

# Using Causal Diagrams to Guide Analysis in Missing Data Problems

Rhian M. Daniel\*, Michael G. Kenward, Simon N. Cousens  
and Bianca L. De Stavola

Faculty of Epidemiology and Population Health,  
London School of Hygiene and Tropical Medicine

## Abstract

Estimating causal effects from incomplete data requires additional and inherently untestable assumptions regarding the mechanism giving rise to the missing data. We show that using causal diagrams to represent these additional assumptions both complements and clarifies some of the central issues in missing data theory, such as Rubin's classification of missingness mechanisms (as missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR)) and the circumstances in which causal effects can be estimated without bias by analysing only the subjects with complete data. In doing so, we formally extend the back-door criterion of Pearl and others for use in incomplete data examples. These ideas are illustrated with an example drawn from an occupational cohort study of the effect of cosmic radiation on skin cancer incidence.

**Keywords:** missing data; causal inference; causal diagram.

---

\*Address for correspondence: Rhian M. Daniel, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. E-mail: rhian.daniel@lshtm.ac.uk

# 1 Introduction

A key aim of medical and epidemiological research is to establish causal links between treatment, or other exposures, and outcomes. The gold-standard approach to achieve this aim is to conduct an ‘ideal’ randomised controlled trial (RCT), where by ‘ideal’ we mean large, double-blind, with no missing data and full compliance. Such ideal RCTs ensure that the observed outcomes in different treatment arms are free from any systematic differences except for those induced by the treatments being compared.

As we move from this ideal, causal inference increasingly requires further assumptions. Causal diagrams<sup>1,2</sup> can represent these and the accompanying theory is useful in informing the design and analysis of studies. Pearl and Greenland et al.<sup>1,2</sup> show how causal diagrams can be used to guide the choice of variables for data collection (and subsequent conditioning in the analysis), in order to make causal inferences more plausible.

Causal diagrams are increasingly used in non-randomised studies, where the main focus is on the control of naturally occurring confounding and investigator-induced selection bias.<sup>3</sup> The case for using causal diagrams to represent the mechanism assumed to give rise to missing data has not been extensively studied. In this article we fill this gap, showing that causal diagrams can complement and clarify some key issues relating to the analysis of incomplete data.

We consider the problem of estimating the causal effect of exposure  $A$  (*e.g.* exposure to cosmic radiation) on outcome  $Y$  (*e.g.* skin cancer), when either  $A$ ,  $Y$  or both, are incompletely observed. We consider whether or not this causal effect can be estimated without bias using only the *complete records*, *i.e.* subjects for whom both exposure and outcome are observed, and refer readers to appropriate alternative methods when this is not the case. We often consider additional covariates  $Z$ , and consider estimating the causal effect of  $A$  on  $Y$  conditional on  $Z$ . In this case, the definition of a complete record becomes a subject for whom  $A$ ,  $Y$  and  $Z$  are all observed.

The assumptions underpinning a causal analysis can be divided in two: (a) causal assumptions, such as ‘missingness is affected by exposure to radiation’ and (b) parametric assumptions, such as ‘the logarithm of the odds of skin cancer increases linearly with age’. The causal effect of  $A$  on  $Y$  will only

be estimated without bias if the assumptions made—both (a) and (b)—are close to being correct. We focus on assumptions of type (a) and assume that assumptions of type (b) (which should be checked using the data) hold.

We start, in §2, with our motivating example, an occupational cohort study of the effect of cosmic radiation on skin cancer incidence. We discuss, informally, how causal diagrams might be used here. This is formalised in §3. In §4, we apply our algorithm to various incomplete data scenarios, before returning—in §5—to our motivating example. The theoretical details, and further examples, are given in the Web Appendix.

## **2 Motivating example: the British Commercial Airline Pilots and Air Traffic Control Officers Study**

### **2.1 The data**

The British commercial airline pilots and air traffic control officers (ATCOs) study is an occupational cohort study set up to compare cause-specific mortality and site-specific cancer incidence rates between British professional pilots, ATCOs and those in the general population.<sup>4</sup> We focus on the estimation of the causal effect of cosmic radiation on skin cancer incidence, with cumulative flying hours serving as a proxy for radiation exposure.

Data on cumulative flying hours were collected using a questionnaire sent to about 27,000 eligible pilots and ATCOs with a response rate of around 50%. In addition, their permission was sought to access Civil Aviation Authority (CAA) medical records and NHS vital and cancer records, and around 92% consented. Outcome data (including skin cancer incidence) are available from these records. Other employment and personal information was collected in the questionnaires.

Estimating the causal effect of interest from these data requires consideration of the measurement error in the exposure, and the possibility of unmeasured confounders, as well as missing data. In this paper, we focus only on the missing data and proceed as if the proxy data are sufficiently close to the true exposure, and that—for this example—data on a sufficient set of confounders have been collected.

We will consider subjects with complete records to be those who responded to the questionnaire and responded to all the questions to be used in the analysis, as well as agreeing for their CAA and NHS records to be accessed. It is customary to treat *unit non-responders* (subjects who don't respond at all) separately from *item non-responders* (subjects who provide partial information), as the mechanisms leading to each are likely to be different. For simplicity, we will not distinguish between item and unit non-response in this article, even though the methods discussed extend naturally to incorporate this distinction.

