Use of causal diagrams to inform the analysis and interpretation of observational studies: An example from the Study of Heart and Renal Protection (SHARP)

N Staplin PhD[1], WG Herrington MD[1], P Judge MBChB[1,2], CA Reith FRCP[1], R Haynes DM[1,2], MJ Landray PhD[1], C Baigent FRCP[1,2], J Emberson PhD[1]

1. Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, UK
2. Medical Research Council Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, UK

Correspondence to:

Dr. Natalie Staplin

Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU),

Nuffield Department of Population Health,

Richard Doll Building, Old Road Campus,

Roosevelt Drive, Oxford OX3 7LF, UK

Email: natalie.staplin@ndph.ox.ac.uk

Phone: +44 1865 743975

Fax: +44 1865 743985

Running title: Causal diagrams in observational studies

Word count:

Abstract: 225

Main body: 3749

Abstract

Observational studies often seek to estimate the causal relevance of an “exposure” to an “outcome” of interest. However, many possible biases can arise when estimating such relationships, in particular bias due to confounding. To control for confounding properly, careful consideration of the nature of the assumed relationships between the exposure, the outcome and other characteristics is required. Causal diagrams provide a simple graphical means of displaying such relationships, describing the assumptions made and allowing for the identification of a set of characteristics that should be taken into account (ie, “adjusted for”) in any analysis. Furthermore, causal diagrams can be used to identify other possible sources of bias (such as selection bias) which, if understood from the outset, can inform the planning of appropriate analyses. In this article, we review the basic theory of causal diagrams and describe some of the methods available to identify which characteristics need to be taken into account when estimating the net effect of an exposure on an outcome. In doing so, we review the concept of “collider bias” and show how it is inappropriate to adjust for characteristics that may be influenced, directly or indirectly, by both the exposure and outcome of interest. A motivating example is taken from the Study of Heart and Renal Protection, in which the relevance of smoking to progression to end-stage renal disease is considered.

1. Introduction

In longitudinal analyses of observational studies (including observational analyses done using trial data) a range of possible biases can arise when assessing the relevance of a particular characteristic ("exposure") to a particular disease ("outcome"). Uppermost among these is bias due to confounding, which may be thought of as a spurious statistical association between the exposure and the disease that arises (wholly or partly) because of some other exposure that is associated with a change in both the exposure of interest and the probability of disease. Statistical methods to control for confounding include adjustment (eg, in a regression model) and stratification (eg, performing analyses separately in subsets of individuals with similar characteristics). While it may be thought that increasing levels of adjustment for characteristics not believed to be involved in the "causal pathway" could only result in an improved estimate of the true aetiological relevance of the exposure to the disease (ie, a reduction in bias due to confounding), in fact, the opposite may be true. Depending on the nature of the relationship between the exposure of interest and the potential confounders, inappropriate adjustment for some of them can lead to the introduction of a bias that didn't previously exist(1).

An understanding of "causal diagrams" aids the identification of possible sources of bias in epidemiological analyses (including confounding, reverse causality and selection bias) and therefore facilitates the identification of an appropriate set of characteristics that would need to be taken into account during any statistical analysis. The purpose of this paper is to provide such an understanding, so that the methods can be considered and applied more widely by those who wish to perform such analyses.

2. A hypothetical example

Consider the following hypothetical example. A conference is attended by 500 delegates. All delegates arrive the day before the plenary session and half have flown long-haul to travel to the conference. Now suppose that everyone who flies long-haul will suffer, to some degree, from jet lag the following day. A drinks reception is held on the evening prior to the morning plenary session, at which a choice of alcoholic and non-alcoholic drinks are available. Suppose the effect of drinking alcohol on jet lag is of interest and the overall 2x2 contingency table of the relationship is as shown in panel A of Table 1 (from which an odds ratio [OR] of 1.00 is estimated, which makes sense since jet lag is a result of flying long-haul rather than alcohol consumption at the conference reception). But, now suppose that this analysis had been done separately among those who had attended the plenary session and those who did not attend. On the face of it, this seems reasonable because both drinking alcohol and being jet lagged may be associated with missing the plenary session. After stratifying by whether the delegate attended the plenary session, the contingency tables could plausibly look like those in panel B of Table 1. The estimated Mantel-Haenszel OR (which takes into account the stratification) is now 0.63 (95% confidence interval (CI) 0.42-0.93), which suggests, quite wrongly, that drinking alcoholic rather than non-alcoholic drinks at the conference reception has a strongly *protective* effect on the risk of jet lag.

The problem is caused by inappropriate stratification by whether the delegate missed the plenary session. In this case, stratification for missing the plenary session introduces bias into the assessment of the relationship between drinking alcohol and jet lag, since missing the plenary session could be influenced by both alcohol consumption and jet lag (i.e. is a "collider"; see below). To understand how this happens, and how it can be avoided, one needs to appreciate the basic theory of causal diagrams.

3. An introduction to causal diagrams

A detailed description of the theory of causal diagrams in epidemiological studies can be found elsewhere(2, 3); only a basic introduction is provided here. We start with some terminology.

