Official Journal of the Asian Pacific Society of Respirology
Respirology

INVITED REVIEW SERIES:
MODERN STATISTICAL METHODS IN RESPIRATORY MEDICINE
SERIES EDITORS: RORY WOLFE AND MICHAEL ABRAMSON

Introduction to causal diagrams for confounder selection

ELIZABETH J. WILLIAMSON,[1,2,3] ZOE AITKEN,[4] JOCK LAWRIE,[3,5] SHYAMALI C. DHARMAGE,[2] JOHN A. BURGESS[2] AND ANDREW B. FORBES[1,3]

[1]*School of Public Health and Preventive Medicine, Monash University,* [2]*Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne,* [3]*The Victorian Centre for Biostatistics (VICBiostat),* [4]*Centre for Women's Health, Gender and Society, Melbourne School of Population and Global Health, The University of Melbourne, and* [5]*Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Melbourne, Victoria, Australia*

ABSTRACT

In respiratory health research, interest often lies in estimating the effect of an exposure on a health outcome. If randomization of the exposure of interest is not possible, estimating its effect is typically complicated by confounding bias. This can often be dealt with by controlling for the variables causing the confounding, if measured, in the statistical analysis. Common statistical methods used to achieve this include multivariable regression models adjusting for selected confounding variables or stratification on those variables. Therefore, a key question is which measured variables need to be controlled for in order to remove confounding. An approach to confounder-selection based on the use of causal diagrams (often called directed acyclic graphs) is discussed. A causal diagram is a visual representation of the causal relationships believed to exist between the variables of interest, including the exposure, outcome and potential confounding variables. After creating a causal diagram for the research question, an intuitive and easy-to-use set of rules can be applied, based on a foundation of rigorous mathematics, to decide which measured variables must be controlled for in the statistical analysis in order to remove confounding, to the extent that is possible using the available data. This approach is illustrated by constructing a causal diagram for the research question: 'Does personal smoking affect the risk of subsequent asthma?'. Using data taken from the Tasmanian Longitudinal Health Study, the statistical analysis suggested by the causal diagram approach was performed.

Key words: causal inference, confounding, directed acyclic graph, observational study.

Abbreviations: SES, socioeconomic status; TAHS, Tasmanian Longitudinal Health Study.

INTRODUCTION

In respiratory health research, we often wish to investigate possible causal relationships between exposures and health outcomes. For example, we might wish to know the effect of individuals' smoking behaviour on their risk of subsequent asthma. Using data from a cohort study, we can estimate this effect from the observed association between personal smoking and subsequent asthma (e.g. the odds or risk ratio for asthma comparing smokers and non-smokers) provided that smokers and non-smokers do not differ in terms of other characteristics associated with the risk of subsequent asthma; otherwise, the estimated exposure effect will be biased. This bias, typically called confounding, is usually dealt with by adjusting for the differing characteristics, provided that they are measured, in a multivariable regression model.[1] Alternatively, if few variables need to be adjusted for, stratification into subgroups defined by these characteristics can be performed. However, in order to apply either of these statistical analyses, we

Correspondence: Andrew B. Forbes, School of Public Health and Preventive Medicine, The Alfred Centre, 99 Commercial Road, Melbourne, Vic 3004, Australia. Email: andrew.forbes@monash.edu

The Authors: Dr Elizabeth J. Williamson is a biostatistician with research interests focusing on methods for causal inference. Ms Zoe Aitken is an epidemiologist working on studies examining the social determinants of health inequalities, with particular interests in epidemiological and quantitative research methodologies. Dr Jock Lawrie is a postdoctoral statistician with research interests in causal inference, experimental design and spatio-temporal modelling. Professor Shyamali C. Dharmage is a clinical epidemiologist and leads the Allergy and Lung Health Unit in the Centre for Epidemiology and Biostatistics at the Melbourne School of Population and Global Health. Dr John A. Burgess is a physician and epidemiologist with research interests in asthma and allergic diseases. Professor Andrew B. Forbes is Head of the Biostatistics Unit, School of Public Health and Preventive Medicine. His research interests are in biostatistical methodology applied to practical problems and collaborative epidemiological and clinical research.

Received 22 November 2013; accepted 3 December 2013.

must first decide which measured variables need to be adjusted for in order to remove confounding, insofar as that is possible using the available data. This variable selection process is often called confounder selection.

