be chosen as a component for a sufficient cause. Given the
correlation among the 100 variables, each chosen variable
is a potential confounding factor for the association between
the exposure and the outcome. The association between the
exposure and the outcome is then estimated using a logistic
regression model, while controlling for (i) all component
causes; and (ii) some of the component causes (selected at
random). The simulations contain 500 replicates, with each
replicate being generated through an independent process.
All simulations are performed using the STATA package
release 12. The procedures involved in each replicate are
outlined below. Details of the simulation procedure, including
the sufficient cause model and the estimation process, are
provided in the Appendix.

(1) Generate a pool of low to moderate correlated ran-
dom variables from the uniform [0,1) distribution:
\( T_{100×50000} = \{T_{i,n}\}, i = (1, 2, 3, \ldots, 100), n = (1, 2, 3, \ldots, 50000) \).

(2) Determine the composition of sufficient causes and the
threshold values of components. The total number for the types of sufficient causes for \( Y \) is randomly
chosen from \( (1, 2, 3, \ldots, 9) \). Components for each
type of sufficient causes are randomly selected from
\( T_{i,n} \), \( i = (2, 3, \ldots, 100) \). \( T_i \) is taken as the exposure,
which is set to be noncausal for \( Y \). For each observa-
tion, a sufficient cause is set to occur, when each of
its components has a value higher than its specific
threshold value. The threshold value is specific for
each component as well as each type of sufficient
cause, and it is randomly chosen from a uniform
[0.5, 0.9] distribution. This allows the threshold values
to vary between components as well as between
different sufficient causes for the same component.
To reflect the fact that exact threshold values are
typically unknown, \( T_{i,n} \) are then dichotomized into
binary form denoted by \( X_{i,n} \), \( i = (1, 2, 3, \ldots, 100) \),
\( n = (1, 2, 3, \ldots, 50000) \), by applying the following
rule: \( X_{i,n} \) is set to 1 if \( T_{i,n} > 0.7 \), and 0 otherwise.

Here, the mean 0.7 of a uniform \([0.5, 0.9]\) variable
is used instead of applying the exact threshold values,
in order to account for unavoidable measurement errors
and misclassifications in confounders and exposures.

(3) Generate competing events for \( Y \), \( E_n, n = (1, 2, 3, \ldots, 50000) \). Note that \( E \) is independent
of \( T \) and \( X \).

(4) Generate small random errors for \( Y \) to represent
measurement errors of outcome and to smooth the
computing process. \( Q \) is a Bernoulli distributed ran-
dom variable, being independent of \( E \) and \( X \) and only
accounts for a small proportion of variance of \( Y \).

(5) Determine the status (occur or not occur) of \( Y \).

(6) Determine the known (not necessary the fact) causal
factors for \( Y \) through a random process.

Details of steps 1 to 6 can be found in the Appendix.

(7) Estimate the effect of \( X_1 \) on \( Y \) when all component
causes are identified. There is no noncausal factor
being mistaken as causal factor. We have

\[
P(Y_n = 1 | X_{i,n}, C) = \frac{\exp(\beta_{1}X_{i,n} + \sum_{i=2}^{100} \beta_{i}X_{i,n}C_{i})}{1 + \exp(\beta_{1}X_{i,n} + \sum_{i=2}^{100} \beta_{i}X_{i,n}C_{i})},
\]

where \( C_i \) indicates whether \( X_i \) is involved in at least one
sufficient cause of \( Y \), that is, \( C_i = 1 \) if true and \( C_i = 0 \)
otherwise. Here, \( \beta_1 \) and \( \beta_i \) are the estimated effects of \( X_1 \) and
each of the component causes on \( Y \), respectively. To estimate
the effect of \( X_1 \) on \( Y \) when only some component causes are
known, and there are some noncausal factors being mistaken
as causal factors, we have

\[
P(Y_n = 1 | X_{i,n}, K) = \frac{\exp(\beta'_{1}X_{i,n} + \sum_{i=2}^{100} \beta'_{i}X_{i,n}K_{i})}{1 + \exp(\beta'_{1}X_{i,n} + \sum_{i=2}^{100} \beta'_{i}X_{i,n}K_{i})},
\]

where \( K_i \) indicates whether \( X_i \) is “known” or suspected to be
involved in at least one sufficient cause of \( Y \), \( \beta'_1 \), and \( \beta'_i \) are the
estimated effects of \( X_1 \) and each of the “known” risk factors
on \( Y \), respectively.