A causal diagram is a graphical tool that enables the visualisation of the relationships between the “exposure of interest”, the outcome being studied, and all other characteristics (variables) that are associated in some way with at least two other variables in the diagram. It encodes the assumptions underlying the epidemiological analysis, and should include all relevant variables (even if they haven’t been measured). Causal diagrams therefore comprise a set of variables (“nodes": often represented by letters) with arrows drawn between them to show the *directions* of the assumed causal relationships. No other assumptions about the nature of these relationships are made however (for example, whether the exposure increases or decreases the value of the outcome, or the magnitude of such an effect). The lack of an arrow between a pair of variables therefore represents the assumption that there is no direct relationship between them.

In a causal diagram, a “cause" is a variable that influences, either directly or indirectly, the value of another variable. Causes are often referred to as “ancestors" of the other variable, with direct causes (see Figure 1) called “parents". By contrast, “effects" are variables that are influenced, either directly or indirectly, by another variable. Effects are often referred to as “descendants", with direct effects (see Figure 1) called “children". For example, consider the relationships between the five variables A—E shown in Figure 1. Looking at variable E, it has A, B and D as causes (or ancestors), but only B is a parent of E. Similarly, the variable A has B, C and E as effects (or descendants), but only B is a child of A. A “path” is said to exist between two variables in a causal diagram if they can be joined through a sequence of single headed arrows, irrespective of their direction, possibly passing through other variables on the way. The paths of most interest in epidemiological studies are generally the “causal

pathways", which are paths starting at the exposure and ending at the disease that *do* follow the direction of the arrows. Directed acyclic graphs (DAGs) are a special form of causal diagram which do not contain any "directed cycles". That is, in a DAG, it is not possible to connect any variable through a path of single headed arrows, when following the direction of the arrows, back to itself. (Time ordering is important in causal diagrams; any causes must precede effects, so no variable can be both the cause and effect of any other variable in the graph.) However, that is not to say that DAGs cannot be used to represent "bi-directional" biological relationships where feedback loops exist, providing the temporality is preserved (eg, see Figure 2 for how a bi-directional relationship between systolic blood pressure and estimated glomerular filtration rate could be represented in a DAG). (Note that the analysis of data with time-varying treatments or exposures may require specialist statistical methods to be employed, such as marginal structural models (4, 5).)

When considering a single exposure of interest (E), an outcome of interest (D) and a single covariate (C) that is associated in some way with both the exposure and the outcome, there are three possible sets of relationships (Figure 3). The first scenario (Figure 3a) is that the covariate C is a cause of both the exposure and the outcome (ie, it is a "confounder"). The second scenario (Figure 3b) is that the covariate lies along the causal pathway from the exposure to the outcome. In this case, C is called an "effect mediator". The final situation (Figure 3c) is where the covariate is a common effect of both the exposure and the outcome. In this case, C is called a "collider".

Paths can be characterised into "causal pathways" and "non-causal pathways". Since the causal pathways represent the associations of interest, they should typically remain "open" in any statistical analysis. (An open path in a DAG is merely a path along which an association can be transmitted.) By contrast, if a path is "blocked" then no association can be transmitted along it. Consequently, it is typically the goal to perform a statistical analysis that blocks all non-causal pathways. If not, then the open non-causal pathway that remains

is referred to as a “biasing” pathway. An association (or a possible association that is to be tested) between two variables requires at least one open path between the variables. An open path can become blocked, and vice versa, by conditioning on particular variables along that path (“conditioning” typically means statistical adjustment [eg, through a regression analysis] or stratification [as in the hypothetical example]). The rules in Table 2 can be used to establish whether a particular causal or non-causal path in a given DAG is open or blocked(6, 7).

Consider now the confounder C_{con} shown in Figure 3a, the effect mediator C_{med} in Figure 3b and the collider C_{col} shown in Figure 3c. The rules in Table 2 can be used to determine whether the paths from E to D through C in these cases are open and hence whether they are biasing pathways. In Figure 3a, the path $E \leftarrow C_{con} \rightarrow D$ is open (rule 2) and is therefore a biasing pathway. Conditioning on C_{con} blocks the path (rule 2) and provides an unbiased estimate of the effect of E on D. In Figure 3b, the path $E \rightarrow C_{med} \rightarrow D$ is also open, but as this path is a causal pathway it should be kept open, assuming the goal is to estimate the total effect of E on D. Conditioning on C_{med} in this situation would block this path (rule 1) and result in estimation only of the *direct* effect of E on D (that is, the part that is not mediated through C_{med}). In Figure 3c, the path $E \rightarrow C_{col} \leftarrow D$ is already blocked (rule 3a) and there are no other non-causal pathways in the DAG. However, if C_{col} were conditioned on, then the path $E \rightarrow C_{col} \leftarrow D$ would be opened (rule 3a) creating a new biasing pathway that didn’t previously exist.

These rules allow us to establish why the estimate of the (hypothetical) causal effect of alcohol consumption on jet lag in Section 2 was so clearly wrong. In this case, missing the plenary session is a collider, as it could have been caused either by alcohol consumption or by suffering from jet lag. Conditioning on it (by creating the two separate 2x2 contingency tables) opened up a new biasing pathway.