A common approach to confounder selection is to apply a stepwise selection procedure. This approach is not recommended for many reasons but particularly because it is based on *P*-values alone.[2] A popular alternative is to use the change-in-estimate criterion where a variable is considered to be a confounder if its omission from a regression model changes the estimated exposure effect by more than a prespecified threshold.[3] A third approach defines confounders as variables which are: (i) associated with the exposure in the source population; (ii) associated with the outcome among the unexposed; and (iii) not on the causal pathway. All of these confounder selection strategies can, in certain circumstances, lead to an increase—rather than the expected decrease—in confounding bias when adjusting for the selected variables.[4,5] Various authors have stressed that background knowledge of causal structures is required for confounder selection; criteria based on statistical association alone are insufficient.[2,5]

In this paper we discuss an alternative approach to confounder selection using causal diagrams, also called directed acyclic graphs. Causal diagrams utilize assumptions regarding the underlying causal relationships between relevant variables to perform confounder selection rather than relying on observed statistical associations.[6–10] We use a causal diagram approach to consider the question of whether personal smoking affects the risk of subsequent asthma. We create a causal diagram for this question based on background knowledge of the underlying causal structure and demonstrate how our diagram can be used to decide which variables must be adjusted for in a multivariable analysis in order to remove confounding bias. We apply the statistical analysis suggested by the causal diagram approach using data from the Tasmanian Longitudinal Health Study (TAHS).

CAUSAL DIAGRAMS

A causal diagram, also known as a causal directed acyclic graph, is a representation of the underlying causal relationships relevant to the research question. Variables, or characteristics, are represented by nodes. Arrows between nodes represent causal effects, depicting the existence—but not the strength—of causal relationships. Causal diagrams contain only unidirectional (single-headed) arrows; bidirectional or undirected arrows cannot be included. They are acyclic, meaning that following a series of arrows in the indicated direction cannot lead back to the original node because variables cannot cause themselves. Finally, a causal diagram must contain all variables that have a causal effect on two or more other variables included in the diagram, even if unmeasured in the dataset.

In our discussion of causal diagram theory, we will assume that the dataset being used for the statistical analysis is so large that we can ignore random error (sampling variability), allowing us to focus on systematic sources of confounding bias in selecting variables to adjust for in the statistical analysis. In our analysis (Section 'Estimating the effect of personal smoking on adult asthma: the TAHS data'), we acknowledge the role of sampling variability by calculating 95% confidence intervals and *P*-values.

Three simplified causal diagrams

To perform confounder selection using a causal diagram, we first need to create a diagram that we believe captures all the causal relationships relevant to our research question. Figure 1 shows three causal diagrams, each representing a different set of causal assumptions that we could propose for the question of whether personal smoking affects subsequent asthma. We note that we believe the true causal scenario to be much more complex (see Fig. 4 for our proposed causal diagram that we use to inform our statistical analysis in Section 'Estimating the effect of personal smoking on adult asthma: the TAHS data'). Although our question of interest concerns the causal effect of smoking on subsequent asthma, in our three simplified examples, we have assumed that this causal effect does not exist, as indicated by the absence of arrows from personal smoking to adult asthma, in order to more clearly explain the features of causal diagrams relevant to confounder selection.

Figure 1a encodes the assumption that childhood asthma has a causal effect on both subsequent smoking behaviour and subsequent asthma of an individual. Further, it assumes that childhood asthma is the only variable that affects both personal smoking and subsequent asthma, as any other such variable would also have to be included in the causal diagram.

Figure 1b shows a slightly more complex causal scenario. In this case, childhood asthma is assumed to have no causal effect on subsequent smoking behaviour, but childhood asthma and personal smoking are connected by a common cause (parental smoking). Similarly, it assumes that the relationship between childhood and adult asthma is not a causal effect of the former on the latter but that both are caused by the underlying atopy of the individual.[11] We suppose that underlying atopy cannot be adequately measured; however, it must still be included in the causal diagram if we believe it affects both childhood and adult asthma.

Figure 1c is very similar to 1b but additionally encodes the assumption that childhood asthma has a causal effect on personal smoking behaviour, as well as being connected through the common cause of parental smoking.

The key assumptions in a causal diagram are in the absence of arrows. In Figure 1b, for example, by omitting an arrow from childhood asthma to personal smoking, we are explicitly asserting our assumption that childhood asthma has no effect on the individual's subsequent smoking status.

Paths and association

Causal diagrams explicitly separate the concepts of causation and association. The underlying causal

