Interaction can also occur when the outcome is a continuous variable, for example, blood pressure or cholesterol levels. The current trend is for very large studies and this is driven by the large samples required for studying gene–environment interactions.

The promise of the new genetics is to clarify the nature of individual and group-level susceptibility to diseases, and the variable response to treatments. Gene variants can act as effect modifiers/interactive factors of the relationship between an environmental exposure and an outcome. The problem is that studies that can accurately assess interaction need to be very large, especially when effects are small (as with most gene variants). Mostly, studies that report there is no interaction are too small to reach such a conclusion, and many others have applied the wrong conceptual approach. If there is heterogeneity in the effects of a risk factor within populations, as is highly likely, there must be effect modification/interaction and vice versa. Failure to seek or notice risk modification can lead to a false measure of population risk and the possibility of missing an important finding, at least for population subgroups. The importance, in practice, of effect modification/interactions is under debate.

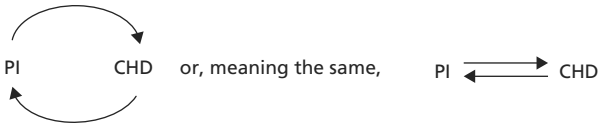
Causal graph methods that are helpful in data gathering, analysis, and interpretation have become available in epidemiology (mainly developed and advanced in other disciplines), and these are considered next. At present, these methods do not incorporate effect modification, and are being developed for this.

5.5 Causal graphs: introducing the directed acyclic graph

A diagram (or graph) that makes explicit the postulated relations between variables is recommended. Producing such a diagram is difficult because it requires considerable understanding, including of the biology, of the topic under study. This step is a component of a disciplined approach to research that includes prior statement of a fully articulated scientific hypothesis (not merely the statistical null hypothesis) and a detailed analysis plan prepared in advance. With a hypothesis, causal diagram, and an analysis plan, we can assess whether the results of the research fit with our prior presumed causal understanding. If yes, that increases confidence that our prior understanding was correct. If not, then it motivates us to alter and improve our hypothesis and causal diagram. This can then be checked with new research. The field of causal diagrams in epidemiology is large, complex, and is developing fast. One important and increasingly used form of causal diagram is called the directed acyclic graph (DAG). The field has been advanced by, among others, Judea Pearl, who is a professor of computing sciences and statistics. Pearl has claimed in his book *Causality* (second edition) that causality has been mathematized. While this claim is not (yet) true for epidemiology and related population health and medical sciences, it merits examination.

The DAG is based on graph theory in mathematics and can be written as an algebraic equation (in a form known as a structural equation model). Equally, algebraic equations can be expressed as DAGs. This property allows DAGs to be used both for expressing potential causal pathways, and for guiding and interpreting data analysis. The field of DAGs has vocabulary that is different but related to that of epidemiology. Some of this vocabulary is in Table 5.4, which also re-expresses it in similar epidemiological terms. The DAG uses lines and arrows (Table 5.5) that follow formal rules.

Assume we have a relationship between two variables (A and B) that we believe is potentially causal. Before examining the DAG approach, we will look at this relationship in general terms. To make this less abstract let us name A as the variable physical inactivity (PI) and the outcome B as coronary heart disease (CHD). It is known that physical inactivity can cause CHD, but also that CHD can cause inactivity. This complex relationship can be designated as circular:



Epidemiology simplifies this kind of problem by examining the possibilities separately. Probably the best way to do this is by studying young people long before they get the outcome, here CHD, which occurs in later life. Therefore, we could, for example, consider the relationship between physical inactivity in early life, childhood, adolescence, and early adulthood and the outcome

Table 5.4 Terminology commonly used in DAGs in relation to similar or related terminology in epidemiology