4. Selecting appropriate sets of covariates for observational analyses

Generally, DAGs drawn to represent the relationship between the exposure and outcome in an epidemiological study will be much more complex than those shown in Figure 3, as there will be many inter-relationships between the various covariates, exposure and outcome, making it difficult to identify all the potentially biasing pathways. While a simple 6-step algorithm exists allowing one to test whether or not a proposed set of covariates to control for removes all biasing pathways(8), applying this can require considerable trial and error. Fortunately, freely available software exists to implement these rules making the selection of an appropriate set of covariates much easier (for example http://www.dagitty.net/ (9)and R package dagR(10)). It is also worth noting that the theory of DAGs can be combined with traditional selection methods(11, 12) to see whether there are any unnecessary covariates that could be removed from the set and therefore possibly improve the precision of the estimated effect of exposure on outcome.

5. A real example: adjustment for baseline albumin:creatinine ratio in analyses of the relevance of smoking to progression to end-stage renal disease

Now that the theory of causal diagrams has been introduced, we shall apply it to a real epidemiological analysis to illustrate the type of problem that can result if causal relationships are not properly considered. The example chosen is an analysis of the relevance of smoking to progression to end-stage renal disease (ESRD) among 6245 patients who were not on dialysis at the time of randomization into the Study of Heart and Renal Protection (SHARP), a large randomized trial of lipid-modification in patients with established renal disease(13).

Interest lies in estimating the relevance of smoking to ESRD and the question is whether or not to adjust (ie, condition) for each participant's urinary albumin-to-creatinine ratio (ACR). Smoking increases ACR(14, 15) and higher ACR concentrations are associated with

increased risk of progression to ESRD(16, 17). But, the causal nature of the association between ACR and ESRD is uncertain. While it is possible that ACR directly influences (ie, causes) ESRD, it is also possible that the association between ACR and ESRD is due to other (unknown) factors that affect both ACR and the risk of reaching ESRD (in which case ACR would merely be a marker of renal progression). These scenarios are illustrated by the DAGs drawn in Figure 4. (In the latter case, exposure to these unknown factors would have to *increase* ACR as well as increase ESRD risk for higher ACR to be associated with increased ESRD risk.) For simplicity, we ignore in Figure 4 all covariates that could otherwise confound the relationship between smoking and ESRD risk and assume that adjustment has already been made for any covariates needed to close any biasing paths between smoking and ESRD.

If Figure 4a represented the truth, then ACR would be an effect mediator and it would be inappropriate to condition on ACR because that would block a causal pathway between smoking and ESRD. (If it were done, the effect of this pathway would be removed from the estimate of the total effect of smoking on ESRD.) If Figure 4b represented the truth, however, then ACR would be a collider on one of the two paths between smoking and ESRD, and, as such, it would still be inappropriate to adjust for it (because doing so would create a biasing pathway between smoking and ESRD). Specifically, if it were known that someone had an increased level of ACR (the impact of conditioning) then knowing that they were also a smoker would reduce the probability that they have been exposed to the other unknown factors. This is because being a smoker is the more likely cause of their increased value of ACR. In other words, given ACR, smokers will be systematically less likely to have been exposed to the other unknown factors (which, of course, cannot be controlled for in the statistical analysis) and their risk of progression to ESRD will, as a consequence, apparently be reduced. Conditioning on ACR in this situation would therefore bias the relative risk (RR) of ESRD associated with smoking downwards and could even result in a RR that (wrongly) suggests smoking has a protective effect on the risk of progression. Another scenario would

be that ACR has some direct effect on the risk of ESRD with other unknown factors influencing both ACR and ESRD (i.e. a combination of the scenarios in Figures 4a and 4b, which is shown in Figure 4c). In this case, it would still be inappropriate to adjust for ACR (as ACR is both an effect mediator and a collider on another pathway between smoking and ESRD).

In SHARP, the aetiological relevance of baseline smoking status (current smokers compared to never smokers) to ESRD was estimated using Cox proportional hazards regression(18). After adjustment for a range of assumed confounders (see the assumed DAG in Figure 5a) current smokers had a similar rate of progression to ESRD as never smokers (hazard ratio [HR] 1.02, 95% CI 0.89 – 1.17) (In DAGs, variables that are "conditioned on" are usually highlighted by drawing a square around them, as shown in the figure). However, if adjustment is also made for ACR (and other factors associated with smoking and ESRD risk: blood pressure, body mass index, current drinking and renal status) rather than only the characteristics thought to be confounders, an apparent *protective* association between current smoking and ESRD risk is observed (HR 0.85, 95% CI 0.74 – 0.98). The DAG in Figure 5b illustrates what has happened. Many of the causal pathways (shown in green in figure 5a) are now blocked as all of these additional factors were assumed to be effect mediators. In addition, a new biasing pathway (shown in red) has arisen by conditioning on ACR, as it is a collider on the path: smoking at entry into SHARP $\rightarrow$ ACR $\leftarrow$ U $\rightarrow$ ESRD. Inclusion of the additional factors (in particular albuminuria) creates what we believe to be an artificial association between smoking and a *reduced* risk of ESRD, which is unlikely to reflect a true protective effect.