3. Results

Data obtained from replicate 1 is used as an example. Table 1
shows details of the sufficient causes and their components
for replicate 1. Overall, the incidence rate (per 1000 observa-
tion units) for \( Y \) is 32.4, while it is 20.2 among unexposed
observations (\( X_1 = 0 \)) and 89.0 among exposed observations
(\( X_1 = 1 \)). This leads to an observed crude exposed-to-
unexposed risk ratio of 4.4, though the exposure is not causal
for \( Y \). Moreover, as shown in Table 2, the strength of asso-
ciation between exposure and confounders is considerably
low, with low level of misclassifications for confounder status.
Table 1: Sufficient causes and their components for replicate 1.

<table>
<thead>
<tr>
<th>Type of sufficient cause</th>
<th>Components (cut-off points)</th>
<th>Observed frequency for the 50,000 observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$X_{17}$ (0.847), $X_{50}$ (0.850)</td>
<td>421</td>
</tr>
<tr>
<td>B</td>
<td>$X_{7}$ (0.521), $X_{29}$ (0.881), $X_{53}$ (0.619)</td>
<td>515</td>
</tr>
<tr>
<td>C</td>
<td>$X_{18}$ (0.754), $X_{30}$ (0.626), $X_{31}$ (0.504), $X_{38}$ (0.642), $X_{91}$ (0.617)</td>
<td>741</td>
</tr>
</tbody>
</table>

As described in the simulation design and the appendix, the total number of sufficient causes and the components of each possible sufficient cause vary between replicates and are determined by independent random process (i.e., sufficient cause A has two components: $X_{17}$ and $X_{50}$; sufficient cause B has three components: $X_{7}$, $X_{29}$, and $X_{53}$).

Table 2: Source and magnitude of bias in replicate 1.

<table>
<thead>
<tr>
<th>Confounder/Component</th>
<th>Correlation with exposure</th>
<th>Percentage of misclassification</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{17}$</td>
<td>0.183</td>
<td>13.4%</td>
</tr>
<tr>
<td>$X_{50}$</td>
<td>0.160</td>
<td>14.4%</td>
</tr>
<tr>
<td>$X_{7}$</td>
<td>0.135</td>
<td>26.7%</td>
</tr>
<tr>
<td>$X_{29}$</td>
<td>0.150</td>
<td>15.5%</td>
</tr>
<tr>
<td>$X_{31}$</td>
<td>0.181</td>
<td>11.6%</td>
</tr>
<tr>
<td>$X_{18}$</td>
<td>0.227</td>
<td>5.89%</td>
</tr>
<tr>
<td>$X_{20}$</td>
<td>0.155</td>
<td>10.8%</td>
</tr>
<tr>
<td>$X_{21}$</td>
<td>0.292</td>
<td>31.2%</td>
</tr>
<tr>
<td>$X_{38}$</td>
<td>0.188</td>
<td>7.9%</td>
</tr>
<tr>
<td>$X_{91}$</td>
<td>0.282</td>
<td>11.4%</td>
</tr>
</tbody>
</table>

1 Measured as the correlation coefficient between binary form of component (occurred or not occurred) and binary form of exposure in the 50,000 observations for replicate 1.

2 Measured as 1 minus the proportion of correct classification of confounder/component status (occurred or not occurred) in the 50,000 observations for replicate 1.

Given that all confounding factors (component causes) are controlled for in the model, the effect of exposure remained significant ($P < 0.001$). Table 3 suggests that the effect of exposure is further biased away from the null when only a subset of real component causes and some noncausal factors are controlled in the model.

Based on the estimates from all replicates, the association between the exposure and the outcome $Y$ is found to be significant in 386 (77%) out of the 500 replicates when all confounders (component causes) are controlled in the model. However, when a subset (rather than all) of real component causes and some noncausal factors are controlled in the model, the association between the exposure and the outcome $Y$ becomes significant in 487 (97%) out of the 500 replicates.

In addition, Figure 1 indicates that when adjusting for all the real component causes, the significantly estimated effect of the exposure is on average substantially smaller than the effects of real component causes. The mean (standard deviation), 25th, 50th, and 75th percentiles of the significant coefficients (natural logarithm of the odds ratio) are 0.22 (0.17), 0.14, 0.18, and 0.25, respectively for the noncausal exposure and are 0.73 (0.79), 0.23, 0.42, and 0.927, respectively, for the real component causes.

