International Epidemiological Association

International Journal of Epidemiology, 2021, 1–7
doi: 10.1093/ije/dyab096
Education Corner

Education Corner

Reflection on modern methods: understanding bias and data analytical strategies through DAG-based data simulations

Chongyang Duan ,[1]* Anca D Dragomir,[2,3] George Luta[3,4] and Lutz P Breitling [5,6]

[1]Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou, China, [2]Department of Oncology, Georgetown University, Washington, DC, USA, [3]Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg, Frederiksberg, Denmark, [4]Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC, USA, [5]Department of Gastroenterology, Endocrinology, Metabolism, and Infectiology, Philipps University of Marburg, Marburg, Germany and [6]Department of Gastroenterology, Rheumatology, and Infectiology, Augsburg University Hospital, Augsburg, Germany

*Corresponding author. Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou, 510515, China. E-mail: donyduang@126.com

Editorial decision 6 April 2021; Accepted 15 April 2021

Abstract

Directed acyclic graphs (DAGs) are increasingly used in epidemiology to identify and address different types of bias. The present work aims to demonstrate how DAG-based data simulation can be used to understand bias and compare data analytical strategies in an educational context. Examples based on classical confounding situations and an M-DAG are examined and used to introduce basic concepts and demonstrate some important features of regression analysis, as well as the harmful effect of adjusting for a collider variable. Other potential uses of DAG-based data simulation include systematic comparisons of data analytical strategies or the evaluation of the role of uncertainties in a hypothesized DAG structure, including other types of bias such as information bias. DAG-based data simulations, like those presented here, should facilitate the exploration of several key epidemiological concepts, DAG theory and data analysis. Some suggestions are also made on how to further expand the ideas from this study.

Key words: Directed acyclic graphs, confounding bias, selection bias, teaching, simulation

Introduction

Directed acyclic graphs (DAGs), a tool that can be used for identifying suitable adjustment sets based on assumed causal relationships, are increasingly used in epidemiology.[1,2] An important advantage of this approach is that harmful adjustment sets, that is those that introduce rather than reduce bias, can be identified in a straightforward manner using DAGs, even for some scenarios that are difficult to deal with using other approaches.[3,4]

In brief, a DAG depicts the causal dependencies between nodes representing a treatment or exposure (say *X*), the outcome of interest (say *Y*) and covariables, by directed

Key Messages

- Directed acyclic graph (DAG)-based data simulations should facilitate the exploration of several key epidemiological concepts, DAG theory and data analysis.
- DAG-based data simulations are suitable for teaching confounding or selection bias with DAG-based approaches and corresponding applied regression analyses.
- DAG-based data simulations can be used for systematic comparisons of data analytical strategies or for the evaluation of the role of uncertainties in a hypothesized DAG structure.

arrows pointing from causes to effects.[1] An important concept is that of a 'backdoor path' defined as 'a non-causal path between treatment and outcome that remains even if all arrows pointing from treatment to other variables (the descendants of treatment) are removed'.[5] An open non-causal path will result in bias. Even though the study of confounding or selection bias using DAGs can be done following a simple set of rules,[1] computer tools employing various algorithms are currently available to assist with this task.[6,7]

The present study illustrates the use of DAG-based data simulations for demonstrating issues related to bias, and the use of data analytical strategies in an educational context. Examples of classical confounding situations as well as a particularly instructive and often used DAG are examined, to exemplify how this approach may be used to present the possibility of harmful adjustment and its impact on regression model estimates.

Methods

DAGs under consideration

In our study, we consider the five DAGs from Figure 1. The function to draw the DAGs used in the present work is included in the publicly available dagR package [cran.r-project.org/web/packages/dagR], and the five DAGs have been described in detail elsewhere including the following motivating real-life examples.[5]

DAG (a) shows a classical confounding situation in which the non-causal path $X \leftarrow C \rightarrow Y$ is open because of the common cause C. This is a very common confounding situation, with an example being the healthy worker bias. Since 'being physically fit' (C) is a cause of both being an active firefighter and having a lower mortality risk, the effect of working as a firefighter (X) on the risk of death (Y) will be confounded by 'being physically fit'.