6. Index event bias in studies of individuals with chronic kidney disease

Causal diagrams allow sources of bias related to the selection of participants into a study to be identified. One such bias that is particularly relevant to studies of chronic kidney disease

(CKD) populations is index event bias. While index event bias is well recognised in the epidemiological literature(19, 20), it is perhaps not sufficiently acknowledged in epidemiological studies of already diseased individuals. It is usually discussed in the context of recurrent events, as paradoxical findings are often observed where well-established risk factors for a disease may seem not to influence recurrence risk amongst individuals who have already had a first event. As progression to ESRD amongst those with CKD is a worsening of a pre-existing condition, studies examining causes of ESRD in CKD populations could suffer from similar issues.

Consider for example the causal diagram in Figure 6. The effect of an exposure measured on entry into the study (E) on progression to ESRD (D) is of interest. The variable S represents the selection criteria for the study (which in this case would be having CKD). Only including individuals that meet this criterion in the study would be the same as conditioning on S. As progression to ESRD is a worsening of pre-existing kidney disease, it is not unreasonable to assume that any risk factors for ESRD will also be risk factors for CKD, and hence will be causally related to the selection criteria. The variable U in Figure 6 represents all such risk factors, which may or may not have been recorded in the study. Prior values of the exposure (P) will influence both the risk of CKD and the value of the exposure at time of entry into the study, introducing an association between E and S. The variable S is a collider on the path from E to D that goes through U, so conditioning on this variable has opened up a biasing pathway between the exposure and the disease. If both the exposure and the unmeasured risk factors are associated with an increased risk of kidney disease, then conditioning on S (which is unavoidable) will create an inverse association between E and U. The consequence of this would be to bias the association between the E and D towards the null.

Adjustment for the shared risk factors U would be required to control for index event bias, although it is always possible that residual bias may exist as some of the risk factors may be unmeasured (or even unknown). However, formulae do exist to estimate the extent of such

residual biases(21, 22) based on assumptions about the distributions and associations of unmeasured factors.

There is an additional source of bias in the causal diagram in Figure 6 that should also be mentioned. As P represents the long-term or “usual” exposure level, this variable could have an effect on progression to ESRD that is not completely explained by the measured baseline value (for example if this value is measured with error). In this case, the biasing pathway $E \leftarrow P \rightarrow D$ may be thought of as representing regression dilution bias, which can easily be accounted for using other established methods(23, 24)

7. Discussion

Sufficiently large properly randomized trials are generally the preferred method of testing the causal effect of an exposure on an outcome of interest. However, epidemiological analyses are still often used to estimate causal effects (either to generate hypotheses to be tested in future trials or when trials are not feasible). For observational data, it is necessary to adjust for variables that could confound the causal effect. Causal diagrams provide a graphical representation of all the assumptions that have been made about the associations between the exposure, the outcome and any other variables (which should be informed by evidence from the literature as well as expert opinions). Based on those assumptions, simple automated methods are available which will identify the appropriate set of variables to include in a statistical analysis, the aim of which is to minimize bias due to observed confounders. Of course, biases may still exist due to unmeasured confounders of the exposure and the outcome or imprecise measurements of confounders (ie, residual confounding). While causal diagrams cannot prevent this residual confounding, they can be used to determine whether the recorded characteristics are likely to be sufficient to adequately adjust for it. Indeed, it may be more likely that the validity of the no residual confounding assumption will be considered if a causal diagram has been constructed.

More generally, causal diagrams allow other potential sources of bias in a study to be recognised, such as index event bias discussed in Section 6. There are many other examples of bias related to the selection of participants in a study or “selection bias”. For instance, the “healthy-participant” effect is already well established in prospective cohort studies as a potential bias when estimating disease prevalence reliably. Selection bias and its representation using DAGs has been discussed in detail elsewhere(3). While this article focuses on the use of causal diagrams in observational analyses, they can also be used to detect potential sources of bias in randomized controlled trials, for example, when participants are unblinded to treatment allocation(25). One limitation of causal diagrams is that they require the direction of causality between two variables to be stated, which may be difficult to establish when multiple biological measurements are conducted at the same time.

It is worth noting that the use of causal diagrams is still appropriate even where there is uncertainty about some of the underlying relationships between the characteristics considered. In such situations, alternative causal diagrams can be considered as sensitivity analyses. Whether explicitly specified or not, *every* epidemiological analysis makes assumptions about the underlying causal relationships between the exposure, outcome and other characteristics. The use of causal diagrams merely states these assumptions and, in doing so, helps avoid potential pitfalls through inappropriate adjustments. Causal diagrams should therefore be considered at all stages when embarking on such analyses, and made available alongside the results to make it clear which assumptions have been made.

Acknowledgements

We thank the participants in SHARP and the local clinical centre staff, regional and national coordinators, steering committee, and data monitoring committee.

