Using directed acyclic graphs to understand confounding in observational studies

Laura Dallolio[1], Rino Bellocco[2], Lorenzo Richiardi[3], Maria Pia Fantini[1] and the Causal Inference In Epidemiology (ICE) SISMEC Working Group[4]

[1] Department of Medicine and Public Health, University of Bologna, Bologna, Italy
[2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[2] Department of Statistics, University of Milano-Bicocca, Milano, Italy
[3] Cancer Epidemiology Unit, CeRMS and CPO Piemonte, University of Turin, Turin, Italy
[4] The SISMEC Working Group on Causal Inference In Epidemiology (ICE) is composed by: R. Bellocco, L. Richiardi, C. Pizzi, D. Zugna, S. A. Romio, M. P. Caria, A. Grotta, L. Dallolio, M. Maule, R. D'Amico, F. Barone-Adesi, M. Bonetti

Corresponding Author:
Laura Dallolio
Department of Medicine and Public Health
University of Bologna
Via San Giacomo 12
40126 Bologna Italy
Phone +39-051-2094832 Fax +39-051-2094839
e-mail: laura.dallolio@unibo.it

Summary
The goal of most epidemiological studies is to determine an unbiased estimate of the effect of being exposed to a given risk factor on a well defined outcome (disease, death) taking into account the effects of confounding.
However, it may not be entirely clear which confounders should be adjusted for in the analysis and which should not, even after using expert knowledge.
Recent developments in epidemiological theory have clearly shown that traditional methods of identifying and adjusting for confounding may be inadequate and so more recently the use of Directed Acyclic Graphs (DAGs) has been advocated.
DAGs are a useful graphical tool for encoding assumptions about causality and deciding apriori which variables require adjustment in the analysis and which not.
However, many clinical problems require complicated DAGs and therefore investigators may continue to use traditional practices because they are discouraged by the apparent complexity. Therefore, the purpose of this manuscript is to provide a simple overview on DAGs and how they can be used to select variables which require adjustment in the analysis.

KEY WORDS: *causal inference, confounding, direct acyclic graphs.*

Introduction

The modern theory of causal diagrams arose within the disciplines of computer science and artificial intelligence by Pearl (1) and Spirtes, Glymour and Scheines (2). The use of causal diagrams in epidemiology was first proposed by Greenland, Pearl, and Robins who showed how the use of such graphs can serve as a visual yet logically rigorous aid for summarizing assumptions in well defined epidemiological research hypotheses.

They have also demonstrated how the use of such graph can aid in planning data collection and analysis, in communicating results, and, with relevance to this paper, in avoiding subtle pitfalls in the selection of confounders (3, 4).

"Confounding is the problem of confusing or mixing of exposure effects with other "extraneous" effects: if at the time of its occurrence, exposure was associated with pre-existing risk for the outcome, its association would reflect at least in part the effect of this baseline

association, not the effect of exposure itself. The portion of the association reflecting this baseline association was called confounding" (5).

The factors responsible for confounding are called confounders; according to a standard textbook a confounder is traditionally defined as any variable that meets the following three necessary (but not sufficient or defining) characteristics:

1) "a confounding factor must be a risk factor for the outcome ";
2) "a confounding factor must be associated with the exposure under study in the source population";
3) "a confounding factor must not be affected by the exposure or the disease. In particular, it cannot be an intermediate step in the causal path between the exposure and the disease" (6).

Bias introduced by confounding could, in principle, produce an effect between exposure and outcome, or could cause an overestimate/underestimate of such effect. Ultimately it could be strong enough to reverse the true direction of the effect. The traditional approach to confounding is to "adjust for it", by including certain covariates in a multiple regression model or by stratification.

Recent developments in epidemiology have shown that traditional methods for identifying and adjusting for confounding (such as comparing adjusted and unadjusted effect estimates or the application of automatic variables selection procedures) may be inadequate (3, 7).

The traditional definition of confounder (a variable that is a risk factor for disease and is associated with exposure but not affected by exposure) has in fact some limitations.