DAG (b) shows confounding caused by the ancestor L of treatment X and the outcome Y sharing a common cause U. The non-causal path $X \leftarrow L \leftarrow U \rightarrow Y$ is open. An example of this situation is confounding by indication. The effect of taking aspirin (X) on the risk of stroke (Y) will be confounded by heart disease (L) since aspirin is more likely to be prescribed to individuals with heart disease, and heart disease and stroke share a common cause, namely atherosclerosis (U).

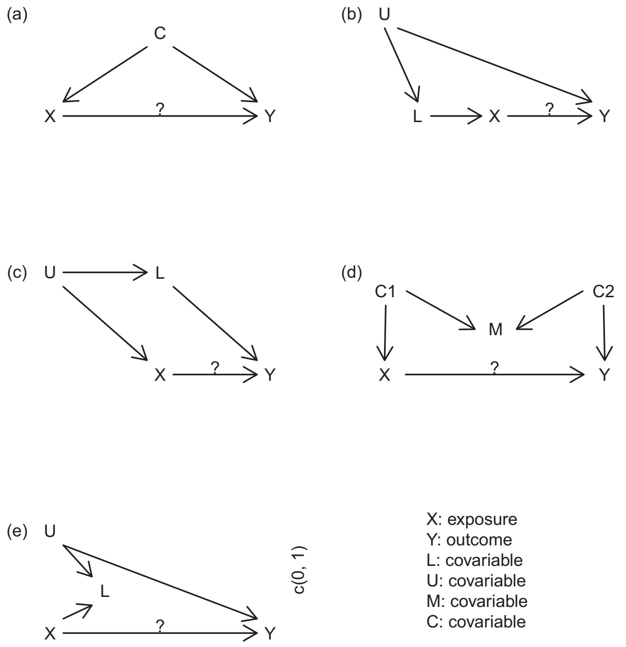

Figure 1 Directed acyclic graphs representing classical confounding [(a), (b) and (c)], and the M-structure frequently used for introducing harmful adjustment [(d) and (e)]

DAG (c) shows confounding that results from the treatment X and the ancestor L of outcome Y sharing a common cause U. The non-causal path $X \leftarrow U \rightarrow L \rightarrow Y$ is open. An example of this situation is confounding by reverse causation. The effect of exercise (X) on the risk of death (Y) will be confounded because exercise (X) is associated with cigarette smoking (L), a known risk factor for death. The association of exercise and cigarette smoking is caused by personality type (or social factors) (U).

DAG (d) is the M-DAG. It corresponds to a situation in which a so-called collider M exists that is causally influenced both by an ancestor of the exposure X and by an ancestor of the outcome Y. Adjusting for M will open the non-causal path $X \leftarrow C1 \rightarrow M \leftarrow C2 \rightarrow Y$ which would be harmful by introducing collider adjustment bias.

DAG (e) is another situation where the adjustment for a collider L will result in bias. For example, consider a study that aims to estimate the causal effect of being physically active (X) on the risk of cervical cancer (Y). A health-conscious personality ($C1$) affects both the possibility of taking a diagnostic test for pre-cancer (M) and of being physically active (X). A pre-cancer lesion ($C2$) affects the possibility of taking a diagnostic test for pre-cancer (M) and the risk of cervical cancer (Y). $C1$ does not affect Y, and $C2$ does not affect X. There is no confounding because there are no common causes of X and Y. However, in a study restricted to subjects taking a diagnostic test for pre-cancer (say, $M=1$), conditioning on a common effect of causes of treatment and outcome will open the non-causal path $X \leftarrow C1 \rightarrow M \leftarrow C2 \rightarrow Y$ and induce bias.

DAG-based data simulation

We will use simulated data to show the bias related to the above DAGs, and the analysis methods to remove the bias. The simulation function that we use is part of the latest version of the dagR package. After specifying the DAG, the coefficients describing the DAG arcs, and the parameters describing the distribution of the data to be simulated, the function simulates the requested number of observations based on the specified DAG structure. In our simulations, all continuous variables were simulated from the normal distribution, and all binary variables were simulated from the binomial distribution. All simulations and statistical analyses were done using R version 3.5.2. and dagR 1.1.3. The R code for the main figures is provided in Supplementary Material, available as Supplementary data at *IJE* online.