Disclosures

Support: SHARP was funded by Merck/Schering-Plough Pharmaceuticals (North Wales, PA), with additional support from the Australian National Health Medical Research Council, the British Heart Foundation, and the UK Medical Research Council. SHARP was initiated, conducted, and interpreted independently of the principal study funder (Merck & Co. and Schering Plough Corp, which merged in 2009). The Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), which is part of the University of Oxford, receives core funding from the UK Medical Research Council, the British Heart Foundation and Cancer Research UK. CTSU has a staff policy of not accepting honoraria or other payments from the pharmaceutical industry, expect for the reimbursement of costs to participate in scientific meetings.

Financial Disclosure: The authors declare that they have no other relevant financial interests.

REFERENCES

1. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol.* 2010;39:417-20.
2. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37-48.
3. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology.* 2004;15:615-25.
4. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA. Methods for dealing with time-dependent confounding. *Statistics in Medicine*. 2013;32(9):1584-618.
5. Fewell Z HM, Wolfe F, Tilling K, Choi H, Sterne JAC. Controlling for time-dependent confounding using marginal structural models. *The Stata Journal*. 2004;4(4):402-20.
6. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *Am J Epidemiol.* 2002;155:176-84.
7. Williamson EJ, Aitken Z, Lawrie J, Dharmage SC, Burgess JA, Forbes AB. Introduction to causal diagrams for confounder selection. *Respirology.* 2014;19(3):303-11.
8. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol.* 2008;8:70-84.
9. Textor J, Hardt J, Knuppel S. DAGitty A Graphical Tool for Analyzing Causal Diagrams. *Epidemiology.* 2011;22:745.
10. Breitling LP. dagR: a suite of R functions for directed acyclic graphs. *Epidemiology.* 2010;21:586-7.
11. Weng HY, Hsueh YH, Messam LLM, Hertz-Picciotto I. Methods of Covariate Selection: Directed Acyclic Graphs and the Change-in-Estimate Procedure. *Am J Epidemiol.* 2009;169:1182-90.
12. Evans D, Chaix B, Lobbedez T, Verger C, Flahault A. Combining directed acyclic graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology. *BMC Med Res Methodol.* 2012;12:156-70.
13. Baigent C, Landray MJ, Reith C, et al. The effects of lowering LDL cholesterol with simvastatin plus ezetimibe in patients with chronic kidney disease (Study of Heart and Renal Protection): a randomised placebo-controlled trial. *Lancet.* 2011;377:2181-92.
14. Halimi JM, Giraudeau B, Vol S, et al. Effects of current smoking and smoking discontinuation on renal function and proteinuria in the general population. *Kidney Int.* 2000;58:1285-92.
15. Pinto-Sietsma SJ, Mulder J, Janssen WM, Hillege HL, de Zeeuw D, de Jong PE. Smoking is related to albuminuria and abnormal renal function in nondiabetic persons. *Ann Intern Med.* 2000;133:585-91.
16. Hemmelgarn BR, Manns BJ, Lloyd A, et al. Relation Between Kidney Function, Proteinuria, and Adverse Outcomes. *JAMA*. 2010;303:423-9.
17. Gansevoort RT, Matsushita K, van der Velde M, et al. Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes. A collaborative meta-analysis of general and high-risk population cohorts. *Kidney Int.* 2011;80:93-104.
18. Staplin N, Haynes R, Herrington WG, et al. Smoking and Adverse Outcomes in Patients With CKD: The Study of Heart and Renal Protection (SHARP). *Am J Kidney Dis*. 2016. doi: 10.1053/j.ajkd.2016.02.052.
19. Dahabreh IJ, Kent DM. Index event bias as an explanation for the paradoxes of recurrence risk research. *JAMA*. 2011;305:822-3.
20. Smits LJ, van Kuijk SM, Leffers P, Peeters LL, Prins MH, Sep SJ. Index event bias-a numerical example. *J Clin Epidemiol.* 2013;66:192-6.
21. Greenland S. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology.* 2003;14:300-6.
22. Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Ann Epidemiol.* 2008;18:637-46.

23. Clarke R, Shipley M, Lewington S, Youngman L, Collins R, Marmot M, et al. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *Am J Epidemiol.* 1999;150(4):341-53.
24. Frost C, Thompson SG. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *J Roy Statist Soc Ser A.* 2000;163:173-89.
25. Shrier I. Estimating causal effect with randomized controlled trial. *Epidemiology.* 2013;24(5):779-81.

Table 1: Hypothetical example illustrating the association between drinking alcohol at the conference reception and suffering from jet lag the following day, before and after stratification for whether the delegate attended the plenary session.

	Jet lag	
	Yes	No
Panel A		
Drank alcohol at reception		
Yes	100	100
No	150	150
Panel B		
Delegate missed plenary session		
Drank alcohol at reception		
Yes	80	60
No	100	30
Delegate attended plenary session		
Drank alcohol at reception		
Yes	20	40
No	50	120

Odds ratio for panel A = $\frac{100/100}{150/150} = 1.$

Mantel-Haenszel odds ratio for panel B $= \frac{(30*80)/270+(120*20)/230}{(100*60)/270+(50*40)/230} = 0.63.$

Table 2: Rules to establish whether a pathway in a causal diagram is "open" or "blocked".