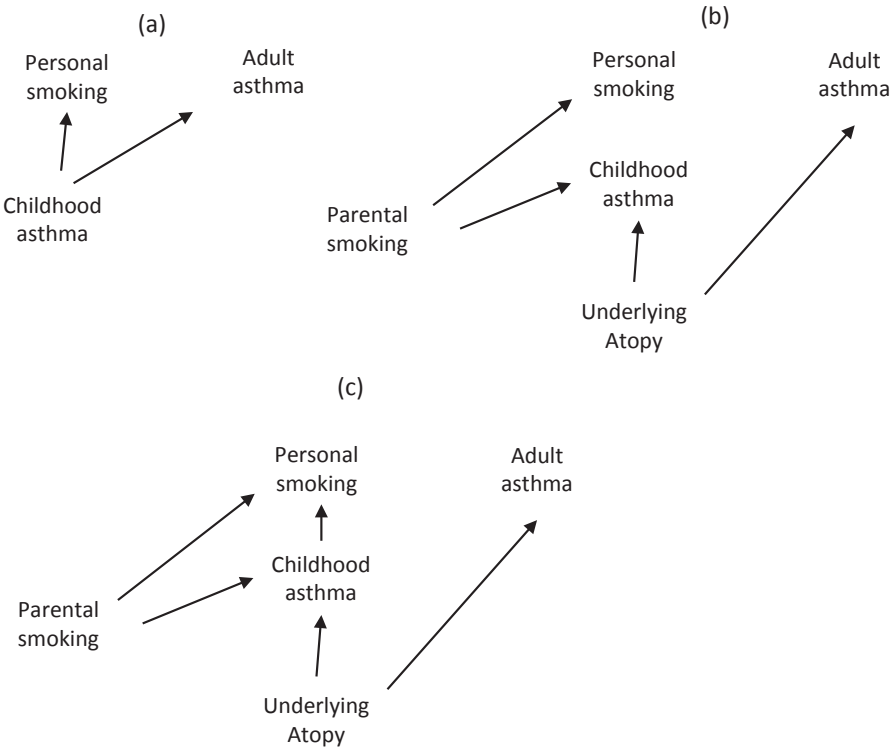

Figure 1 Causal diagrams showing three possible underlying relationships connecting personal smoking and subsequent asthma. We assume that all characteristics shown are measured, in the data to be used for analysis, except underlying atopy which is unmeasured.

structures shown in the causal diagram inform us about which statistical associations between the variables in the diagram are present, assuming that the diagram correctly depicts the underlying scenario. This allows us to form valid estimates of causal effects from the observed associations (either unadjusted or adjusted for measured variables) in our data.

We begin by using the causal diagram to decide whether, under the assumptions of the diagram, an unadjusted comparison of the outcome between exposed and unexposed participants (i.e. the unadjusted exposure-outcome association) will be a biased estimate of the causal effect of the exposure on the outcome. Using the causal diagram, we can determine whether spurious (non-causal) exposure-outcome association is present. If no such spurious association exists, then the unadjusted comparison will be unbiased. If spurious association exists, this unadjusted comparison will be biased. In this case, adjustment for measured variables may help to remove the bias (see Section 'Removing confounding bias by adjustment').

In causal diagrams, a *path* from the exposure (personal smoking) to the outcome (adult asthma) is a series of arrows connecting the exposure and outcome, irrespective of the direction of arrows. A *causal path* from the exposure to the outcome is a path starting at the exposure in which each arrow is followed from tail to head until it reaches the outcome (e.g. personal smoking → bronchial hyperreactivity → adult asthma). These causal paths are the only paths in the diagram that depict causal effects of the exposure on the outcome. All other paths from exposure to outcome are non-causal. The three causal diagrams in Figure 1 contain no causal paths from personal smoking to adult asthma. Under these three scenarios, any observed association between exposure and outcome is spurious (non-causal). Figure 1a and b each contain one non-causal path from personal smoking to adult asthma, while Figure 1c contains two such paths: Path 1 (personal smoking ← parental smoking → childhood asthma ← atopy → adult asthma) and Path 2 (personal smoking ← childhood asthma ← atopy → adult asthma).

Suppose that Figure 1a depicts the true underlying relationship between personal smoking and adult asthma. In this scenario, childhood asthma plays the role of a traditional confounder—it is a common cause of the exposure and outcome. This results in a 'mixing of effects' of smoking and childhood asthma on adult asthma.[12] For example, childhood asthmatics may be less likely to smoke but more likely to have adult asthma. Then, an unadjusted comparison of smokers (a small proportion of whom were childhood asthmatics) and non-smokers (a larger proportion of whom were childhood asthmatics) would falsely conclude that the non-smokers have a higher risk of asthma, even though personal smoking truly has no causal effect on adult asthma. In statistical terms, the confounding by childhood asthma creates spurious (non-causal) association between personal smoking and adult asthma. Therefore, under Figure 1a, an unadjusted analysis will be biased.

In causal diagram terminology, the path in Figure 1a (personal smoking ← childhood asthma → adult asthma) is *open*. This means that the relationships shown within that path create association between the variables at each end of the path (the exposure and outcome).

Now suppose instead that Figure 1b depicts the true relationship between personal smoking and adult asthma. Despite there being a path from personal smoking to adult asthma in this diagram, personal smoking will not be associated with adult asthma, so an unadjusted comparison of asthma status between smokers and non-smokers will correctly conclude that personal smoking has no causal effect on adult asthma; this unadjusted analysis is unbiased. In the path of Figure 1b (personal smoking ← parental smoking → childhood asthma ← atopy → adult asthma), childhood asthma is called a *collider* variable; the arrows on this path 'collide' at childhood asthma. In epidemiological terms, childhood asthma is a common effect (of parental smoking and atopy). Importantly, association is not transmitted across common effects (collider variables). Just because two factors share an effect does not mean that they themselves are associated. Thus, the path in Figure 1b is *closed*; although the path connects the exposure to the outcome, this connection does not create association between these two variables.

Paths in causal diagrams are either open (they transmit association) or closed (do not transmit association). Figure 2 lists a set of rules that can be applied to each path in a causal diagram to decide if it is open or closed.