One is that it applies only to the classical condition in which there is just one variable to consider (8). Another one is that, while every confounder satisfies all the three traditional criteria, some nonconfounders satisfy them as well. In other words the three traditional criteria for defining a confounder are necessary but not sufficient. In some cases, adjusting for such nonconfounders that meet the above definition is harmless, but in others it introduces bias (9).

Methods to aid in identifying sufficient sets of variables for control have been developed using graphical causal models (or Directed acyclic graphs-DAGs) (3, 10).

The strength of using DAGs is that traditional criteria of confounding usually agree with graphical criteria; that is, one should choose the same set of covariates for adjustment using either set of criteria. Nonetheless, there are cases in which the criteria disagree, and when they diverge, it is the conventional criteria that will fail (see the appendix for a better explanation of this concept).

In the following section, we illustrate the causal graph theory starting from an example of confounder identification using DAG to elucidate the hypothetical relationship between maternal alcohol use in pregnancy and low intelligent quotient (IQ) scores at age 5.

Causal graphs

Causal inference generally requires expert knowledge and untestable assumptions about the causal network linking exposure, outcome and other variables (10). A causal diagram can be constructed by abstracting the causal assumptions embedded in a narrative description of the hypothesized relations among the study variables.

To illustrate the idea, consider in Figure 1 a causal diagram illustrating the hypothetical relationship between high maternal alcohol use in pregnancy (AL) and low intelligence quotient (IQ) scores in childhood.

In estimating the causal effect of high prenatal alcohol exposure on IQ at age 5, we also consider the effects of socioeconomic status (SES), being born small for gestational age (SGA), smoking during pregnancy (SM) and other unmeasured factors (U).Unmeasured variables might include genetic factors associated with cognitive and behavioural outcomes in the mother which are in turn related to low IQ in the child and also environmental factors such as poor maternal rearing behaviours which are associated with poor quality of the postnatal environment leading to low IQ score at age 5.

Although residual confounding due to the unmeasured factors is of considerable importance in consideration of this association, a review of the literature suggests that such factors are extremely difficult to measure and therefore seldom measured.

For the purpose of this example we will ignore these factors and assume that they will not confound the estimate of interest.

In the terminology of causal diagrams, variables in the graph are called *nodes* or *vertices* and any line or arrow connecting two variables is called an *arc* or an *edge*. The arrows represent causal relations; whenever the arrow is lacking we assume that there is no direct causal relations.

A variable X affects a variable Y *directly* if there is an arrow from X to Y.

A variable X affects Y *indirectly* if there is a head to tail sequence of arrows from X to Y.

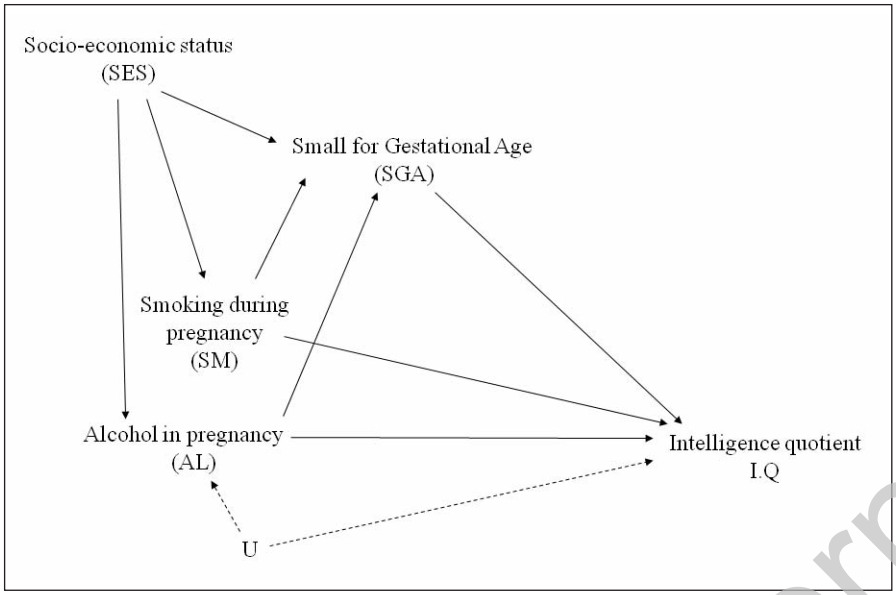

Figure 1. Causal diagram illustrating the hypothetical relationship between high maternal alcohol use in pregnancy and low intelligence quotient.