We define the missingness indicator  $R$  to be 1 for the subjects with complete records, and 0 otherwise. We will only consider settings in which  $R$  may be causally affected by other variables, and not situations in which other variables are affected by  $R$ . In many settings (including our motivating example) this assumption is reasonable, although in some prospective studies, the act of being measured could itself affect a subject's subsequent behaviour.

## 2.2 Informal application of causal diagrams

Figure 1 shows a possible causal diagram for our example. Further possible diagrams are discussed in the Web Appendix. Arrows between variables denote the assumed direction of causal influence. Thus, Figure 1 represents the assumptions that both age and exposure to cosmic radiation have a causal effect on both the probability of developing skin cancer and a subject's propensity to respond (*i.e.* the probability of having a complete record). In addition, age has a causal effect on the exposure, and there are unmeasured influences on both the outcome and the propensity to respond (for more detail see §5). The magnitude of these causal influences is not specified, and thus an effect of magnitude zero (and hence independence of cosmic radiation and skin cancer, for example) is permitted. It is the omission of arrows in causal diagrams that represent our assumptions. So, in Figure 1 it is assumed that exposure to cosmic radiation is independent of genetic factors.

Conditioning on a common effect of two independent causes induces an association between them within strata of the conditioning variable. For some intuition as to why, consider measuring the association between sport-

ing and academic ability at a selective school where both are used as entry criteria. Even if uncorrelated in the population, they will be negatively correlated within the school, since a pupil with low academic ability is likely to have high sporting ability, and vice versa. A common effect of two independent causes could take the form of a *collider* (so called because two or more arrowheads ‘collide’ there; *e.g.*  $R$  in Figure 1), but also includes descendants of colliders (an example is given in the Web Appendix). Conditioning on a variable is denoted by placing a square box around it, and associations induced as a result of conditioning are denoted by dashed lines. Thus conditioning on  $R = 1$  in Figure 1 results in the graph shown in Figure 2. Note that this is no longer a causal diagram, since it represents non-causal (as well as causal) associations between variables.

It is tempting to treat Figure 2 as if it were a causal diagram and apply the *back-door criterion* to determine whether the causal effect of interest can be estimated without bias using only the complete records. Instructions on how to apply the back-door criterion are given in Greenland et al.<sup>2</sup> Briefly, we look for a path from the exposure to the outcome, other than the causal path (exposure to cosmic radiation→skin cancer), which does not contain a collider. If no such path exists, the causal effect of interest is estimated without bias. If such a path exists, it must be blocked, for example by conditioning on a variable on that path. In Figure 2, many back-door paths exist. Two of these can be blocked by conditioning on age and hair colour, but others remain.

### 2.3 The need for greater formality

The previous paragraph is informal for many reasons. First, we are treating Figure 2 as if it were a causal diagram. In other words, we are applying the back-door criterion to Figure 1 and including  $R$  in the conditioning set. However, this is not permitted when  $R$  is affected by the exposure.<sup>2</sup> Also, the dashed lines induced by conditioning on  $R$  are added prior to deleting the arrows emanating from the exposure, which again is not correct. Finally, in one of the examples included in the Web Appendix, a ‘back-door path’ is evident only when additional sources of variation in the outcome, independent of all other variables in the diagram, are included in the diagram. However, only common causes of two or more variables already in the diagram need

be included, implying that these additional nodes are unnecessary.

### 3 Guidelines for the use of causal diagrams in missing data problems

Two conditions need to be satisfied for the causal effect of exposure  $A$  on outcome  $Y$  to be identified from the complete records alone, conditional on a sufficient set of variables  $\mathcal{Z}$ . The technical details are given in the Web Appendix. A summary of the algorithm for determining whether these conditions are satisfied is given below. First we give a few definitions.

#### 3.1 Preliminary definitions

**Definition 1.** A *causal diagram*  $\mathcal{G}$  consists of nodes denoting variables, and arrows between nodes denoting the assumed direction of causal influence. Any variable which is the common cause of two or more variables in  $\mathcal{G}$  must be in  $\mathcal{G}$ .

Let  $\mathcal{V}_0$  be the subset of variables in  $\mathcal{G}$  which would have been observed on all subjects had there been no missing data. If there are unmeasured nodes in  $\mathcal{G}$ , such as those shown in Figure 1,  $\mathcal{V}_0$  will not contain every node in  $\mathcal{G}$ .

**Definition 2** (path). If  $W_1$  and  $W_m$  are disjoint nodes in  $\mathcal{G}$ , a *path*  $W_1W_2\dots W_m$  from  $W_1$  to  $W_m$  is a sequence of nodes such that, for each  $k = 1, \dots, m - 1$ , there is either an arrow from  $W_k$  to  $W_{k+1}$  or from  $W_{k+1}$  to  $W_k$  in  $\mathcal{G}$ .