An unadjusted comparison of the outcome between exposure groups will be an unbiased estimate of the causal effect of the exposure if there is no spurious exposure-outcome association. This spurious association can only arise through open non-causal paths from the exposure to the outcome. Thus, an unadjusted comparison of the outcome between the exposure groups will be unbiased only if all non-causal paths between the exposure and outcome are closed. In Figure 1a, the single non-causal path from personal smoking to adult asthma is open (see rule 3, Fig. 2); thus, the unadjusted analysis is biased. In Figure 1b, the single non-causal path from personal smoking to adult asthma is closed (rule 5, Fig. 2); thus, the unadjusted analysis is unbiased.

When, as is typical, there are multiple non-causal paths between the exposure and the outcome, each path must be examined separately to determine if it is open or closed in order to determine whether there may be spurious association transmitted between the exposure and outcome. Suppose Figure 1c, for example, depicts the true causal scenario. In this case, there are two non-causal paths from personal smoking to adult asthma. One contains a collider variable (Path 1: personal smoking ← parental smoking → childhood asthma ← atopy → adult asthma), and the other does not (Path 2: personal smoking ← childhood asthma ← atopy → adult asthma). Therefore, Path 1 is closed (rule 5, Fig. 2), but Path 2 is open (rule 3, Fig. 2). Note that the definition of a collider is path-specific. Childhood asthma is a collider in Path 1 but not in Path 2. Because Path 2 is open, spurious association exists between personal smoking and adult asthma; thus, an unadjusted analysis will be biased.

Removing confounding bias by adjustment

If an unadjusted comparison of the outcome between exposure groups gives a biased estimate of the causal effect of the exposure, it may be possible to remove the confounding bias by performing an analysis adjusting for a set of measured variables. We would fit a multivariable regression model[1] (e.g. a logistic regression model) for the outcome, including the exposure and the selected measured variables as explanatory variables. The estimated exposure effect from this model (e.g. the exposure odds ratio) will provide an unbiased estimate of the causal effect of

A causal path from exposure to outcome
1. Is open (by definition it does not contain any collider variables)
2. Should be left open (do not adjust for any variables on these causal paths)

A non-causal path from exposure to outcome containing no collider variables
3. Is open if no variables on the path are adjusted for
4. Is closed if one or more variables on the path are adjusted for

A non-causal path from exposure to outcome containing one collider variable
5. Is closed if no variables on the path are adjusted for
6. Is closed if only non-collider variables are adjusted for
7. Is open if the collider variable,* is the only variable on the path adjusted for
8. Is closed if the collider variable,* and one or more other (non-collider) variables are adjusted for

A non-causal path from exposure to outcome containing more than one collider variable
9. Is closed if no variables (or only non-collider variables) on the path are adjusted for
10. Is closed if at least one collider variable,* is not adjusted for
11. Is open if all the collider variables,* but no non-collider variables, are adjusted for
12. Is closed if all collider variables,* and one or more other (non-collider) variables are adjusted for

Figure 2 Rules to decide whether a particular path is open or closed in a causal diagram. *The same rules apply if, instead of adjusting for a collider, we adjust for a variable that is caused by that collider.

the exposure provided that adjustment for the selected group of variables is sufficient to remove confounding bias. In causal diagram terms, this will be the case provided that adjustment for these variables closes all non-causal paths from the exposure to the outcome, thereby removing all spurious exposure-outcome association, under the causal assumptions encoded in the proposed causal diagram.

Under Figure 1a, for example, we have seen that the unadjusted analysis is biased because there is an open non-causal path from personal smoking to adult asthma. This path is closed by adjusting for childhood asthma (rule 4, Fig. 2). Thus, under Figure 1a, a comparison of adult asthma between smokers and non-smokers adjusting for childhood asthma status (e.g. via a logistic regression model of adult asthma on smoking and childhood asthma status) would correctly conclude that smokers and non-smokers have the same odds (and so the same risk) of adult asthma.

Under Figure 1b, we have seen that an unadjusted analysis would be unbiased. However, suppose that we decided to adjust for childhood asthma as it fulfils the traditional definitions of a confounder (it is associated with exposure and outcome and is not on the causal pathway). Unfortunately, if we adjust for childhood asthma, we will open the non-causal path between personal smoking and adult asthma (rule 7, Fig. 2). This creates spurious exposure-outcome association, thereby introducing bias into the estimated exposure effect. Therefore, adjusting for childhood asthma introduces bias.