For example in Figure 1, SES affects SM directly and IQ indirectly.

A *path* between two variables is any noncrossing and nonrepeating sequence traced out along edges starting with a variable and ending with another one, regardless of the direction of arrowheads.

For example the succession of arrows between AL and SGA is called the path between AL and SGA.

Directed paths are the special case in which all the edges in the path flow head to tail.

Any other path is an *undirected path*.

In Figure 1, the path AL-SGA-IQ is directed, but AL-SES-SGA-SM-IQ is not.

A variable is said to be a *child* of another variable if it is caused by this variable, i.e in figure 1, AL is a child of SES and U; conversely SES and U are *parents* of A.

More generally, the *descendents* of a variable X are variables affected, either directly or indirectly, by X. Conversely, the *ancestors* of X are all variables that affect X directly or indirectly.

In Figure 1, SES has three children (AL, SM, SGA) and four descendents (AL, SM, SGA, IQ); and IQ has four parents (U, AL, SM, SGA) and five ancestors (U, AL, SM, SGA, SES). When tracing out a path, a variable on the path where two arrowheads meet is called a *collider* on that path ($\rightarrow C \leftarrow$). In Figure 1, SGA is a collider on the path from SM to SES.

A path is said to be *open* or *unblocked* or *active* unconditionally if there is no collider on the path. Otherwise, if there is a collider on the path, it is said to be *closed* or *blocked* or *inactive* and the collider blocks the path.

In Figure 1 the path AL-SES-SGA-SM-IQ is a blocked path because it collides at SGA, whereas the path AL-SES-SM-IQ is open.

Associations can be propagated only across a noncollider ($\rightarrow C \rightarrow$ or $\leftarrow C \rightarrow$) on a path. Metaphorically, it is possible to think of associations as water flowing through the graph: water can flow along some open (unblocked) paths but not along closed (blocked) paths (11). But the open and closed paths are switched by conditioning (stratifying) on the variable.

In other words, stratifying on a variable which is a noncollider closes the path (stratifying or conditioning is like to place a valve at the node that make it impossible for water to flow), whereas stratifying on a collider opens the path.

Then, associations can be propagated across a noncollider unless we do completely stratify on it, associations can be also transmitted across a collider if we stratify (condition) on it or a descendant of the collider itself.

It is not necessary to include all causes of variables in the diagram but it is important to include any key variables, otherwise causal graph interpretations can be severely misleading.

That is to say, selection of variables for modelling using causal graphs does not preclude the need to consider unmeasured confounders.

When variables have not been measured, it is helpful to denote pathways in and out of their node by dashed edges. In figure 1, U means the presence of unmeasured variables that cause both AL and IQ. All the graphs are

considered *acyclic* which means they contain no feedback loops; this means that if a variable X causes Y, Y cannot also cause X at the same moment (that is, no variable can cause itself). For this reason causal diagrams are also called directed acyclic graphs. Extension to time dependent variables can relax this assumption (12)**.**

Using DAGs to graphically represent confounding

One of the most attractive features of a causal graph is that it allows one to describe the causal structure that gives rise to confounding: in this setting, confounding is defined as the bias that arises when the exposure and the outcome share a common cause.

In graph theory, a path like E←C→D that links E and D through their common cause C is referred to as a *backdoor path* (Figure 2).

More specifically, undirected paths from E to D are termed back-door (relative to E) if they start with an arrow pointing into E (i.e, E←C).

Then confounding can be defined as the presence of a common cause (C) of the exposure E and the outcome D, or, equivalently, the presence of an unblocked back-door path between E and D.

To better understand this concept, let us consider Figure 2.

If the common cause C did not exist, then the only path between exposure and outcome would be E→D, and thus the entire association between E and D would be due to the causal effect of E on D. But the presence of the common cause C creates an additional source of association between the exposure E and the outcome D, which we refer to as confounding for the effect of E on D.

A back-door path is blocked if it contains a collider and unblocked if there is not a collider on the path. In figure 2 the path E-C-D is an unblocked back-door path.