**Definition 3** (directed path). The path  $W_1W_2\dots W_m$  is *directed* if all arrows go from  $W_k$  to  $W_{k+1}$ .

**Definition 4.** If there is an arrow from  $W_i$  to  $W_j$  in  $\mathcal{G}$ ,  $W_j$  is a *child* of  $W_i$ , and  $W_i$  a *parent* of  $W_j$ . If there is a directed path from  $W_i$  to  $W_j$  in  $\mathcal{G}$ ,  $W_j$  is a *descendant* of  $W_i$ , and  $W_i$  an *ancestor* of  $W_j$ .

For example, in Figure 1, there is a directed path from ‘Red hair’ to  $R$  via ‘Unmeasured behavioural factors’. ‘Age’ is a parent of ‘Skin cancer’,  $R$

is a child of ‘Exposure to cosmic radiation’,  $R$  is a descendant of ‘Red hair’, and ‘Unmeasured genetic factors’ is an ancestor of  $R$ .

Each child-parent family in  $\mathcal{G}$  (containing  $n$  nodes  $W_1, \dots, W_n$ ) corresponds to a function

$$W_i = f_i(pa(W_i), \varepsilon_i) \quad i = 1, \dots, n \quad (1)$$

from a nonparametric structural equations model, where  $\{\varepsilon_i : i = 1, \dots, n\}$  are independent unobserved random disturbances, and  $pa(W_i)$  are the parents of  $W_i$  in  $\mathcal{G}$ .

**Definition 5** (do operator).  $\check{w}_j$  denotes the act of *intervening* on  $W_j$  and setting its value to  $w_j$ .  $W_j = \check{w}_j$  is verbalised “do  $W_j$  equals  $w_j$ ” or “set  $W_j$  equal to  $w_j$ ”.

**Definition 6** (causal effect). For any  $l \neq k$ , the *causal effect* of  $W_l$  on  $W_k$ , denoted  $pr(w_k | \check{w}_l)$ , is a function from  $W_l$  to the space of probability distributions on  $W_k$ . For each  $w_l$ ,  $pr(w_k | \check{w}_l)$  gives the probability of  $W_k = w_k$  induced by intervening on  $W_l$  and setting its value to  $w_l$ . This probability is calculated by removing  $W_l = f_l(pa(W_l), \varepsilon_l)$  from (1) and replacing  $W_l$  with  $w_l$  in all other equations.

Thus, in general, the causal effect of  $W_l$  on  $W_k$  is a comparison of the different probability distributions for  $W_k$  obtained under the (hypothetical) interventions we could perform on  $W_l$ . However, often the term *causal effect* is used in connection with a specific function of  $pr(w_k | \check{w}_l)$ , such as the causal mean difference

$$\begin{aligned} & E(W_k | \check{w}_l = 1) - E(W_k | \check{w}_l = 0) \\ &= \sum_{w': pr(w_k = w') > 0} w' pr(w_k = w' | \check{w}_l = 1) - \sum_{w': pr(w_k = w') > 0} w' pr(w_k = w' | \check{w}_l = 0), \end{aligned}$$

or, for binary  $W_k$ , the causal odds ratio

$$\frac{pr(w_k = 1 | \check{w}_l = 1) pr(w_k = 0 | \check{w}_l = 0)}{pr(w_k = 0 | \check{w}_l = 1) pr(w_k = 1 | \check{w}_l = 0)}.$$

In this article, we use causal effect to mean the full function  $pr(w_k | \check{w}_l)$ , although we also consider the properties of particular causal measures, such as the causal odds ratio.

The probability  $pr(w_k | \check{w}_l)$  is fundamentally different from  $pr(w_k | w_l)$ . The former is the probability of observing  $W_k = w_k$  given that we force  $W_l$  to take the value  $w_l$ , whereas the latter is the probability of observing  $W_k = w_k$  given that we happen also to observe  $W_l = w_l$  (*i.e.* the familiar conditional probability function). Consider the variables ‘Exposure to cosmic radiation’ ( $A$ ), ‘Age’ ( $L$ ) and ‘Skin cancer’ ( $Y$ ) in Figure 1, but—for the sake of this discussion—let us assume that there is no arrow from  $A$  to  $Y$ , *i.e.* no causal effect of the exposure on the outcome. Suppose that older people tend to have a higher exposure and a higher incidence of skin cancer. Now suppose we could intervene on  $A$  by coating all aeroplanes in a substance that absorbs cosmic radiation, thereby setting  $A = 0$  for all subjects. According to our causal diagram (together with the assumption that  $A$  has no causal effect on  $Y$ ), this intervention would have no effect on  $Y$ . Thus, knowing that this intervention had been performed and that therefore a subject had zero exposure would tell us nothing about  $pr(Y = 1)$ . That is, according to Figure 1 (and the additional assumption of no causal effect),  $pr(Y = 1 | \check{a}) = pr(Y = 1)$ . However, if—rather than intervening on  $A$ —we merely *observe* that a particular subject has a low exposure, this tells us that the subject is likely to be younger and thus less likely to have skin cancer, that is  $pr(Y = 1 | a) \neq pr(Y = 1)$ . That  $pr(w_k | w_l) \neq pr(w_k | \check{w}_l)$  is a mathematical representation of the phrase “association is not causation”.