To gain some intuition for this association introduced by conditioning on a collider variable, let us consider a slightly unrealistic version of Figure 1b. Suppose that the two arrows pointing to childhood asthma are deterministic: if a child is 'atopic' or has parents who smoke (or both), the child will certainly have childhood asthma; otherwise the child will not. If we consider only the subgroup of participants who have childhood asthma (which is another way of adjusting for childhood asthma), then any child whose parents do not smoke must be atopic. Conversely, any child who is non-atopic must have parents who smoke. Thus, after adjusting for childhood asthma, parental smoking status provides information about atopy status and vice versa; these characteristics are now strongly associated. This in turn leads to spurious association between personal smoking and adult asthma. This phenomenon is often referred to as *M-bias* (due to the 'M' shape that can be formed by drawing Figure 1b upside-down) or, more generally, *collider-stratification* bias.[4]

Table 1 lists various analysis strategies that might be used to estimate the effect of personal smoking on adult asthma under the three scenarios depicted in Figure 1 and whether each strategy leads to a biased or unbiased estimate of the exposure effect. We see that under Figure 1b, an unadjusted analysis, or an analysis adjusting for both childhood asthma and parental smoking, will lead to an unbiased estimate of the effect of personal smoking on adult asthma, whereas adjusting only for childhood asthma will produce a biased estimate of effect. In Figure 1c, both non-causal paths are simultaneously closed only when both childhood asthma and parental smoking are adjusted for; thus, these variables must be adjusted for in the statistical analysis in order to obtain an unbiased estimate of the exposure effect.

Further considerations

It is important to stress that the previous sections assume that our proposed causal diagram correctly depicts the underlying scenario. If our causal diagram is incorrect, our list of variables to adjust for may be incomplete or the adjustment may even increase confounding bias. It might therefore be important to

Table 1 Summary of how different choices of variables to be adjusted for in the statistical analysis will affect bias in the estimated effect of personal smoking on asthma under the three causal diagrams of Figure 1

Variables adjusted for†	Status of each path		Estimated exposure effect†
Figure 1a			
Path P1: *personal smoking ← childhood asthma → adult asthma*			
None (i.e. unadjusted analysis)	P1 = Open		Biased
Childhood asthma	P1 = Closed		Unbiased
Figure 1b			
Path P1: *personal smoking ← parental smoking → childhood asthma ← atopy → adult asthma*			
None (i.e. unadjusted analysis)	P1 = Closed		Unbiased
Childhood asthma	P1 = Open		Biased
Childhood asthma and parental smoking	P1 = Closed		Unbiased
Figure 1c			
Path P1: *personal smoking ← parental smoking → childhood asthma ← atopy → adult asthma*			
Path P2: *personal smoking ← childhood asthma ← atopy → adult asthma*			
None (i.e. unadjusted analysis)	P1 = Closed,	P2 = Open	Biased
Childhood asthma	P1 = Open,	P2 = Closed	Biased
Childhood asthma and parental smoking	P1 = Closed,	P2 = Closed	Unbiased

† Within a multivariable regression model.

consider several different causal diagrams that might show the true scenario and to assess the robustness of the estimated effects to these causal assumptions.

The conclusion, from the confounder selection process described earlier, will sometimes be that no set of measured variables is sufficient to remove confounding bias. It may be possible to perform further data collection to address this problem. If not, any statistical analysis of the data must be cautiously interpreted in the light of the unmeasured confounding variables.

We have assumed that we are interested in the total effect of the exposure on the outcome. If interest lies in isolating particular pathways through which the exposure is thought to affect the outcome—often called mediation analysis—extra care is needed in selecting an appropriate analysis.[13]

Clinical interest often lies in assessing the exposure effect within a subgroup of individuals. For the smoking-asthma question, the predominant focus may be on the effect of personal smoking on subsequent asthma among people who had childhood asthma (i.e. investigating asthma remission). Restricting the analysis to this subgroup is simply a different way of adjusting for childhood asthma. Under Figure 1b, for example, the observed smoking-asthma association among childhood asthmatics will give a biased estimate of the exposure effect, unless adjustment for parental smoking is additionally performed (via a multivariable regression model fitted on the subgroup of childhood asthmatics).

Checking the proposed adjustment

The methods described earlier will often be able to identify an *adjustment set*—a set of measured variables such that adjustment for these variables will ensure all non-causal paths between the exposure and outcome are closed, thereby providing a valid estimate of the exposure effect under the proposed causal diagram. Because identifying this adjustment set can be a tricky process, there are various ways of checking that adjustment for these variables will indeed close all the necessary paths.

Shrier and Platt[14] describe six simple steps, originally developed by Pearl,[8] that can be used to check that adjustment for the selected adjustment set removes all spurious exposure-outcome association under the assumptions encoded by the proposed causal diagram. These are summarized briefly in Figure 3.

Various computational tools exist to aid in the use of causal diagrams. For example, DAGitty[15] is an excellent freely available online program that can identify a suitable adjustment set, when given the proposed causal diagram.