Rule Number	Type of path	Number of colliders on path	Aim of adjustment	How to establish if path is open or blocked
1	Causal	N/A	Leave open (for estimation of total causal effect)	The path will be open providing that no variables along it are conditioned on. (Otherwise, it will be blocked.)
2	Non-causal	None	Block non-causal pathway	The path will be blocked if *at least one* variable along it is conditioned on. (Otherwise, it will be open.)
3a	Non-causal	One	Block non-causal pathway	The path will be open if the *only* variable conditioned on is the collider‡. (Otherwise, it will be blocked.)
3b	Non-causal	More than one	Block non-causal pathway	The path will be open if *all* the collider variables‡ (*and* no non-collider variables) are conditioned on. (Otherwise, it will be blocked.)

‡ Or descendant(s) of the collider(s).

Figure legends

Figure 1: Causal diagram to illustrate causes, ancestors and parents, and effects, descendants and children.

Figure 2: Example of how to represent a possible b-directional relationship in a DAG. SBP = systolic blood pressure. eGFR = estimated glomerular filtration rate.

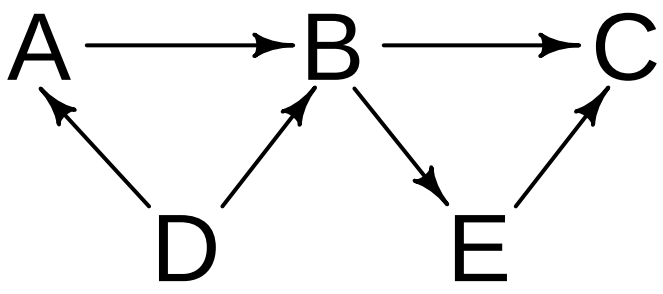
Figure 3: Causal diagrams showing possible underlying relationships for a covariate (C) that is associated with both the exposure of interest (E) and the outcome of interest (D)

Figure 4: Causal diagrams that represent three possible relationships between smoking, ESRD and ACR in SHARP. ESRD = end−stage renal disease. ACR = albumin:creatinine ratio. SHARP = Study of Heart and Renal Protection.

Figure 5: Causal diagram showing assumed associations between baseline smoking status, ESRD and baseline characteristics in SHARP. ESRD=end-stage renal disease. ACR=urinary albumin:creatinine ratio. BP=blood pressure. BMI=body mass index. U=other unknown factors. *Age, sex, ethnicity, country and education would also be causes of BMI, current drinking, BP, renal status and ACR. $Prior diseases would also be causes of renal status and ACR. Boxes around variables indicate that they have been adjusted for in analyses. Open causal pathways are highlighted by green arrows (e.g. smoking status at entry into SHARP $\rightarrow$ ACR $\rightarrow$ ESRD in panel a)) and biasing pathways are indicated by red arrows (e.g. smoking status at entry into SHARP $\rightarrow$ ACR $\leftarrow$ U $\rightarrow$ ESRD in panel b)).

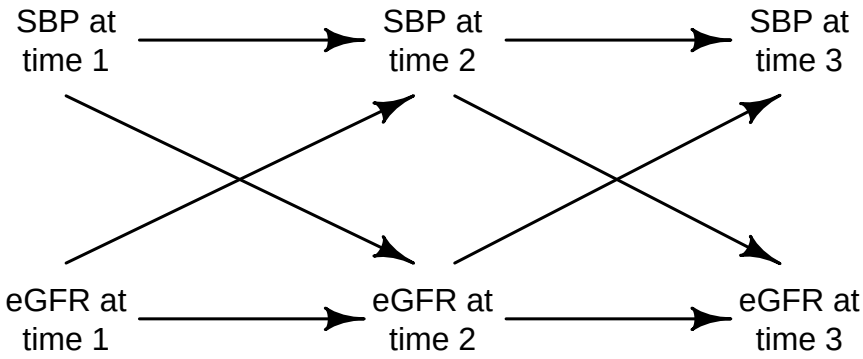
Figure 6: Causal diagram to illustrate issue of index event bias in observational analyses restricted to participants with chronic kidney disease

Figure 1: Causal diagram to illustrate causes, ancestors and parents, and effects, descendants and children.



Variable	All causes (ancestors)	Direct causes (parents)	All effects (descendants)	Direct effects (children)
A	D	D	B, C, E	B
B	A, D	A, D	C, E	C, E
C	A, B, D, E	B, E	−	−
D	−	−	A, B, C, E	A, B
E	A, B, D	B	C	C

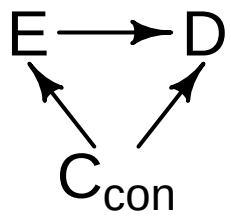
Figure 2: Example of how to represent a possible bi−directional relationship in a DAG



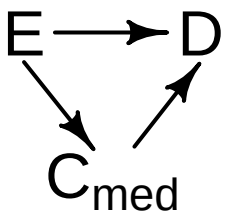
SBP = systolic blood pressure.
eGFR = estimated glomerular filtration rate.