**Definition 7** (conditional causal effect). The *conditional causal effect* of  $W_l$  on  $W_k$ , given  $W_m$ , is denoted  $pr(w_k | \check{w}_l, w_m)$ , and is defined as the conditional probability of  $W_k = w_k$  given  $W_m = w_m$  induced by intervening on  $W_l$  and setting its value to  $\check{w}_l$ .

### 3.2 Algorithm for determining whether or not the causal effect of $A$ on $Y$ given $Z$ can be identified from the complete records

1. Draw a causal diagram ( $\mathcal{G}$ ) for the problem.
2. Extend it (to  $\mathcal{G}^+$ ) by adding parents of  $A$  and parents of descendants of  $A$  (except for parents of  $R$ ).
3. Take  $\mathcal{G}^+$  and draw a dashed line between any pair of variables that are both parents of  $R$ , or that share a child which is an ancestor of  $R$ . This is  $\mathcal{M}^+$ .

4. Draw a dashed line between any pair of variables that are both parents of a variable in  $\mathcal{Z}$ , or that share a child which is an ancestor of a variable in  $\mathcal{Z}$ .
5. Look for a generalised path (where a generalised path can consist of arrows in any direction and dashed lines) from  $A$  to  $Y$ , not passing through  $R$ , that either (i) starts with an arrow *into*  $A$ , or (ii) contains a dashed line. Does it contain a collider, and/or pass through a member of  $\mathcal{Z}$ ? If the answer is ‘yes’ for every generalised back-door path, **condition 1. is satisfied**. We call this condition the *generalised back-door criterion*.
6. Return to  $\mathcal{G}$ . Remove all arrows *into*  $A$ .
7. Draw a dashed line between any pair of variables that are both parents of a variable in  $\mathcal{Z}$ , or that share a child which is an ancestor of a variable in  $\mathcal{Z}$ .
8. Look for a path from  $R$  to  $Y$ , not passing through  $A$ . Does it contain a collider, and/or pass through a member of  $\mathcal{Z}$ ? If the answer is ‘yes’ for every such path, **condition 2. is satisfied**.

If conditions 1 and 2 are satisfied then we show in the Web Appendix that the causal effect of  $A$  on  $Y$  given  $\mathcal{Z}$  can be identified from the complete records alone.

Intuitively, condition 1. ensures that any association seen between  $A$  and  $Y$  is causal. Suppose this association is estimated from a generalised linear model, then the coefficient of  $A$  can be given a causal interpretation if the first condition holds. Condition 2. concerns the ‘intercept’. In order to identify  $pr(y|\check{a}, z)$  from the observed data, we must additionally be able to estimate the distribution of  $Y$  under the intervention  $\check{a} = 0$ , and for this condition 2. is needed.

## 4 Examples: applying the algorithm to various incomplete data settings

In this section, we look at examples of missing data mechanisms, use causal diagrams to represent them, and demonstrate how the algorithm above confirms and clarifies our understanding of the suitability of complete records analyses in these settings.

### 4.1 Missingness completely at random (MCAR)

An important distinction in the missing data literature is that between *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR).<sup>7</sup> A variable  $Z$  is MCAR if the probability that  $Z$  is observed, given the full data  $\mathcal{V}_0$ , is independent of  $\mathcal{V}_0$ . The MCAR assumption says that  $pr(R = 1 | v_0)$  is independent of the value of  $v_0$ . Suppose that the missing questionnaires in our motivating example were missing as a result of a postal strike. It might then be reasonable to assume that the missing questionnaires are MCAR, *i.e.* that the fact that a questionnaire is missing is unrelated to the unseen answers written on that questionnaire.

In a causal diagram for this simple situation,  $R$  would be isolated from the rest of the graph. Thus, conditioning on  $R = 1$  would have no consequence. More formally, the empty set satisfies the generalised back-door criterion, and there are no paths from  $Y$  to  $R$ . Thus, both conditions (§3.2) are satisfied with  $\mathcal{Z} = \emptyset$ . This confirms that a complete records analysis is valid when the mechanism is MCAR.

### 4.2 Missingness driven only by exposure

Figure 3(1a.) represents a causal diagram ( $\mathcal{G}$ ) with an arrow from  $A$  to  $R$  denoting that the probability of an incomplete record depends on  $A$ .

This includes the situation in which  $Y$  is *missing at random* given (fully-observed)  $A$ . To define the term *missing at random* we must first be more specific about what  $R = 0$  implies. In particular, we suppose that  $Y$  is missing for some subjects, and that there is at least one variable in  $\mathcal{V}_0$  which

is completely-observed.  $Y$  is *missing at random* (MAR) if the probability that  $R = 1$ , given the full data  $\mathcal{V}_0$ , is a function only of the observed part of  $\mathcal{V}_0$ , and not of the potentially missing value of  $Y$ . Suppose that  $A$  is employment status, and we believe that retired pilots were less likely to return the questionnaire than those still employed, but that apart from this, non-response was not related to any other variable, then assuming that data on current employment status are available from the CAA database, the data would be MAR given employment status.