In Figure 1 the path AL-SES-SGA-SM-IQ is a *blocked backdoor* path because it collides at SGA.

In contrast the path AL-SES-SGA-IQ is an *unblocked back-door* path because neither SES nor SGA are colliders on this path.

In a DAG all unblocked back-door paths are biasing paths.

In order to identify the causal effect of an exposure E on an outcome D, all the back-door paths between the two variables must be blocked.

In Figure 2 to identify the causal effect of E on D, the back-door path between the two variables must be blocked stratifying or conditioning on C.

We graphically represents the controlling (eg, regression adjustment, stratification, restriction) by placing a box around the controlled variable (Figure 3).

Conditioning on C closes the path and removes C as a source of association between E and D.

In other words stratifying on a single variable (which is a noncollider) is equivalent to removing that variable or node from the graph.

An important subtle idea arises when colliders are included in a set of stratifying variables because controlling for a collider can open biasing path.

As we report above the open and closed paths are switched by conditioning (stratifying) on the variable. This means that in Figure 4 if we seek to remove confounding by stratifying the population solely by C (thereby removing this node), a new pathway is inadvertently opened between E and D (the new pathway is indicated in Figure 4 as a dashed non-directional arc).

When, as in reality, there are complex DAGs, a simple graphical algorithm called the "*back-door criterion*" allows researchers to determine whether confounding exists and whether a set of measured variables

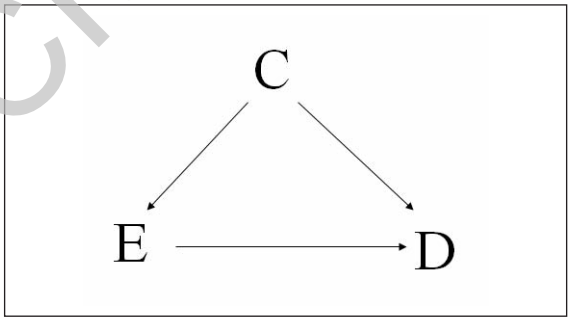

Figure 2. Causal graph indicating the presence of confounding due to an unblocked backdoor path from E to D.

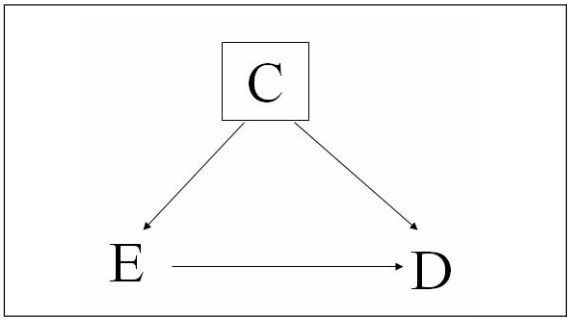

Figure 3. Causal graph indicating the absence of confounding after controlling for C.

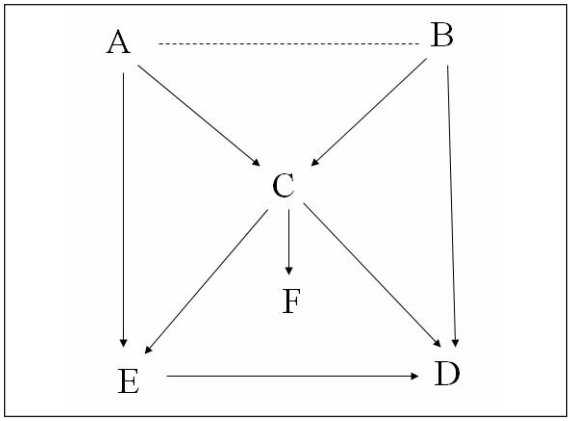

Figure 4. Causal diagram illustrating the consequence of conditioning for C.

S is sufficient to identify (validly estimate) the causal effect of X on Y.

To determine whether confounding exists, we follow two simple steps. Given any DAG:

1. Delete all arrows from exposure (that is exposure effects);
2. In this reduced graph, determine whether there is any unblocked backdoor path from the exposure to the outcome. If such a path exists, the causal relationship is confounded by the effects of the other variables; if there is no such path, there is no confounding.