Results

Simulations for understanding the basic concepts

A simple example was simulated for each of the five DAGs shown in Figure 1, and linear regression models were used to show that biased results will be generated when confounding exists, and also ways to remove the bias. All variables considered in the DAGs were simulated from Normal $(0, 0.3^2)$ distributions, the effect of X on Y was set to be null, and the sample size n was set to be 100 000 (the sample size was large so that we could reduce the random error to a negligible level).

For DAGs (a), (b) and (c), if we ignore the confounding covariables [C for (a); L and U for (b) and (c)] and estimate the effect of X on Y without adjusting for these covariables, the results are biased (Table 1). The bias can be removed by adjusting for the specified covariate(s). For DAGs (d) and (e), the situation is fundamentally different. There is no confounding to start with, and we can estimate the effect of X on Y without adjusting for any covariables. If we thoughtlessly adjust for colliders [M for (d), and L for (e)], we obtain biased results (Table 1), although we can overcome this harmful adjustment by controlling for additional covariables.

Simulations for quantifying the bias

Classical confounding

DAG (a), representing the classical confounding situation, is shown in Figure 1. The first question we may ask is whether the bias is constant or proportional to the x-y-effect (the effect of X on Y which is defined as the true linear regression coefficient of X on Y). Figure 2a shows the impact of varying the true direct x-y-effect for the classical confounding scenario while keeping all other parameters constant. The raw regression estimates of the x-y-effect overestimated the x-y-effect by approximately a constant amount across all simulations, an indication of constant bias, regardless of the x-y-effect.

The second question may be how the bias will change if we increase the c-x-effect (the effect of C on X) or c-y-effect (the effect of C on Y). Our intuition may be that the bias will also increase. However, Figure 2b and c shows that the coefficients obtained by linear regression of Y on X had quite different patterns. When the effect of C on X increased, the estimated x-y-effect increased sharply and then decreased [panel (b)]. By contrast, when the effect of C on Y increased, the estimated x-y-effect increased linearly [panel (c)]. When we analysed the simulated data, we found that when we increased the c-x-effect or c-y-effect, the variance of X or Y will also be increased [panel (b)], with the co-increase of the c-x-effect and the variance of X resulting in the pattern of estimated x-y-effect from panel (b). If we increased the c-x-effect but at the same time reduced the noise of X to prevent the variance of X from changing, then the estimated x-y-effect increased linearly [panel (d)].

M-DAG

Simulations based on an M-DAG structure (Figure 1d) can be easily used to demonstrate the 'harmful' (biasing) effect of adjusting for M, and to show that additionally adjusting for either $C1$ or $C2$ generally removes that bias (with adjustment for $C1$ resulting in somewhat unstable, fluctuating estimates of the x-y-effect) (Figure 3). Based on the