Figure 3(1a.) also includes the situation in which  $A$  is incomplete and the missingness mechanism depends only on  $A$ . This is a case of  $A$  being *missing not at random*: if  $A$  is neither MCAR nor MAR,  $A$  is MNAR. Given the full data, the probability that  $A$  is missing depends on the potentially unobserved value of  $A$ . Figure 3(1a.) is a special case of MNAR, in which missingness depends *only* on  $A$ . An example is if high exposure to cosmic radiation increases the probability that pilots return their questionnaires (perhaps since increased exposure leads to more interest in the study).

The extended causal diagram ( $\mathcal{G}^+$ ) corresponding to Figure 3(1a.) is shown in Figure 3(1b.). Conditioning on  $R = 1$  in the modified extended diagram (shown in Figure 3(1c.)) does not introduce any dashed lines. Again, the empty set satisfies the generalised back-door criterion, and the only path from  $Y$  to  $R$  passes through  $A$ . Thus, both conditions (§3.2) are satisfied by  $\mathcal{Z} = \emptyset$ .

Without using causal diagrams, the implications for analysis of different missingness mechanisms can be illustrated using a simple artificial example with two continuous variables, an exposure  $A$  and an outcome  $Y$ , where  $Y$  is (apart from random error) a linear function of  $A$  (see Figure 4). If  $Y$  is MCAR, then the complete records form a random subset of the full data, and any aspect of the joint distribution of  $A$  and  $Y$  (such as the mean of  $Y$  or the causal effect of  $A$  on  $Y$ ) can be consistently estimated using the complete records.

This is not true when  $Y$  is MAR given  $A$ . If  $Y$  is more likely to be missing for high values of  $A$ , then the mean of the observed  $Y$ -values will be biased downwards as an estimate of the mean of  $Y$ . However, the causal effect of  $A$  on  $Y$  *can* be consistently estimated (for example using a linear regression of  $Y$  on  $A$ ) using only the complete records. This can be seen in Figure 4(A.),

where  $Y$  is not observed if  $A > 2.4$  (hence MAR). The cut-off line drawn at  $A = 2.4$  does not distort the unexplained variation in  $Y$ . Likewise, if  $A$  is MNAR dependent only on  $A$ , then the mean of the observed  $A$ -values will be biased but the causal effect of  $A$  on  $Y$  can be consistently estimated from the complete records.

Such extreme “cut-off” mechanisms are unlikely to occur in practice but their simplicity helps to illustrate the features also present in more plausible mechanisms.

The fact that, when exposures and/or covariates are MNAR given only themselves, a complete records analysis is valid, is not always well-understood, but has been demonstrated, for example, by Rathouz<sup>12</sup> and confirmed in simulation studies by Giorgi et al.<sup>13</sup> Illustrating the missingness mechanism using a causal diagram makes this considerably clearer.

### 4.3 Missingness driven by exposure and covariates

Figure 3(2a.) contains an additional variable  $L$  predictive both of missingness and  $Y$ . Without conditioning on  $R$  (*e.g.* when data are complete), an unadjusted analysis for the causal effect of  $A$  on  $Y$  is unbiased since there are no open back-door paths through  $L$ . However, upon extending the diagram (Figure 3(2b.)) and conditioning on  $R = 1$  (Figure 3(2c.)), an association is induced between  $A$  and  $L$  through conditioning on their common child,  $R$ . This opens up the generalised back-door path  $A - -L \rightarrow Y$  and thus, for a complete records analysis to be unbiased, we must condition on  $L$ . In fact  $\mathcal{Z} = \{L\}$  satisfies both conditions (§3.2) and thus  $pr(y|\check{a}, l)$  can be identified from the complete records alone.

This is consistent with the advice often given, that “if a variable is predictive of both outcome and missingness, it should be appropriately incorporated into the analysis”.

The marginal causal effect  $pr(y|\check{a})$  is related to the conditional causal effect via

$$pr(y|\check{a}) = \sum_l pr(y|\check{a}, l) pr(l)$$

but  $pr(l)$  is not identifiable from the complete records, since  $L$  directly affects

$R$ . Thus this marginal causal effect cannot be identified.

#### 4.4 Missingness driven by both outcome and exposure

Figure 3(3a.) shows missingness depending on both  $Y$  and  $A$ . Conditioning on  $R = 1$  creates (see Figure 3(3c.)) a dashed line between  $A$  and  $Y$ , which is itself a generalised back-door path, as well as the generalised back-door path  $A - -U_2 \rightarrow Y$ . Neither of these contains a measured variable and thus neither can be blocked. The conditions (§3.2) are not satisfied by any  $\mathcal{Z} \in \mathcal{V}_0$ . Whether  $A$  or  $Y$  (or both) is missing, this causal diagram represents a MNAR mechanism, where missingness depends on both  $A$  and  $Y$ . Complete records analyses are not valid in such settings and thus we should consider the robustness of our inferences to plausible MNAR mechanisms using sensitivity analyses (see, for example, Molenberghs and Kenward<sup>14</sup>, part V).