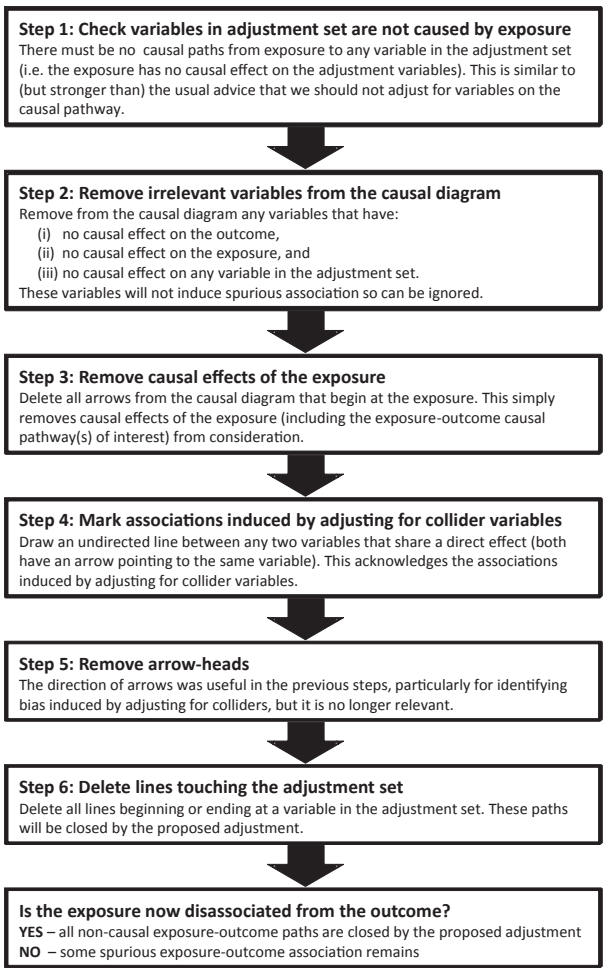

Figure 3 Summary of Shrier and Platt's[14] six-step rule (originally from Pearl[8]).

ESTIMATING THE EFFECT OF PERSONAL SMOKING ON ADULT ASTHMA: THE TAHS DATA

The TAHS data

In this section, we estimate the effect of personal smoking on asthma remission (no adult asthma) among participants who reported asthma during childhood. We use data taken from the TAHS, a population-based longitudinal cohort study of 8683 children born in 1961 and attending school in Tasmania in 1968.

At study enrolment in 1968,[16] parents provided information on their child's respiratory health including asthma (age at onset and the number of asthma attacks during the 12 months prior to enrolment), bronchitis and pneumonia history, together with information on their own respiratory health, smoking history and occupation. Each child underwent lung function testing. In the 2004 follow-up survey, the participant's adult asthma status, smoking history and occupation (reflecting socio-economic status) were documented.

The original data have been used in an extensive investigation of risk factors for asthma remission; clinical interpretations of these analyses have been reported previously.[17] This analysis is for illustrative purposes only and uses a subsample of 194 participants from the TAHS data who reported asthma during childhood.

Our proposed causal diagram

Our proposed causal diagram to address the smoking-asthma research question is shown in Figure 4. This was created by drawing on substantive knowledge of the underlying causal relationships, combining subject-matter expertise and results of previous research.[5] We have now included an arrow from personal smoking to adult asthma as this arrow represents our research question—the causal effect we wish to estimate. Figure 4 adds four characteristics to Figure 1c: chronic bronchitis/poor lung function, parental asthma, socioeconomic status (SES) and sex. All variables shown in Figure 4 are measured in the TAHS data other than underlying atopy.

Confounder selection using the proposed causal diagram

Because Figure 4 is simply an extension of Figure 1c (with additional nodes and paths), we know that both parental smoking and childhood asthma must be adjusted for. Chronic bronchitis/poor lung function, SES and sex play a similar role to childhood asthma in Figure 1a. They are each the only variable in an open non-causal path that contains no collider variables. Thus, we must adjust for these variables. Adjusting for parental smoking, childhood asthma, chronic bronchitis, poor lung function, SES and sex closes all non-causal paths between personal smoking and adult asthma. Thus, adjustment for these variables is sufficient to remove confounding bias under the assumptions depicted in Figure 4. We apply the steps of Shrier and Platt[14] (Fig. 3) in Appendix S1 in the online supporting information to ensure that this proposed adjustment closes all necessary paths.

Statistical analysis

We have used the proposed causal diagram to select a set of confounders that we need to adjust for. However, some modelling decisions remain. Although our causal diagram suggests that we should adjust for SES, for example, it does not convey information about whether the asthma–SES relationship is linear across categories or not, nor does a standard causal diagram convey information about effect-modification, although some work has been done