Table 1 Simulation results for the five DAGs under consideration

DAGs	Path	The estimated effect of X on Y			
	Models and true parameters[a] $n = 100\,000$	Covariables adjusted for	Status of the path	Analysis results[b]	Estimated effect
(a)	Path: X←C→Y				
	$X = \alpha_x + \beta_{cx}C + e_x$	–	Open	0.194	Biased
	$Y = \alpha_y + \beta_{cy}C + \beta_{xy}X + e_y$	C	Closed	−0.004	Unbiased
	$\beta_{xy}=0,\ \beta_{cx}=\beta_{cy}=0.5$				
(b)	Path: X←L←U→Y				
	$L = \alpha_l + \beta_{ul}U + e_l$	–	Open	0.092	Biased
	$X = \alpha_x + \beta_{lx}L + e_x$	L	Closed	−0.005	Unbiased
	$Y = \alpha_y + \beta_{uy}U + \beta_{xy}X + e_y$	U	Closed	−0.002	Unbiased
	$\beta_{xy}=0,\ \beta_{ul}=\beta_{lx}=\beta_{uy}=0.5$	L, U	Closed	−0.004	Unbiased
(c)	Path: X←U→L→Y				
	$L = \alpha_l + \beta_{ul}U + e_l$	–	Open	0.100	Biased
	$X = \alpha_x + \beta_{ux}U + e_x$	L	Closed	0.003	Unbiased
	$Y = \alpha_y + \beta_{ly}L + \beta_{xy}X + e_y$	U	Closed	0.001	Unbiased
	$\beta_{xy}=0,\ \beta_{ul}=\beta_{ux}=\beta_{ly}=0.5$	L, U	Closed	0.003	Unbiased
(d)	Path X←C1→M←C2→Y				
	$M = \alpha_m + \beta_{c1m}C1 + \beta_{c2m}C2 + e_m$	–	Closed	−0.003	Unbiased
	$X = \alpha_x + \beta_{c1x}C1 + e_x$	M	Open	−0.037	Biased
	$Y = \alpha_y + \beta_{c2y}C2 + \beta_{xy}X + e_y$	M, C1	Closed	−0.004	Unbiased
	$\beta_{xy}=0,$	M, C2	Closed	−0.002	Unbiased
	$\beta_{c1m}=\beta_{c2m}=\beta_{c1x}=\beta_{c2y}=0.5$				
		M, C1, C2	Closed	−0.004	Unbiased
(e)	Path X→L←U→Y				
	$L = \alpha_l + \beta_{ul}U + \beta_{xl}X + e_l$	–	Closed	−0.006	Unbiased
	$Y = \alpha_y + \beta_{uy}U + \beta_{xy}X + e_y$	L	Open	−0.104	Biased
	$\beta_{xy}=0,\ \beta_{ul}=\beta_{xl}=\beta_{uy}=0.5$	U	Closed	−0.005	Unbiased
		L, U	Closed	−0.003	Unbiased

DAGs, directed acyclic graphs.

[a] The true effect is defined as the true linear regression coefficient β. The true effect of X on Y is β_{xy}, the others were defined in the same way. α is the intercept and e is the added noise. (All variables are normal, for more simulation details see Supplementary Table S1, available as Supplementary data at *IJE* online).

[b] Linear regression model was used to estimate the effect of X on Y, Y was the dependent variable, and X and the covariables were the independent variables.

analysis of the simulated data, Table 2 shows that biased estimates of the effect of *X* on *Y* resulted within each stratum of *M* [cases (b) and (c)] and there was no bias without stratification [case (a)]. If we further stratified the data by *C*1 [cases (d) to (g)] and *C*2 [cases (h) to (k)], the bias within each stratum was reduced to 0. It should be noted that within each stratum of *C*1, *X* was distributed within a very small range which resulted in the unstable estimation of the x-y-effect when adjusting for *C*1 (Supplementary Figure S1, available as Supplementary data at *IJE* online).

Binary *X* and *Y*

For binary outcome *Y* but normally distributed *X*, similar results were observed (Supplementary Results part 1 and Supplementary Figure S2, available as Supplementary data at *IJE* online). By contrast, for binary *X* with normal *Y* or binary *Y*, there was a monotone but nonlinear increase (for the classical confounding DAG) or decrease (for the M-DAG) of the bias with the increase of the confounder effect on *X*, regardless of whether we increase the prevalence of $X = 1$ with the confounder effect on *X* or if we keep it fixed (Supplementary Results part 1 and Supplementary Figures S3–S5, available as Supplementary data at *IJE* online).

Discussion

The results presented in this study show how DAG-based data simulations could be used to examine and demonstrate fundamental aspects of the application of DAGs to epidemiological data analysis. The consistent results from these easy-to-understand examples suggest that the DAG functions integrated into R may not only be useful for hands-on teaching of epidemiological applications of DAGs and related regression modelling, but also for quantifying the bias and harmful adjustment for DAGs that are more complex than