#### 4.5 Missingness driven only by outcome

When missingness depends only on  $Y$  (see Figure 3(4a.)), conditioning on  $R = 1$  (see Figure 3(4c.)) does not induce a dashed line between  $A$  and  $Y$ , but the generalised back-door path  $A - -U_2 \rightarrow Y$  is still created and cannot be blocked, and thus the conditions (§3.2) remain violated. Note that this generalised back-door path would not have been uncovered had we not started by extending  $\mathcal{G}$  to include  $U_2$ .

Returning to Figure 4(B.), the same picture is revealed. Using the line  $Y = 6$  as a cut-off *does* distort the unexplained variation, and there is a corresponding bias in the estimate of the causal effect. Extremely high  $Y$ -values for a given  $A$  are not observed under this MNAR mechanism, whereas extremely low  $Y$ -values for a given  $A$  are observed. This causes the attenuation seen in the estimate of the causal effect. This is mirrored in Figure 3(4c.): by conditioning on  $R = 1$ , we induce an association between  $U_2$  and  $A$ . When both  $U_2$  and  $A$  are positively correlated with  $R$ ,  $U_2$  and  $A$  will be negatively associated within strata of  $R$  and the role of  $U_2$  in a complete records analysis will be similar to that of a negative confounder for the causal effect of  $A$  on  $Y$ .

If the outcome is fully-observed, then principled methods for MAR in-

complete data, such as direct likelihood,<sup>8</sup> multiple imputation,<sup>9,10</sup> or inverse probability weighting<sup>11</sup> might be considered. However, if the outcome is incomplete, then sensitivity analyses would again be advisable.

One apparent exception to the discussion relating to Figure 3(4a.) is a well-conducted case-control study, where patients are selected with different probabilities according to the binary outcome  $Y$  (case or control) but, as a consequence of its reversibility, the odds ratio for the effect of  $A$  on  $Y$  in the selected subjects is known to be unbiased.<sup>15</sup> The causal diagram for a case-control study is precisely that seen in Figure 3(4a.). There is no contradiction here: when the causal diagram suggests that a particular causal effect ( $pr(y|\tilde{a})$ ) is estimated with bias, this does not exclude the possibility that a particular many-to-one function of this causal effect (in this case, the causal odds ratio) could be estimated without bias.

Another case requiring special attention is Figure 3(4a.) when the arrow from  $A$  to  $Y$  is removed (when the causal null hypothesis holds). In this case, condition 1. is satisfied, but condition 2. is not.  $A$  and  $Y$  remain independent even after conditioning on  $R = 1$  (by condition 1.), but the distribution of  $Y$  is distorted (by the arrow from  $Y$  to  $R$ —condition 2.), and thus  $pr(y|\tilde{a}) = pr(y)$  cannot be estimated without bias from the complete records. This agrees with the intuition given in §3.2.

Figure 3(5a.) is the same as Figure 3(4a.), except that a measured variable  $L$  has been added, which affects  $Y$  but nothing else.  $L$  can be thought of as a measured component of  $U_2$ . Informally, conditioning on  $R = 1$  induces confounding through both  $L$  and  $U_2$ . We can condition on  $L$  and eliminate some of the bias. While this is in accord with another common piece of advice, “condition on as many covariates as possible to get closer to MAR”, the theory of causal diagrams exposes the potential danger associated with this way of thinking: adjusting for variables which are affected by the exposure and/or outcome can introduce bias (*cf.* conditioning on  $R$  in Figures 3(2a.–5a.)). Assuming that we take care to avoid introducing bias in this way, controlling for as many variables predictive of  $Y$  as possible is beneficial as it reduces the unexplained variation. This is analogous to reducing the ‘spread’ of the points about the straight line in Figure 4(B.). This reduces (but does not eliminate) the bias in the coefficient of the exposure in the estimate of the causal effect.

## 4.6 Missingness driven only by covariates

Finally, Figure 3(6a.) is the same as Figure 3(2a.), except that the effect of the exposure on  $R$  has been replaced with an effect of a cause  $M$  of the exposure on  $R$ . Thus, missingness is driven only by the covariates  $M$  and  $L$ , neither of which is a confounder of the relationship between  $A$  and  $Y$ , but one of which affects the exposure, the other the outcome. This sort of causal diagram has been the focus of many discussions.<sup>2</sup> Conditioning on  $R = 1$  (see Figure 3(6c.)) induces an association between  $M$  and  $L$ , implying that we must additionally condition on either  $M$  or  $L$  (or both) in order for condition 1. (§3.2) to be satisfied. However, for condition 2. to be satisfied, we must condition on  $L$ . In this case, the conditional causal effect of  $A$  on  $Y$  given  $L$  can be estimated from the complete records (but not the conditional causal effect of  $A$  on  $Y$  given  $M$ ). The symmetry of Figure 3(6a.) with respect to  $L$  and  $M$  is misleading; when designing a study in such a situation, it would be far more important to plan measurement of  $L$  than of  $M$ .