DAG terminology	Epidemiology terminology for similar ideas
Ancestors	Distal causes
Back door pathway	Confounding variable(s) creating the association
Block/blocking	Eliminating an association through, for example, adjusting, stratifying, etc.
Blocked path	An association that has been eliminated as it has been controlled for, e.g. by adjusting for confounders
Blocking	Presence of confounding or selection bias
Child	Effect, outcome
Collapsibility	The measure of the association is not affected whether examining stratified or overall, actual (crude) rates
Collider	A variable that is caused by both the exposure and outcome under study
Collider stratification bias	See M-bias
Conditioning	General term to include stratification, standardization, and adjustment in a model (conditioning means holding a variable constant)
Descendants	Effects on potential causal path including mediators
D-connected	The postulated causal path is open (see open path and path)
D-separation or D-unconnected (directional separation)	The postulated causal path is closed
Endogenous selection	See M-bias
Identification	Analysis of associations to separate error/bias/confounding from causal effects
M-bias	Berkson's bias/selection bias. It arises from, for example, adjusting for a collider
Nodes	Variables
Open path	Potential causal relationship, i.e. association
Parent	Proximal cause
Path	The route to potential causality, i.e. from A to B
Vertices	Variables

Table 5.5 Symbols used in the construction of directed acyclic graphs (DAGs)

Symbol	Name of symbol
———	Edge. This edge is undirected, i.e. there is no arrow to give the direction
>	The direction indicator for an edge
⟶	Directed edge (arrow) indicating association (arrow can be thickened to denote stronger association)
X	Variable or in some notations variables conditioned on X (the practice of putting variables in a box is not always followed in epidemiology)
. . . . .	Association induced by collider bias
< ····· >	Bidirected edge, denoting an association induced by confounding (the line can be solid)

CHD in middle age and beyond. In this case we postulate and study that physical activity leads to CHD i.e. PI → CHD. It is not plausible that CHD in middle age leads to inactivity earlier. So in doing this, we have simplified our research.

In contrast to this example from a chronic disease, nearly all infectious diseases have the simple A → B relationship where A is the infectious agent and B is the outcome; for example, A is the measles virus and B is the illness. It is not plausible that measles illness causes the acquisition of the measles virus.

So, for causal analysis we usually start by simplifying the matter under study. Simplifying means the construction of models (in the same way that an architect may create a model building). The DAG is a model. The equation describing the DAG is also a model. The DAG needs specification of exposure (A) and outcome (B) and, ideally, on the ancestors and descendants (see Table 5.4) of A and B.

A relationship between A and B can be expressed as A – B, so A and B are connected by a line that we call an edge. The variables are called nodes. Does A cause B or does B cause A (the problem of reverse causation)? As it is designed to help causal analysis, the DAG method does not permit two-headed (bidirectional) arrows. The word acyclic means there are no cycles in the graph. When complex cyclical relationships are to be studied, other types of causal graphs need to be used (theory is available). Of course, we could create two DAGs, one for A → B and the other for A ← B. We could plan and run the analyses separately. Let us assume that our interest is in A causes B, then we can express that as

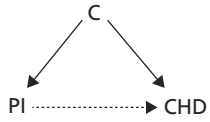
$$A \rightarrow B$$

If A → B, it could either increase B (positive association) or decrease B (negative association). The DAG can be used to show this, for example, by having one colour for positive relationships and another for negative ones or plus/minus signs. Clearly, the researchers need a high level of knowledge to set out these kinds of decisions in advance.

Let us assume that our understanding of the relationship between physical inactivity and CHD is not based on our reading of fabricated, seriously biased, or chance research results (i.e. that it is a reasonable proposition).

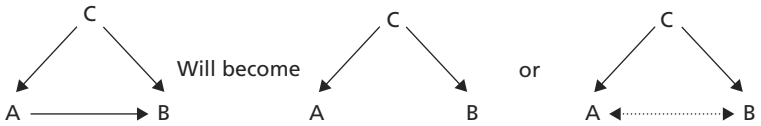
What else do we need to consider in this DAG to make it informative? First, we must consider the possibility that a third variable (C) is creating the relationship A → B, that is, there is confounding (known as a backdoor path in DAG terms). The confounding factor could be age, sex, socioeconomic status, or similar variables.

This can be expressed as

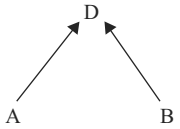


Here the dashed line denotes the possibility of a confounded relationship. As we already know if we adjust for C, and C is a confounding factor creating this relationship, then the association between A and B will greatly reduce or even disappear. The dashed line will disappear. This is described as closing the backdoor path.

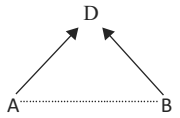
If we control for a confounding variable and the association between A and B disappears, we can redraw the DAG either without an arrow between A and B, or with a dotted line between A and B to signify that relationship is confounded.