Figure 3: Causal diagrams showing possible underlying relationships for a covariate (C) that is associated with both the exposure of interest (E) and the outcome of interest (D)

a) C is a 'confounder'



b) C is an 'effect mediator'



c) C is a 'collider'

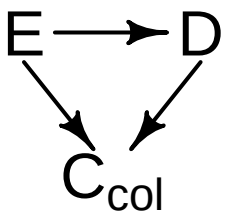
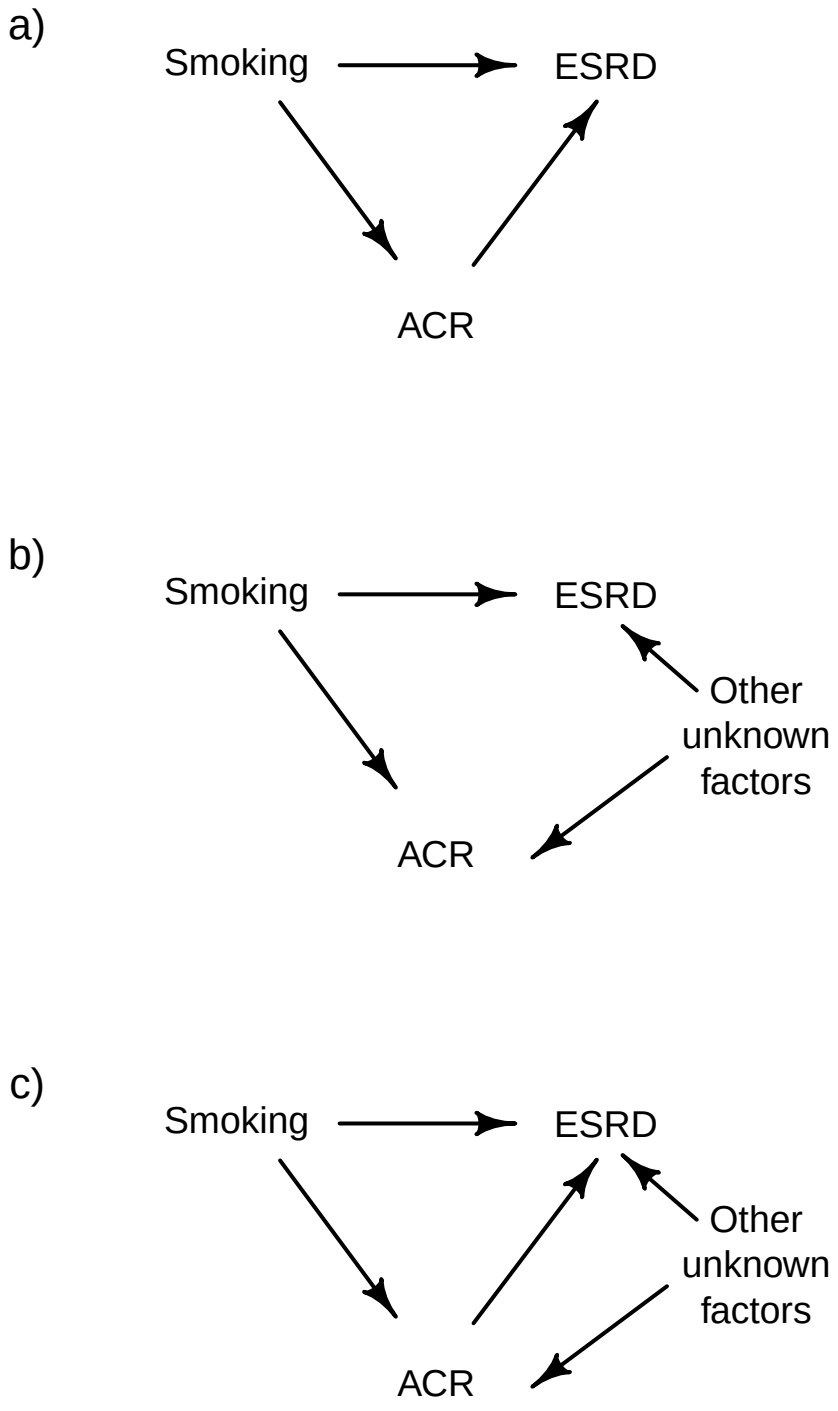