## 5 Application: the British Commercial Airline Pilots and ATCOs Study

In the light of §3.2, we return to Figure 1 to give a more formal interpretation. First, in order to change the causal diagram to an *extended* causal diagram, we should include two additional nodes: one representing all causes of ‘Exposure to cosmic radiation’ and the other representing all other causes of ‘Skin cancer’ (not already in the diagram). It transpires in this case that these additional nodes have no bearing on any subsequent argument, and thus we have omitted them.

In addition to checking for unblocked generalised back-door paths (condition 1.), we must also check condition 2.: that there be no unblocked paths from ‘Skin cancer’ to  $R$  except through ‘Exposure to cosmic radiation’.

We had already seen that the first condition given in §3.2 could not hold in Figure 1. More formally, there are two generalised back-door paths that cannot be blocked (‘exposure—Unmeasured behavioural factors→skin cancer’ and ‘exposure—Unmeasured behavioural factors←Unmeasured genetic factors→skin cancer’), while the paths via age and red hair can be blocked

by conditioning on these variables.

Note that ‘Red hair’ plays a similar rôle to  $L$  in Figure 3(5a.) and the same argument for adjusting for hair colour (along with age) to reduce some of this bias applies.

Condition 2. also fails, since two paths from skin cancer to  $R$  remain open (‘skin cancer←Unmeasured behavioural factors→ $R$ ’ and ‘skin cancer←Unmeasured genetic factors→Unmeasured behavioural factors→ $R$ ’). Collecting details on behavioural risk factors such as use of sun beds and hours spent sunbathing, and conditioning on these variables, would therefore be required to reduce the bias induced by missingness.

A discussion of other possible causal diagrams for this example is included in the Web Appendix.

## 6 Discussion

In this article we have described a general graphical tool giving sufficient conditions under which the causal effect of an exposure  $A$  on an outcome  $Y$  can be identified (possibly conditionally on other variables  $\mathcal{Z}$ ) using only the collected variables ( $\mathcal{V}_0$ ) in the subjects with complete records. Although more sophisticated approaches than merely a complete records analysis are readily available and are, in general, to be advocated, it is important to know when a complete records analysis would suffice.

We reviewed (briefly, in §2.2) the use of causal diagrams to adjust for confounding using the back-door criterion. In §3, we extended this algorithm to the missing data setting. The theory is given in the Web Appendix. Further work is required, in particular with regards to the necessity of the conditions stated in §3.2. Nevertheless, we have shown these conditions to be sufficient and conjecture that they are also necessary.

Our approach is fully-integrable with the existing causal diagrams framework to deal with confounding. For this, the original causal diagram  $\mathcal{G}$  is sufficient. In moving from  $\mathcal{G}$  to  $\mathcal{M}^+$ , we have included additional nodes and additional (dashed) lines, but no node nor arrow has been removed, and thus the identification of the variables to control for confounding is not af-

fects (*cf.* age in Figure 1). When the data are incomplete, our algorithm encompasses the original back-door criterion found in Pearl<sup>1</sup> and Greenland et al.<sup>2</sup>

We considered a possible causal diagram for our motivating example (others are discussed in the Web Appendix), a study of the effect of exposure to cosmic radiation on skin cancer incidence in a population of airline pilots and air traffic control officers. Given the assumptions of Figure 1, we showed that a complete records analysis of these data, even after adjusting for age and hair colour, would be biased. When planning another similar study, Figure 1 could be used to identify new questions to be included in the questionnaire, to provide observed variables on some of these generalised back-door paths.

As with all graphical approaches, the conclusions are only valid if the assumptions implied by the diagram are close to being correct. In the absence of good background knowledge of the subject area and the plausible causal mechanisms at play, including knowledge of the mechanisms giving rise to incomplete data, it is unfeasible to attempt a causally-interpretable analysis of the data.

One feature of our approach is that it concerns full causal effects rather than particular causal measures. It is possible for a particular causal measure (such as the causal odds ratio in a case-control study) to be identifiable from the complete records even when the full probability distribution of the outcome under varying exposure levels (*i.e.* the causal effect) cannot be identified. This is a consequence of the non-parametric nature of causal diagrams. The advantage of causal diagrams is their generality. They can be used to illustrate simply the relationships between many variables, and, using general rules, to focus attention on variables that lie on important pathways. The price to be paid for this generality is the lack of sensitivity to properties of particular causal measures, which must be established independently.

In summary, we have shown how Pearl's theory of causal diagrams can be used to determine whether a causal effect can be estimated without bias by analysing only the subjects with complete data. When this is not possible, the modified extended diagrams introduced in this article provide an intuitive tool to help understand how and why a complete records analysis is biased.