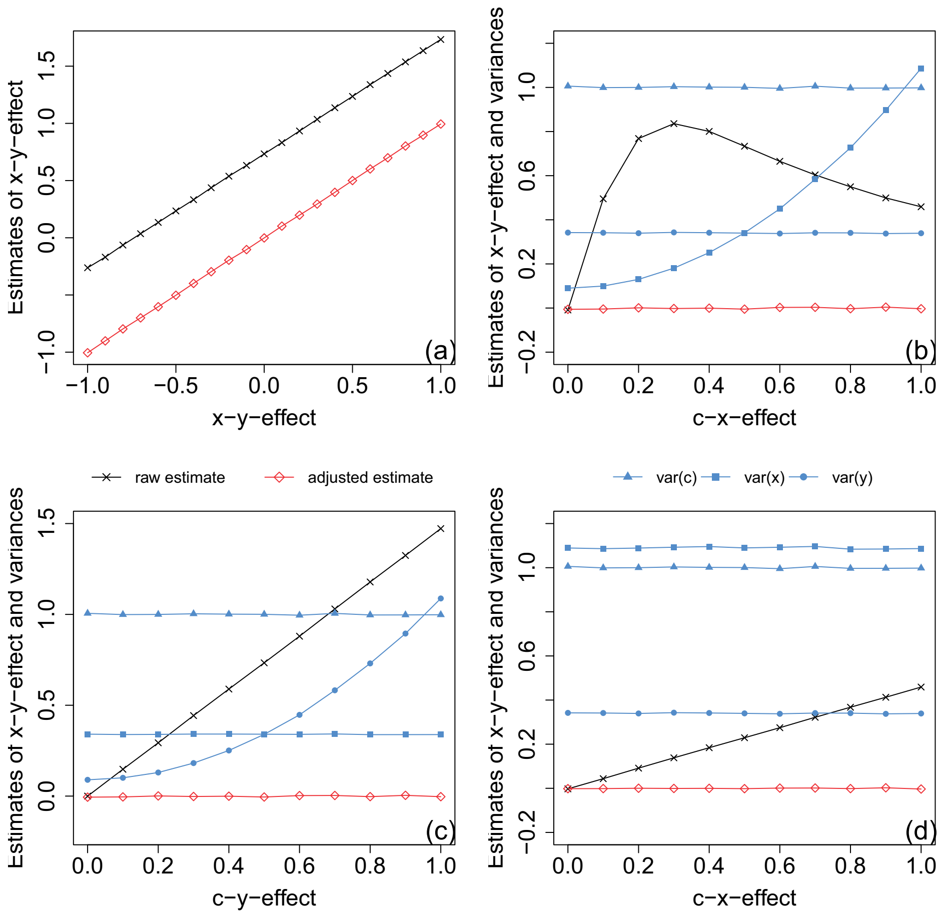

Figure 2 Estimated effects for the classical confounding DAG, depending on the simulated direct x-y-effect [panel (a)] and the simulated direct effect of confounder *C* on exposure *X* [panel (b) and panel (d)] or outcome *Y* [panel (c)]

Models used in the simulation: $X=\alpha_x+\beta_{cx}C+e_x$ and $Y=\alpha_y+\beta_{cy}C+\beta_{xy}X+e_y$, where β_{xy} is the x-y-effect, β_{cx} is the c-x-effect, β_{cy} is the c-y-effect, standard deviation of *C* is denoted as c-noise SD, standard deviation of e_x is denoted as x-noise SD and standard deviation of e_y is denoted as y-noise SD. For simulation details see Supplementary Table S2, available as Supplementary data at *IJE* online). Crosses joined by black line = regression estimates of x-y-effect without adjustment for *C*; diamonds joined by red line = regression estimates of x-y-effect adjusted for *C*; solid triangles joined by blue line = variance of *C*; solid squares joined by blue line = variance of *X*; solid circles joined by blue line = variance of *Y*

those investigated in this introductory paper. In addition to the confounding and selection bias discussed in this study, the DAG-based simulations can also be used to address other types of bias, such as information bias.

Some points to make on teaching classical confounding

Already the simulations based on the classical confounding DAG may serve to demonstrate some important points regarding regression analyses. In particular, decreasing the effect of the confounder on the outcome vs on the exposure affected very differently the regression estimation of the effect of *X* on *Y*. For normally distributed *X*, the variance of *X* plays an important role in the regression estimation of the effect of *X* on *Y*, and we need to consider the change in the variance of *X* when evaluating the effect on the bias of changing the effect of the confounder on the outcome or the exposure. For a real-life data analysis situation, if we want to evaluate the bias due to an unmeasured confounder C, the variance of *X* would be considered fixed, and the bias caused by C will increase linearly with the true effect of C on X. But for binary *X*, the story is different. The dependence of the variance of *X* on the prevalence of *X* presents a challenge when approaching this problem analytically.[8] Straightforward simulation results, on the other hand, show that the bias caused by *C* will increase monotonically with the true effect of C on X.