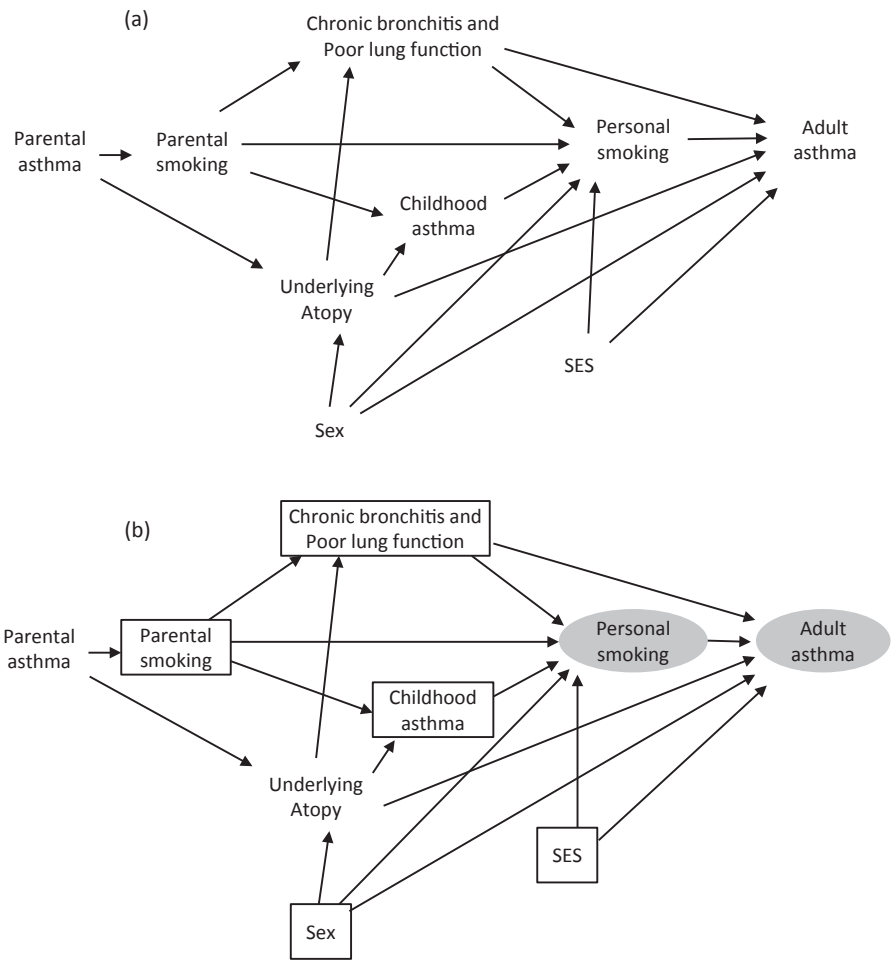

Figure 4 (a) Proposed causal diagram to investigate the hypothesized causal effect of personal smoking on subsequent adult asthma. (b) The same diagram with the exposure and outcome shaded and the proposed adjustment set indicated by boxes. In the Tasmanian Longitudinal Health Study data, all variables in these diagrams are measured other than underlying atopy.

incorporating such relationships.[18,19] Such modelling decisions must be considered independently of the causal diagram. To keep the analysis simple, we have not considered non-linearities or interactions between variables.

Also for some nodes in our causal diagram, we must decide which variables best represent these nodes. For example, do the relationships in Figure 4 represented by the node 'childhood asthma' depend only on the presence of childhood asthma, or do they additionally depend on the severity? We believe both are relevant; thus, our variable(s) must encapsulate both of these aspects. By restricting the sample to childhood asthmatics, we have already adjusted for the presence of childhood asthma. To additionally account for the severity of asthma, we adjust for the number of asthma attacks as an indicator of severity.

To estimate the effect of personal smoking on subsequent asthma remission among participants reporting asthma during childhood, we fitted a multivariable logistic regression model[1] for the outcome of asthma remission (yes/no). We included personal smoking (ever/never), and the confounders selected using the causal diagram (poor childhood lung function, chronic bronchitis, number of asthma attacks, sex, number of parents reporting smoking and SES) as independent variables in the model.

Results

Of the 194 participants analysed, 119 (61%) were adult smokers. Of the smokers, 86 (72%) had subsequent asthma remission (no adult asthma) compared with 51 (68%) of the never smokers.

In this subsample of data, the unadjusted odds ratio comparing asthma remission between adult smokers and non-smokers was odds ratio = 1.23 (95% confidence interval 0.65–2.30, $P = 0.53$), suggesting that smokers might have increased odds of asthma remission but that a lack of association is entirely plausible. This was reduced to odds ratio = 1.12 (95% confidence interval 0.56–2.27, $P = 0.75$) by adjusting for the confounders selected by the causal diagram approach.

Conclusion from analysis of TAHS data

In this subsample of data, we find little evidence that adult smoking affects the odds of subsequent asthma remission. The initial suggestion of a protective effect of adult smoking was largely removed by adjustment for variables believed to be causing confounding bias.

The validity of our analysis depends on our postulated causal diagram being correct. If we have omitted key variables or relationships, then our adjustment set may be incomplete. Furthermore, we have assumed that our measured variables perfectly capture the relationships shown in our causal diagram. Because this analysis is for illustration only, we have not further investigated these issues.

DISCUSSION

Causal diagrams (directed acyclic graphs) are a useful way to communicate the causal assumptions underlying an analysis investigating the relationship between an exposure and outcome. They provide a simple algorithm to decide which variables should be adjusted for in a multivariable analysis in order to isolate the causal effect of interest. However, a causal diagram that incorrectly portrays the underlying scenario can lead to poor decisions concerning which variables should be adjusted for. A causal diagram, therefore, is only as useful as the substantive knowledge used to create it. Given the complexity of many health research studies, it may be naïve to assume that an appropriate diagram can always be created. However, where causal assumptions in the diagram are uncertain, we could postulate a few competing diagrams, which may result in different analyses, and present the competing estimates of effect each being valid under certain explicitly stated assumptions.