When both our variables of interest are associated with a third variable, presumed causally in that direction, this variable is called a collider. If there is a collider D, then we have



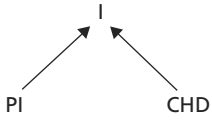
If we control for collider D, we induce a spurious association between A and B, which is denoted as



The dotted line has no arrows, which distinguishes it from confounding.

It is possible that both physical inactivity and CHD cause an effect that is important in the relationship, for example, social isolation (I)

So



I (social isolation) is called a collider. The collider blocks the path from P1 to CHD but adjusting for it opens it up, the opposite of adjusting for confounding (closing the back door path). Here a biased association may be created between PI and CHD because of adjustment for social isolation. This is not at all an unfamiliar concept (even though it is still an unfamiliar term) in epidemiology,

but the use of DAGs has made it clearer. We have already discussed this in Chapter 4 with the example of Berkson's bias. In this example, let us assume that, in fact, there is no association between PI and CHD as in the above diagram (no line connecting the two). If we condition on the collider I, we would hold it constant. One of several ways of doing that is to study only socially isolated people, that is, stratify. In this group, there will be an association between PI and CHD, if these two variables are among the causes of social isolation, which seems plausible. The term *collider stratification bias* includes a range of biases that have varying names in epidemiology (e.g. Berkson's bias).

In analysis, our aim is to help separate non-causal and causal associations so we can see we must not control for colliders (or their descendants or ancestors) and mediators (or their ancestors or descendants), but we should control for confounders. If the study is free of error/bias, and all confounding factors are included, then the DAG reflects a causal structure and the resulting analysis reflects a causal relationship between A and B. Of course, this ideal state is not actually achieved.

We may ask how physical inactivity causes CHD. One possibility, among others, is that it does so through obesity (O).

$$PI \rightarrow O \rightarrow CHD$$

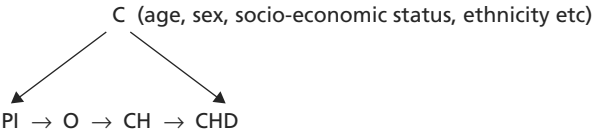
Obesity might, for example, work through raised cholesterol (CH).

$$PI \rightarrow O \rightarrow CH \rightarrow CHD$$

Obesity and cholesterol would, therefore, be postulated to be intermediate variables (synonym, mediators). If, however, we adjust statistically for the intermediate variable O and the relationship between PI and CHD remains completely unaltered then it is not, actually, acting as an intermediate variable. If the association weakens or disappears, then the case for O being a mediator is strengthened. However, this interpretation requires that the outcome, here CHD, does not cause the intermediate (i.e. obesity), which common sense tells us is plausible but we make the assumption this in not true. (In our DAG we may either use different colours for the lines/arrows for different kinds of variables, or use different line widths or use broken lines.)

Therefore, our revised causal proposition might be that after adjusting for potential confounding factors, we think physical inactivity leads to CHD through obesity and cholesterol.

So, this DAG will be

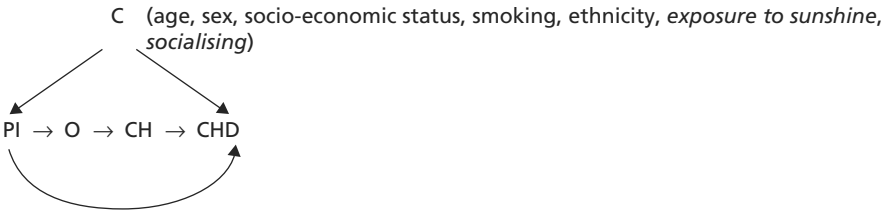


We can see the DAG has helped to think through and clearly express, in a way that everyone can understand, complex relationships between confounders (C), the exposure/risk factor (PI), intermediate variables (O, CH) and the outcome (CHD). Our knowledge and data sets both tend to be incomplete, so this is a provisional model. The model should also include uncertainties. Uncertainties can be added, for example, as U for unmeasured variables and error terms.