Because now we understand the implication of stratification by a collider, it is possible to use a graphical algorithm for checking whether a set of variables is sufficient for adjustment.

This algorithm is known as the "*backdoor test for sufficiency*".

Given a subset S={S_1,…..S_n} of variables that contain no descendant of the exposure or the outcome, the steps are as follows:

1. Delete all arrows from exposure (that is exposure effects);
2. Draw undirected arcs to connect every pair of variables that share a child that is either in S or has a descendant in S (that is, put in all arcs generated by control of S);
3. In the new graph, determine whether there is any unblocked backdoor path from the exposure to the outcome that avoids passing through any node in the set of stratification factors. If no such path is found, confounding is controlled by the proposed factors; if there is such a path, stratification by these factors is not sufficient to remove all confounding.

Applying the "back-door criterion" to Figure 4, we can easily identify if there is confounding between E and D. Because there are three unblocked back-door paths (E-A-C-D, E-C-B-D and E-C-D), it is necessary to adjust.

The backdoor test for sufficiency allows us to answer the question: what is the smallest subset from the covariates A, B and C that would be sufficient for adjustment to estimate the effect of E on D?

C alone is not sufficient because after removing the arrow coming from E and after linking every pair of variables that share a child or a descendant in S (in this case the set of variables S is only C), a new unblocked backdoor path E-A-B-D is created.

While S={C, A}, or S={C, B} are sufficient.

Applying the backdoor test algorithms to the DAGs in Figure 1, we are now able to answer the research question: is there any confounding in the association between alcohol in pregnancy and IQ? In order to apply the "back-door criterion" to Figure 1, we have to delete all lines emanating from AL, then check if there is any unblocked backdoor path from the exposure to the outcome.

After deleting all exposure effects, there are four unblocked backdoor paths: AL-SES-SM-IQ, AL-SES-SM-SGA-IQ, AL-SES-SGA-IQ, AL-U-IQ (see Figure 5). Because there is confounding, we could decide to control for SES.

The "backdoor test for sufficiency" allow us to answer to the question: is conditioning on SES sufficient to estimate the causal effect of alcohol in pregnancy (AL) on intelligent quotient (IQ) scores at age 5?

Given S={SES}and according to step 2 of the "backdoor test for sufficiency" we should add undirected arcs connecting every pair of variables which share a child that is either in SES or has a descendant in SES (step 2). Nevertheless, as SES is not a child or a descendant of any other variables, we do not need to add undirected arcs. Finally we have to determine if there is any unblocked backdoor path from AL to the IQ that avoids passing through S (step 3).

As any path is found (because after conditioning on SES, the open paths AL-SES-SM-IQ, AL-SES-SGA-IQ, AL-SES-SGA-SM-IQ have been blocked), adjustment for SES alone is sufficient to remove all confounding (apart the confounding due to U that remains unknown) (Figure 6).

Note that, while it is sufficient to adjust for SES alone, it is not sufficient to control only for SM or SGA. It is also correct to control for {SES, SGA}, {SES, SM}, {SGA, SM} or {SES, SM, SGA}.