Introducing harmful adjustment

A reasonable starting point would be the Berkson's bias-like phenomenon which is at the core of the harmful effect

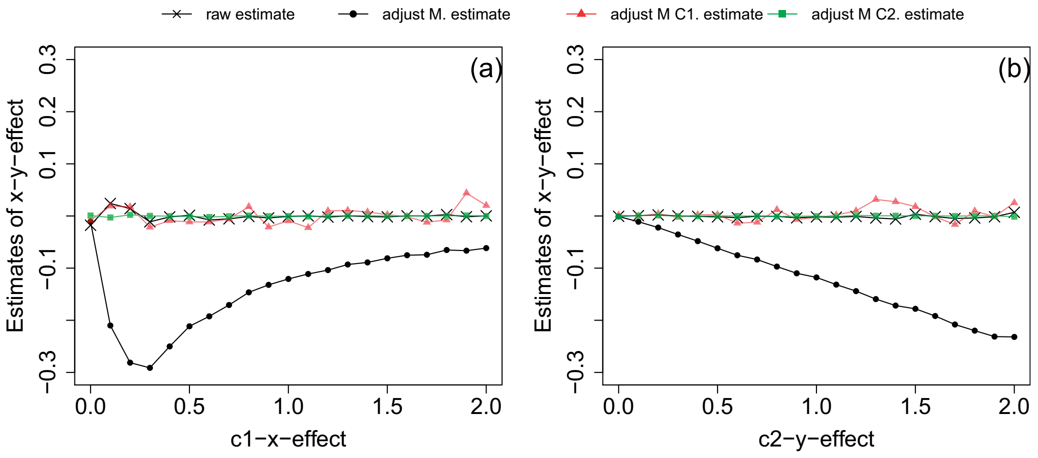

Figure 3 Estimated effects for the M-DAG, depending on the simulated direct effect of *C*1 on *X* [panel (a)], and *C*2 on *Y* [panel (b)]. Binary *C*1, *C*2 and *M* were simulated. Models used in the simulation were: $X{=}\alpha_x{+}\beta_{c1x}C1{+}e_x$, $Y{=}\alpha_y{+}\beta_{c2y}C2{+}\beta_{xy}X{+}e_y$ and logit(Prob($M{=}1$))$=\alpha_m{+}\beta_{c1m}C1{+}\beta_{c2m}C2$, where β_{xy} is the x-y-effect, β_{c1x} is the c1-x-effect, β_{c2y} is the c2-y-effect, β_{c1m} is the direct effects of *C*1 on *M*, β_{c2m} is the direct effects of *C*2 on *M*, the standard deviation of e_x is denoted as x-noise SD, and the standard deviation of e_y is denoted as y-noise SD. For simulation details see Supplementary Table S3, available as Supplementary data at *IJE* online). Crosses joined by black line =regression estimates of x-y-effect without harmful M-adjustment; solid circles joined by black line = regression estimates of x-y-effect with harmful M-adjustment; solid triangles joined by red line = regression estimates of x-y-effect adjusted for *M* and *C*1; solid squares joined by green line = regression estimates of x-y-effect adjusted for *M* and *C*2

Table 2 Simulation results for M-DAG structure

Estimated effect of X on Y (the true effect is null)[a]

Case	Stratum analysed	Status of the path	Analysis results[b]	Estimated effect
a)	–	Closed	−0.005	Unbiased
b)	M = 0	Open	−0.124	Biased
c)	M = 1	Open	−0.127	Biased
d)	C1 = 0 and M = 0	Closed	0.033	Unbiased
e)	C1 = 0 and M = 1	Closed	0.025	Unbiased
f)	C1 = 1 and M = 0	Closed	−0.053	Unbiased
g)	C1 = 1 and M = 1	Closed	−0.027	Unbiased
h)	C2 = 0 and M = 0	Closed	−0.001	Unbiased
i)	C2 = 0 and M = 1	Closed	−0.000	Unbiased
j)	C2 = 1 and M = 0	Closed	−0.001	Unbiased
k)	C2 = 1 and M = 1	Closed	0.001	Unbiased

M-DAG, directed acyclic graph with a collider *M* that is causally influenced both by an ancestor of the exposure *X* and by an ancestor of the outcome *Y*.