If we know of variables that are potentially important in the relationship but have not been measured, they should be added to the DAG so we can immediately see the limitations of the study. So, for example, it may not be physical inactivity that is important but the fact that physically inactive persons may spend less time outdoors, thereby being less exposed to sunshine or to the social benefits of meeting neighbours and people on the street. These could be unmeasured confounders or mediators. Our study may well have no data, either because these variables were

forgotten, omitted deliberately, or most probably it was not feasible to collect the data, possibly for reasons of time and cost. The DAG helps by being explicit on the limitations of the causal analysis.

Again, these unmeasured variables could be picked out using a different colour or different kind of line. Here, these are shown as unmeasured confounders in *light grey*. These unmeasured factors are treated here as confounders, but they may actually be on the causal path. Investigators need knowledge and judgement to decide where the variables belong.



Of course, physical inactivity may have effects on CHD either directly (without mediators as shown by the direct arrow) or through a completely different pathway than obesity (e.g. through endothelial dysfunction). The effect of the direct and indirect paths can be quantified. This extra, endothelial pathway, could be added (but this is beyond our purposes here). There may be confounding in the relationships between the intermediate variables and both exposure and outcome, which needs to be considered. Ideally, the DAG (and statistical model) should include confounders not only between PI and CHD, but also between other variables on the path.

The DAG is also telling us an important story about the variables that are not shown. The DAG is telling us that the investigators do not think these are important in this particular causal path. For example, diet is not mentioned. If that is not an oversight, then we can deduce the investigators think that diet has no relevance here. If so, excluding diet is the correct decision. Variables that do not connect to causal and confounding variables on the DAG should be excluded, even if they connect to the outcome through some other causal path, for example, smoking cigarettes and CHD. Ideally, investigators would explain why variables like this were not included. DAGs do not include statistical interaction, although this area is being developed.

A variable may be a moderator, mediator, and a confounder variable (e.g. socio-economic position). The elements of such a variable that are thought to be moderators or mediators need to be separated from those thought to be confounders, and should be labelled differently, for example, SEP1, (income), SEP2 (education), SEP3 (housing tenure), etc.

Software is available for drawing DAGs, for example, DAGitty (www.dagitty.net). This software uses colour coding with confounders in red, mediators in green, and no effect in grey. Currently, a variable can only have one status in this software. The DAG software shows which variables are, and which are not, essential for measuring the association between A and B.

Exemplar 5.1 illustrates a study that shows how the process works.

Exemplar 5.1: Factors associated with prediabetes (Bardenheier *et al.* 2013)

Bardenheier *et al.* used the cross-sectional United States National Health and Nutrition Examination Surveys (NHANES) 2001–2006 to test a hypothetical causal model for prediabetes in people aged 50 years or more. NHANES had data on 2230 eligible people without diabetes. The study was motivated by observing that while many risk factors for prediabetes had been identified, they had not been examined simultaneously as a coherent system or model. They identified a statistical method called structural equation modelling (see glossary) as a way of testing a hypothetical model. Their hypothetical models before and after analysis can

be examined in their original coloured format at this URL: http://care.diabetesjournals.org/content/36/9/2655.full.pdf

The variables that were measured were shown in rectangles, and those that were not measured but reflected in or derived from observed variables were shown in ovals and are known as latent variables. The arrows indicated the postulated direction of association.

Based on the investigators' knowledge, itself reflecting prior research, 10 major variables were identified as either direct or indirect predictors on the causal path to prediabetes. Arrows showed 27 paths from these variables to the outcome. Some of the variables were composite ones based on several measured variables, for example, socio-economic position (a latent variable; SEP), while others were single (and therefore measured); for example, high blood pressure.

The model was assessed using structural equation modelling with factor and path analysis. Factor analysis can group interrelated variables (factors are the lowest common denominators for a number). Path analysis assesses the direct and indirect effects of the factors identified. The factor analysis approach reduced the number of individual variables, which has statistical advantages, as well as easing examination of the model.

In the model, there were variables that had no number (i.e. age, race/ethnicity, and sex). These were identified as potential confounding variables. They were not entered as potential direct effects, while family history was. The authors justified this as follows: 'Because age, sex, and race/ethnicity are strong, non-modifiable confounders related to most of the other factors in the model, their direct effects, while included, are not shown in the graphic of the final model. Although family history is non-modifiable, it is specific to diabetes risk and therefore is examined as a factor of interest.'