4. Discussion

In observational studies, when a statistical significant association arises between an exposure and the outcome in the multivariate analysis, it is usually considered as supportive evidence for causal relationship [8]. We adopt the sufficient cause model in the simulation process to investigate how likely a significant association between the exposure and the outcome may be observed when there is no causal association between the two in an observational study setting. The results indicate that significant associations between the exposure and its noncausal related outcomes are presented in more than 70% of the situations, even when assuming that all confounders (causal factors) are known to researchers and controlled for in the simulation analysis. In reality, many component causes of a disease are unknown [8, 20–23].

Moreover, results from the simulation study suggest that under the conventional multivariate analysis approach, residual confounding effects remain strong enough to influence the observed associations and an observed significant association provides only limited evidence for a causal relationship. Therefore, new methods are required to handle residual confounding effects. The simulation design adopted in this study can also serve as a platform to evaluate the performance of such methods.
Table 3: Estimates from multivariate analysis in replicate 1.

<table>
<thead>
<tr>
<th>Model adjusted for all component causes</th>
<th>Model adjusted for randomly selected component causes and noncausal factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds ratios</td>
<td>95% Confidence interval</td>
</tr>
<tr>
<td>$(X_i)$ Exposure</td>
<td>1.31</td>
</tr>
<tr>
<td>$X_3$</td>
<td>—</td>
</tr>
<tr>
<td>$X_7$</td>
<td>1.61</td>
</tr>
<tr>
<td>$X_{11}$</td>
<td>—</td>
</tr>
<tr>
<td>$X_{14}$</td>
<td>—</td>
</tr>
<tr>
<td>$X_{17}$</td>
<td>2.45</td>
</tr>
<tr>
<td>$X_{18}$</td>
<td>4.55</td>
</tr>
<tr>
<td>$X_{20}$</td>
<td>2.67</td>
</tr>
<tr>
<td>$X_{21}$</td>
<td>1.49</td>
</tr>
<tr>
<td>$X_{23}$</td>
<td>—</td>
</tr>
<tr>
<td>$X_{29}$</td>
<td>2.68</td>
</tr>
<tr>
<td>$X_{32}$</td>
<td>—</td>
</tr>
<tr>
<td>$X_{37}$</td>
<td>—</td>
</tr>
<tr>
<td>$X_{38}$</td>
<td>3.10</td>
</tr>
<tr>
<td>$X_{50}$</td>
<td>2.41</td>
</tr>
<tr>
<td>$X_{53}$</td>
<td>1.90</td>
</tr>
<tr>
<td>$X_{57}$</td>
<td>—</td>
</tr>
<tr>
<td>$X_{59}$</td>
<td>—</td>
</tr>
<tr>
<td>$X_{90}$</td>
<td>—</td>
</tr>
<tr>
<td>$X_{91}$</td>
<td>2.28</td>
</tr>
</tbody>
</table>

—: variables not included in the model.

There are several advantages of our simulation design. Firstly, although all component causes and sufficient causes are determined through a random process, they are all tracked and measured, unlike collected data where most pieces of information on component causes and sufficient causes are unknown and unmeasurable. Secondly, for specific exposures and outcomes, information from existing literature can be easily adopted into the simulation design. Thirdly, the simulation design can be adjusted to fit specific prior assumptions on the distributions and correlations among component causes and the exposure as well as compositions of sufficient causes. Hence it is possible to obtain estimates on the effects of the exposure under different prior assumptions.

5. Conclusion

This study demonstrates that even when all confounding factors are known and controlled for using conventional multivariate analysis, the observed association between exposure and outcome can still be dominated by residual confounding effects. An observed significant association apparently provides limited evidence for a causal relationship.

Appendix

Details of Steps 1 to 6 in Simulation Procedure

(i) $T_{i,n} = V_{i,n} P_i + U_{i,n}(1 - P_i)$ is a linear combination of a variable component ($V_i$) and a unique component ($U_i$) for each $i$, both being uniform [0,1] distributed random variables, and $P_i$ is a random proportion drawn from a uniform [0.3, 0.6] distribution. The range [0.3, 0.6] is chosen in order to set a low to moderate level of correlation among $T$. The mean (standard deviation), 25th, 50th, and 75th percentiles of the correlation coefficients for the matrix $T$ are 0.35 (0.13), 0.25, 0.33, and 0.45, respectively.

(ii) $G_{j,n}$ is set to 1 if $T_{i,n} > A_{j,i}$ and 0 otherwise, where $A_{j,i}$ takes on a random value drawn from a uniform [0.5, 0.9] distribution. The mean (standard deviation), 25th, 50th, and 75th percentiles of the correlation coefficients for the matrix $G$ are 0.16 (0.07), 0.11, 0.15, and 0.20, respectively.