Causal diagrams have been used in many contexts including: defining and identifying selection bias,[20] accounting for measurement error,[21] baseline adjustment in analyses of change scores,[22] and understanding and identifying biases in indirect treatment comparisons,[23] analyses of time-varying exposures,[24] mediation analyses,[13] and the estimation of direct and indirect effects.[25] Causal diagrams have also been applied to understand biases arising due to missing data.[26]

In summary, causal diagrams or directed acyclic graphs are an invaluable tool for confounder selection in analyses involving non-randomized exposures. By explicitly stating the causal assumptions underlying the variable selection process, causal diagrams increase transparency and facilitate communication and debate concerning the validity of estimated causal effects.

Acknowledgements

We thank Associate Professor Julie Simpson from the Melbourne School of Population and Global Health, The University of Melbourne and Professor Michael Abramson from the School of Public Health and Preventive Medicine, Monash University for useful comments on the manuscript, and Professor Dallas English from the Melbourne School of Population and Global Health, The University of Melbourne for years of discussion and debate about causal diagrams in epidemiology. We thank the TAHS Steering Committee for providing us with a random subset of the data from the TAHS cohort that was funded by the National Health and Medical Research Council, Australia, ID#299901. This work was supported under a National Health and Medical Research Council Centre of Research Excellence grant, ID#1035261, to the Victorian Centre for Biostatistics (ViCBiostat).

REFERENCES

1 Kasza J, Wolfe R. Statistical regression models: interpretation of commonly-used models. *Respirology* 2014; **19**: 14–21.
2 Greenland S, Neutra R. Control of confounding in the assessment of medical technology. *Int. J. Epidemiol.* 1980; **9**: 361–7.
3 Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am. J. Epidemiol.* 1993; **138**: 923–36.
4 Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003; **14**: 300–6.
5 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; **155**: 176–84.
6 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**: 37–48.
7 Pearl J. Causal diagrams for empirical research. *Biometrika* 1995; **82**: 669–710.
8 Pearl J. The art and science of cause and effect. In: *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2000; 331–58.
9 Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes M, Kaufman J (eds) *Methods in Social Epidemiology*. Jossey-Bass, San Francisco, CA, 2006; 387–422.
10 Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL (eds) *Modern Epidemiology*, 3rd edn. Lippincott Williams & Wilkins, Philadelphia, PA, 2008; 183–212.
11 Martin PE, Matheson MC, Gurrin L, Burgess JA, Osborne N, Lowe AJ, Morrison S, Meszaros D, Giles GG, Abramson MJ *et al*. Childhood eczema and rhinitis predict atopic but not nonatopic adult asthma: a prospective cohort study over 4 decades. *J. Allergy Clin. Immunol.* 2011; **127**: 1473–9.
12 Rothman KJ, Greenland S, Lash TL. Validity in Epidemiologic Studies. In: Rothman KJ, Greenland S, Lash TL (eds) *Modern Epidemiology*, 3rd edn. Lippincott Williams & Wilkins, Philadelphia, PA, 2008; 128–47.
13 Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *Int. J. Epidemiol.* 2013; **42**: 1511–9.
14 Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med. Res. Methodol.* 2008; **8**: 70.
15 Textor J, Hardt J, Knüppel S. DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology* 2011; **5**: 745.
16 Gibson HB, Silverstone H, Gandevia B, Hall GJ. Respiratory disorders in seven-year-old children in Tasmania: aims, methods and administration of the survey. *Med. J. Aust.* 1969; **2**: 201–5.
17 Burgess JA, Matheson MC, Gurrin LC, Byrnes GB, Adams KS, Wharton CL, Giles GG, Jenkins MA, Hopper JL, Abramson MJ *et al*. Factors influencing asthma remission: a longitudinal study from childhood to middle age. *Thorax* 2011; **66**: 508–13.
18 VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. *Epidemiology* 2007; **18**: 561–8.
19 Weinberg CR. Can DAGs clarify effect modification? *Epidemiology* 2007; **18**: 569–72.
20 Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–25.
21 VanderWeele TJ, Hernan MA. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *Am. J. Epidemiol.* 2012; **175**: 1303–10.
22 Glymour MM, Weuve J, Berkman LF, Kawachi I, Robins JM. When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *Am. J. Epidemiol.* 2005; **162**: 267–78.
23 Jansen JP, Schmid CH, Salanti G. Directed acyclic graphs can help understand bias in indirect and mixed treatment comparisons. *J. Clin. Epidemiol.* 2012; **65**: 798–807.
24 Platt RW, Schisterman EF, Cole SR. Time-modified confounding. *Am. J. Epidemiol.* 2009; **170**: 687–94.
25 Cole SR, Hernan MA. Fallibility in estimating direct effects. *Int. J. Epidemiol.* 2002; **31**: 163–5.
26 Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat. Methods Med. Res.* 2012; **21**: 243–56.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1 Application of the six-step rule to verify the adjustment set for the TAHS data.