The analysis used advanced statistics that is both beyond the scope of the book and unnecessary to understand the key points.

The structural equation model indicated that in the SEP factor (group of variables) the number of family members did not contribute. In the poor diet factor, saturated fats and processed meats did not contribute. The latent constructs were shown to be correlated with each other. To create a model fitting the data better, total cholesterol and BMI were removed. Then the direct effect of diet on high-density lipoprotein (HDL) cholesterol was dropped. Age, sex, and ethnicity were shown to have direct effects on most factors.

The model was reconstructed with the 10 postulated directly causal variables.

Concluding remarks

The outcome of prediabetes is not a complex one compared to many chronic diseases or syndromes such as CHD or asthma. The paper shows, however, how complex the putative causal relationships are. Yet, the authors have also taken a pragmatic decision to treat three major variables (age, sex, ethnicity) as confounding variables, rather than as direct effects or intermediate variables. This simplifies the thinking greatly. The great merit of this paper is the immense effort the authors have expended in creating a credible, theory-based causal diagram. They have then used data to assess the model. The data suggested a slightly different model that provided the best fit between the postulated model and the calculated model. The authors have, after this mammoth effort, made no claims to having produced a definitive model. Rather, recognizing the limitations of their data, in particular the cross-sectional design, they have urged further examination of the model, but with cohort data. They have also showed in what respect this model aligns with the published literature. This paper exemplifies how causal epidemiology might be done.

In my view, all papers working with associations aiming to contribute to the causal basis of a subject should provide the causal diagram based on current evidence (in the introduction) and the refined version following their research (in the discussion), as Bardenheier and colleagues did.

Source: data from Bardenheier *et al.* (2013) A novel use of structural equation models to examine factors associated with prediabetes among adults aged 50 years and older: National Health and Nutrition Examination Survey 2001–2006. *Diabetes Care*, Volume 36, Issue 9, pp. 2655–62.

DAGs and related diagrams help to design studies and to analyse data. By pointing to data which are needed, they can make research more efficient. Using DAGs is also changing basic concepts in epidemiology, an example being the realization that adjusting for a collider can cause bias, that is, a spurious association. If a variable is both a collider and a confounder then adjustment for it removes confounding, but can induce bias. Obviously, stratifying a sample into separate groups is a selection bias, but so is selecting a sample (except perfectly randomly), or non-response, missing data, subgroup analysis, or adjusting for a variable in regression analysis. So, selection bias (and hence colliders) are unavoidable in the practice of epidemiology.

The strength of the DAG approach is that it openly presents the assumptions guiding the analysis and interpretation of the data. The DAG, unlike the statistical analysis, is a causal model, albeit a postulated one.

DAGs can be very complicated so variables can be grouped, for example, as demographic variables, or as behaviour-related ones to help simplify the diagram.

The output is not a causal truth, but an opportunity for the investigators to assess whether the data fit the model. If there is a close match with most of the variances explained, then it would be reasonable to infer that the DAG is reflecting a causal structure, at least as portrayed in these data. The ultimate goal is to strip out bias and error, so what is left must be causal. This process is called identification.

The field of causal diagrams is enriching epidemiology and we are seeing these diagrams increasingly, though mostly in specialist journals. They have not, however, 'mathematized causality' (a claim by Judea Pearl) in epidemiology, which is still based on judgement as discussed in section 5.6. DAGs have the potential to contribute substantially to the formation of such judgements.

An even more complex approach is the logic diagram, which tries to produce a model that shows the full complexity of reality including non-causal, non-linear relationships. One of the best known of these diagrams is that produced by the Foresight Report on obesity (see http://www.noo.org.uk/NOO_about_obesity/causes). (Government Office for Science 2007).

5.6 Guidelines (sometimes erroneously called criteria) for epidemiological reasoning on cause and effect

5.6.1 Comparison of epidemiological and other guidelines for causal reasoning

Turning epidemiological data into an understanding of cause and effect is challenging and perhaps the most difficult aspect of the subject. Unfortunately, there is a widespread tendency to reach easy, but often premature or wrong conclusions. The commonest problem is either to declare, or interpret, an association as causal when it quite possibly a result of confounding or reverse causality. This problem may be becoming more common, partly because of media involvement. It is not easy to present the nuances of data interpretation in a news sound bite or even press release.