(iii) $X_{i,n}$ is set to 1 if $T_{i,n} > 0.7$ and 0 otherwise, where 0.7 is the expected value of the uniform [0.5, 0.9] distribution. The mean (standard deviation), 25th, 50th, and 75th percentiles of the correlation coefficients for the matrix $X$ are 0.19 (0.07), 0.13, 0.17, and 0.24, respectively.
(2) Determine sufficient cause compositions and their components.

(i) Components for nine possible sufficient causes for \( Y \) are determined. Let \( C_{ij}, j = (1, 2, 3, \ldots, 9), \ i = (1, 2, 3, \ldots, 100) \) indicate whether \( T_i \) is a component of the \( j \)th possible sufficient cause: if \( T_i \) is component of the \( j \)th possible sufficient cause, \( C_{ij} = 1 \), and 0 otherwise. \( C_{ij} \) takes on a random value drawn from the Bernoulli distribution with probability of success \( H_i \), which is derived (rescaled) from a gamma distribution with both shape parameter and scale parameter equal to 1. For each sufficient cause if the components are less than 2, that is, for a given \( j \) if \( \sum_i C_{ij} < 2 \), then all components are redetermined through the same random process.

(ii) Determine whether a possible sufficient cause occurs. Let \( O_{jn} = 1 \) when all components for the \( j \)th possible sufficient cause become active or occur in the \( n \)th observation; that is, \( \sum_j O_{jn} C_{ij} = \sum_j C_{ij} \); otherwise \( O_{jn} = 0 \), \( i = (2, 3, \ldots, 100) \), \( j = (1, 2, \ldots, 9) \), and \( n = (1, 2, 3, \ldots, 50000) \).

(iii) Choose real sufficient causes from the nine possible sufficient causes. Let \( F_j, j = (1, 2, 3, \ldots, 9) \) denote whether the \( j \)th possible sufficient cause is a real sufficient cause for \( Y \). If the \( j \)th possible sufficient cause is a real sufficient cause, then \( F_j = 1 \) and 0 otherwise. \( F_j \) takes on a random value drawn from the Bernoulli distribution with probability of success 0.5. If there is no real sufficient cause assigned, that is, \( \sum_j F_j < 1 \), then the real sufficient causes for \( Y \) are redetermined through the same random process.

(3) Determine competing events. Let \( E_n \) denote the competing events for outcome \( Y \), \( n = (1, 2, 3, \ldots, 50000) \). \( E \) is a Bernoulli distributed random variable with a probability of success 0.001, value of success (competing events occurred) being 1, and value of failure (competing events not occurred) being 0. \( E \) is independent of \( X \).

(4) Determine small random errors for \( Y \). Let \( Q_n \) denote a small random error of \( Y \), \( n = (1, 2, 3, \ldots, 50000) \). \( Q \) is a Bernoulli distributed random variable with a probability of success 0.001, value of success being 1, and value of failure being 0. \( Q \) is independent of both \( E \) and \( X \).

(5) Determine the status of outcome \( Y \). Let \( Y_n = [0, 1] \), \( n = (1, 2, 3, \ldots, 50000) \) denote the outcome not occurred or occurred, respectively. Value of each \( Y_n \) is determined as follows. For each observation \( n \), \( Y_n = 1 \) (outcome occurred) if \( Q_n = 1 \), or for \( j = (1, 2, 3, \ldots, 9) \), \( \sum_j O_{jn} F_{jn} \geq 1 \) when \( E_{jn} = 0 \); otherwise \( Y_{jn} = 0 \) (outcome not occurred).

(6) Determine the known/suspected causal factors, in other words, potential confounding factors.

Let \( K_i, i = (2, 3, 4, \ldots, 100) \) denote the researcher’s knowledge (not necessary the fact) on \( X_i \) in relation to its confounding effect on the association between \( X_1 \) and \( Y \). \( K_i \) is a random value drawn from the Bernoulli distribution with a probability of success \( 0.1 + \sum_j C_{ij} F_j/10 \), value of success being 1, and value of failure being 0. \( \sum_j C_{ij} F_j \) is the total number of real sufficient causes that included \( X_i \) as a component.

**Conflict of Interests**

The authors declare that there is no conflict of interests regarding the publication of this paper.

**References**


Submit your manuscripts at http://www.hindawi.com