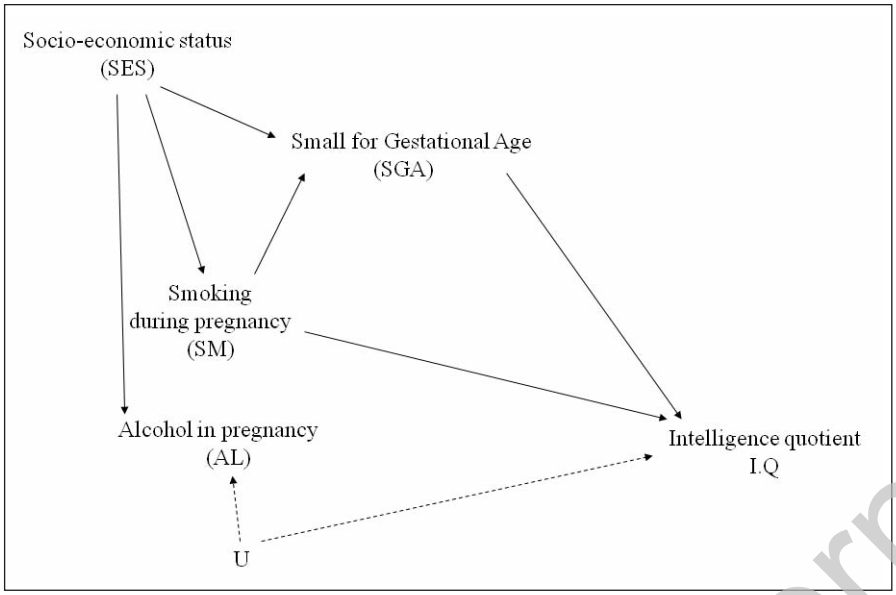

Figure 5. Back-door criterion applied to the DAG of the relationship between maternal alcohol use in pregnancy and low intelligent quotient (IQ) scores at age 5.

Knowing that confounding can be removed when we control solely for SES it is particularly valuable in case it is difficult to collect information on SGA or SM.

Conclusions

Directed acyclic graph models were developed as a theory of statistical causal inference.
We reckon, as Dawid warns (13), that there is more and more in need of explicit, methodological and philosophical justification to use them to explain causal relationships.
But we believe they are a powerful tool to help researchers to choose which covariates should be included in traditional statistical approaches in order to minimize the magnitude of the bias in the estimate produced (14). Furthermore, they are a graphical tool to display the web of causation that is not captured by statistical conventional models.
Although this manuscript is limited to how to select vari-

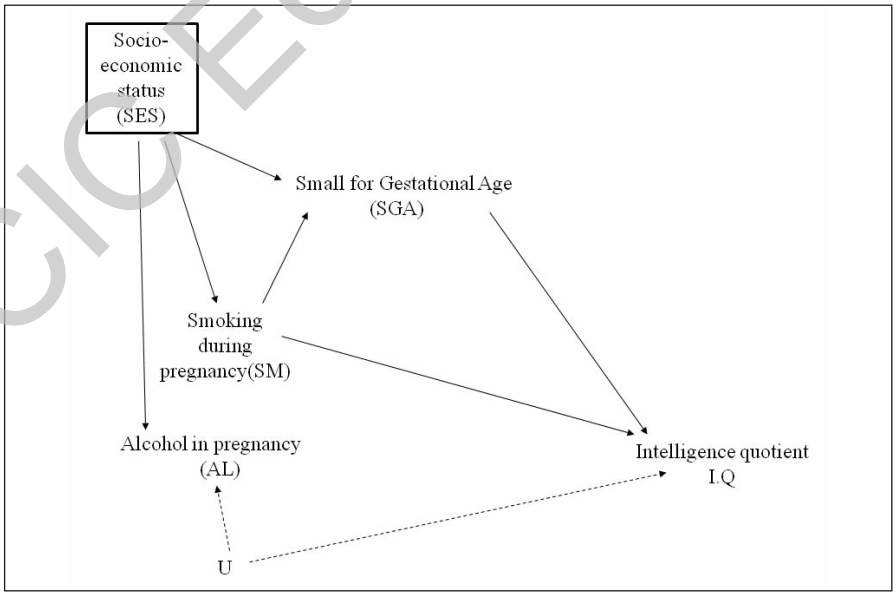

Figure 6. Back-door test for sufficiency applied to the DAG of the relationship between maternal alcohol use in pregnancy and low intelligent quotient (IQ) scores at age 5.

ables which require adjustment using back-door criterion, DAGs can be also used for distinguishing and reasoning about selection bias (15).

As in fact Greenland pointed out, another advantage of using causal diagrams is they are the easiest way to remember the logic of biases.

Critical points are the difficulty of reporting the structure of the effect modification and the absence of quantification of the associations involved (i.e. if X is a strong cause of Y and a weak cause of Z, this information is lost in a DAG in which Y and Z are simply children of X).

Though causal diagrams are a useful tool to think conceptually about a causal inference problem, there is a need for research in which DAGs and quantitative approach are explored together.

Acknowledgments

We thank Ron Gray for his important contribution to this paper and Silvia Candela, Nicola Caranci, Paola Rucci and Elisa Stivanello for their helpful comments to previous drafts of this article.