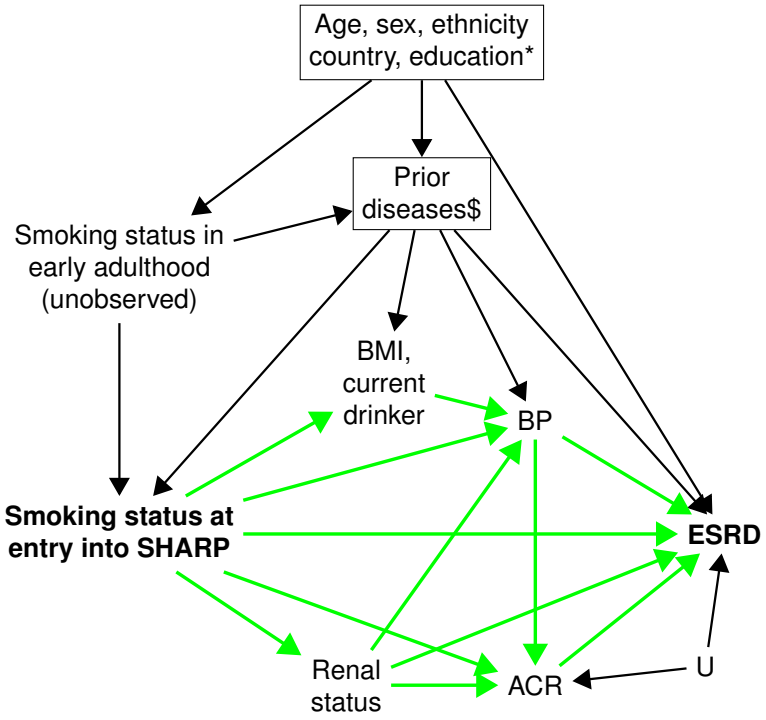
Figure 4: Causal diagrams that represent three possible relationships between smoking, ESRD and ACR in SHARP



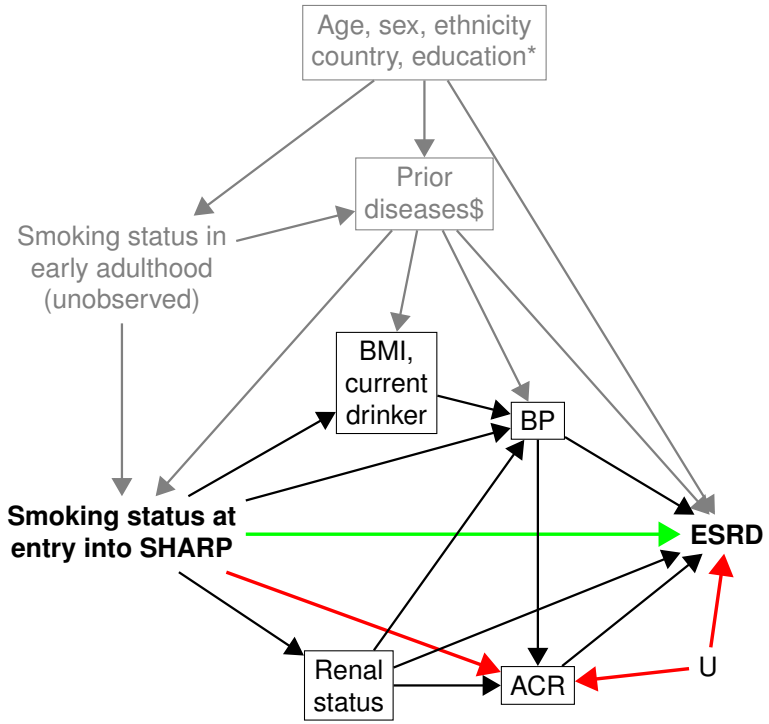
ESRD = end−stage renal disease.
ACR = albumin:creatinine ratio.
SHARP = Study of Heart and Renal Protection.

Figure 5: Causal diagram showing assumed associations between baseline smoking status, ESRD and baseline characteristics in SHARP

a) Adjustment for variables considered to be confounders keeps all causal pathways open and blocks all non-causal pathways

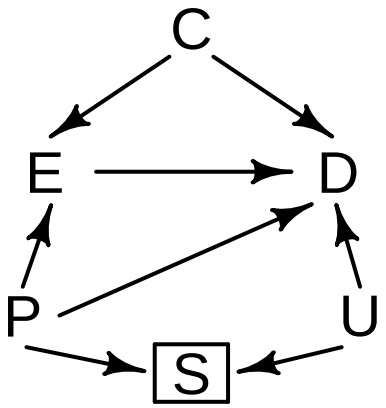


b) Adjustment for effect mediators and colliders blocks causal pathways and creates a biasing pathway



ESRD=end-stage renal disease. ACR=urinary albumin:creatinine ratio. BP=blood pressure. BMI=body mass index. U=other unknown factors. *Age, sex, ethnicity, country and education would also be causes of BMI, current drinking, BP, renal status and ACR. $Prior diseases would also be causes of renal status and ACR. Boxes around variables indicate that they have been adjusted for in analyses. Open causal pathways are highlighted by green arrows (e.g. smoking status at entry into SHARP -> ACR -> ESRD in panel a)) and biasing pathways are indicated by red arrows (e.g. smoking status at entry into SHARP -> ACR <- U -> ESRD in panel b)).

Figure 6: Causal diagram to illustrate issue of index event bias in observational analyses restricted to participants with chronic kidney disease.



E = exposure at study entry

D = progression to end−stage renal disease

C = confounders

P = long−term (or "usual") values of exposure prior to study entry (unmeasured)

S = selection criteria (i.e. having chronic kidney disease)

U = risk factors for kidney disease (possibly unmeasured)