[a] Models used in the simulation: logit(Prob($M = 1$))$=\alpha_m{+}\beta_{c1m}C1{+}\beta_{c2m}C2$, $X = \alpha_x{+}\beta_{c1x}C1{+}e_x$, $Y = \alpha_y{+}\beta_{c2y}C2{+}\beta_{xy}X{+}e_y$. *M*, C1, and C2 are binary and *X* and *Y* are normal. The true effect of *X* on *Y* is β_{xy}, the others were defined in the same way. α is the intercept and *e* is the added noise. The simulated prevalences of C1, C2 and the collider *M* were 50%, β_{xy} was 0, β_{c1x} and β_{c2y} were 1, β_{c1m} and β_{c2m} were log(5), α_x and α_y were 0.5, and the SDs of the noise e_x and e_y added to *X* and *Y* were both 0.1.

[b] Stratification analysis was used to adjust the effect of *M*, C1 and C2. For each stratum, the linear regression model was used to estimate the effect of *X* on *Y*, *Y* was the dependent variable, and *X* was the independent variable.

of adjusting for a collider.[1] Although DAG papers such as[1] provide easy–to-understand numerical examples, simulations may be worthwhile to show that this is a rather universal phenomenon.

The performance of adjusting for either covariable in addition to the collider (i.e. adjustment sets {*M*, *C*1} or {*M*, *C*2}) can be demonstrated to be very similar for a wide range of parameter combinations. Relevant differences may only become apparent if the DAG-based simulated data involves very strong correlations, which are unlikely to be encountered with real-life epidemiological data. Whether performance differences are considered relevant, of course, often depends on subjective and subject matter considerations.

Limitations

Only simple DAGs and data analytical approaches were used in the present paper to avoid distracting the readers from the main issues of interest. As such, the results presented in this study could also be derived theoretically using linear regression theory. The advantage of the simulation-based approach is that there is no need for advanced knowledge of statistical theory. The dagR suite can readily accommodate more complex DAGs, and the simulation-based approach might also be feasible for situations in which the complexity of causal relationships renders current statistical approaches either impractical or impossible to use.

In conclusion, confounding and selection bias and related regression analyses are core topics of every comprehensive epidemiology course, and the inclusion of DAG methodology in such courses is becoming more and more common. Although the rules of DAGs are easy to teach, what the adjustment for different variables will do for a real dataset is often not intuitive to students, and so using simulation can help them understand the concepts better because they can see the data generated based on the assumed DAG. R is often the statistical software of choice, due to its free availability and wide use in relevant standard textbooks.[9–12] The analyses presented above indicate how DAG-based simulation might help introduce prospective epidemiologists to the use of DAGs for bias analysis and related regression model estimation using simulated data.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported by the National Natural Science Foundation of China [grant number 81803327].

Conflict of Interest

None declared.

References

1. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
2. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press, 2009.
3. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol* 2008;**8**:70.
4. Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL (eds). *Modern Epidemiology*. 3rd edn. Philadelphia, PA: Lippincott Williams & Wilkins, 2008, pp.183–209.
5. Hernán MA, Robins JM. *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC, 2020.
6. Knüppel S, Stang A. DAG program: identifying minimal sufficient adjustment sets. *Epidemiology* 2010;**21**:159.
7. Breitling LP. dagR: a suite of R functions for directed acyclic graphs. *Epidemiology* 2010;**21**:586–87.
8. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology* 2016;**27**:368–77.
9. Fox J. *Applied Regression Analysis and Generalized Linear Models*. New York, NY: SAGE Publications, 2015.
10. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. New York, NY: Springer, 2015.
11. Steyerberg EW. *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating*. New York, NY: Springer, 2009.
12. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer, 2013.