References

1. Pearl J. Causality: Models, Reasoning and Inference. Cambridge University Press 2009.
2. Spirtes P, Glymour C and Scheines R. Causation, Prediction, and Search, 2nd Edition, The MIT Press, 2000.
3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48
4. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology. 2001;12(3):313-20.
5. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. Epidemiol Perspect Innov. 2009 Sep 4;6:4.
6. Rothman KJ, Greenland S, Lash TL, Modern Epidemiology, 3rd Ed Lippincott Williams & Wilkins March 2008
7. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol. 2002 Jan 15;155(2):176-84.
8. Maldonado G, Greenland S. Estimating causal effects. International Journal of Epidemiology 2002; 31:422-429.
9. Glymour M. M, Greenland S. Causal Diagrams. In Modern Epidemiology, 3rd Ed Lippincott Williams & Wilkins March 2008
10. Hernan MA, Robins J. Causal Inference available in http://www.hsph.harvard.edu/faculty/miguel-hernan/files/HernanRobinsJuly09.pdf
11. Jewell N. P. Statistics for Epidemiology. Chapman & Hall/CRC 2004.
12. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000 Sep;11(5):550-60.
13. Dawid A. Philip Beware of the DAG! available in http://clopinet.com/isabelle/Projects/reading/Dawid_NIPS08_causality_preprint.pdf
14. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol. 2008 Oct 30;8:70.
15. Richiardi L, Barone-Adesi F, Merletti F, Pearce N. Using directed acyclic graphs to consider adjustment for socioeconomic status in occupational cancer studies. J Epidemiol Community Health. 2008 Jul;62(7):e14.

Appendix

Why conventional rules for confounding are not always reliable

In Figure 7 we reported an example from the Chapter

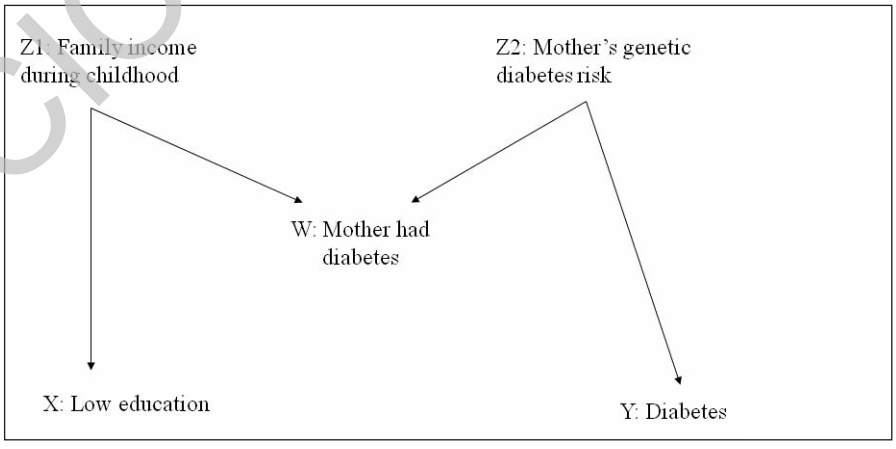

Figure 7. A DAG under which traditional confounder-identification rules fail

"Causal Diagrams" in Modern Epidemiology to illustrate why conventional rules for confounding are not always reliable.

The Figure reflects the assumptions that maternal diabetes is associated with the subject's education via the common cause Z_1 "family income" (the reasoning is that if a subject was poor as a child, his or her mother was poor as an adult, and this poverty also increased the mother's risk of developing diabetes). So, the first traditional requirement for confounding is satisfied.

Maternal diabetes is associated with the subject's diabetes via the common cause Z_2, the genetic factor. The second traditional criteria is satisfied. Maternal diabetes W is not affected by exposure X or outcome Y. The third traditional requirement is satisfied.

According to the traditional definition, to estimate if educational attainment affects the risk of type II diabetes, we should adjust for W.

Differently, according to the graphical criteria, we do not need to adjust for mother's diabetes because the path between X to Y is already blocked at W and then it is not a biasing path.

Conditioning on W alone opens the confounding path $X\text{-}Z_1\text{-}W\text{-}Z_2\text{-}Y$, in this sense adjustment for W would be one form of overadjustment.