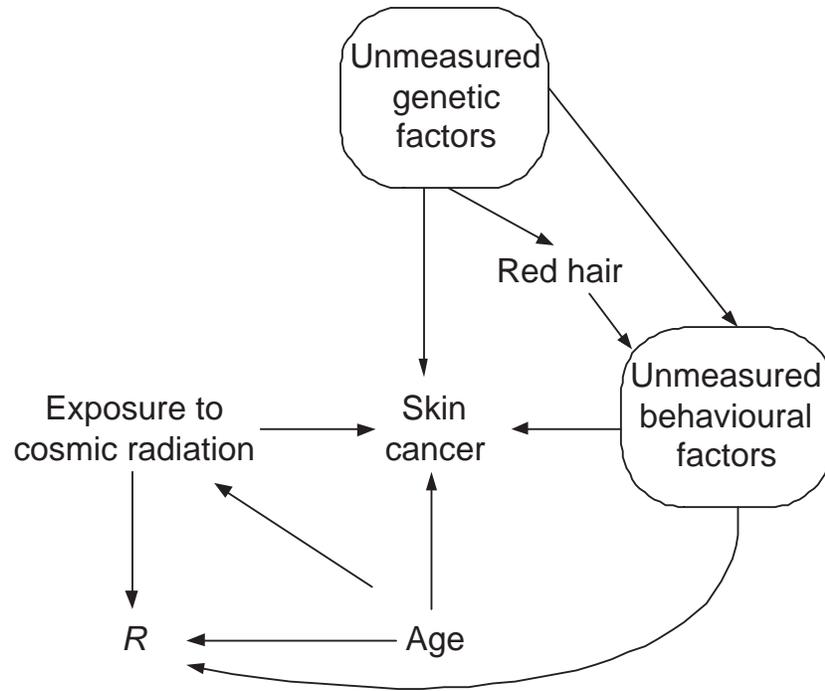
## Acknowledgements

We would like to thank Prof. Stijn Vansteelandt for some illuminating discussion on this work, Prof. Isabel dos Santos Silva for her guidance on the British airline pilots and ATCOs study used as an example in this paper, and Prof. John Whittaker for his comments on an earlier version of the paper. We are also extremely grateful to two anonymous referees and a member of the editorial board for their very careful and constructive comments which have greatly improved the main article and the theoretical details given in the Web Appendix. This work was supported by a grant from the Medical Research Council, UK (Grant number: G0701024).

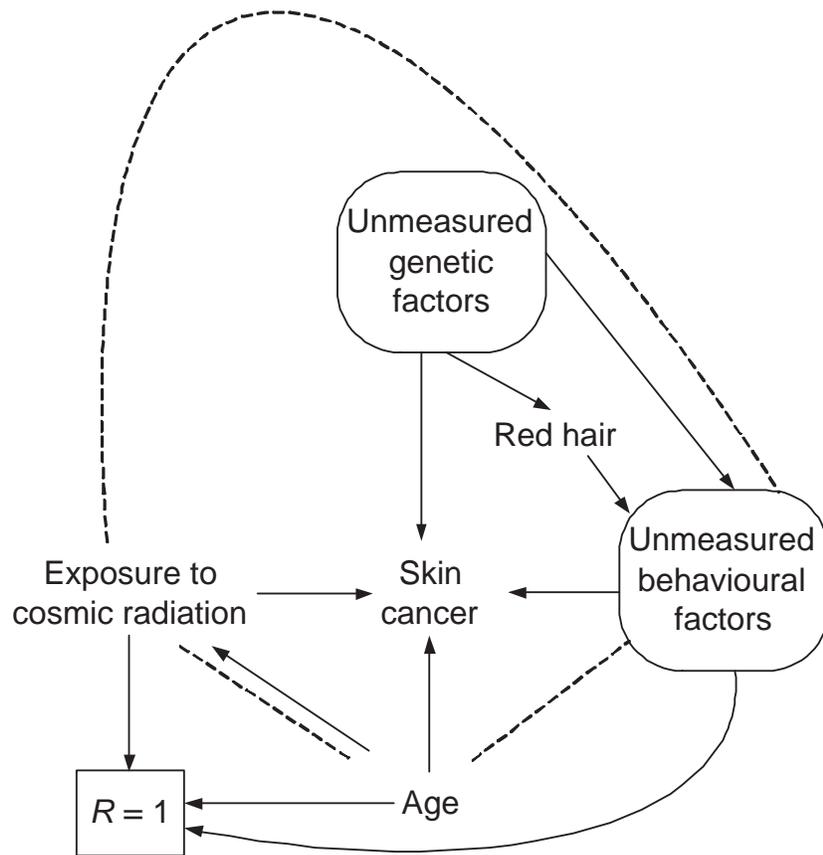
## References

1. Pearl J. Causal diagrams for empirical research. *Biometrika* 1995; **82**:669–709.
2. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiological research. *Epidemiology* 1999; **10**:37–48.
3. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**:615–625.
4. Pizzi C, Evans SA, De Stavola BL, Evans A, Clemens F, dos Santos Silva I. Lifestyle of UK commercial aircrews relative to air traffic controllers and the general population. *Aviation, Space, and Environmental Medicine* 2008; **79**:964–974.
5. Pearl J. *Causality*. 2nd ed. Cambridge University Press, 2009.
6. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. *Formalising the use of causal diagrams in missing data problems*. Technical Report, London School of Hygiene and Tropical Medicine, 2010.
7. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.
8. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley, 2002.

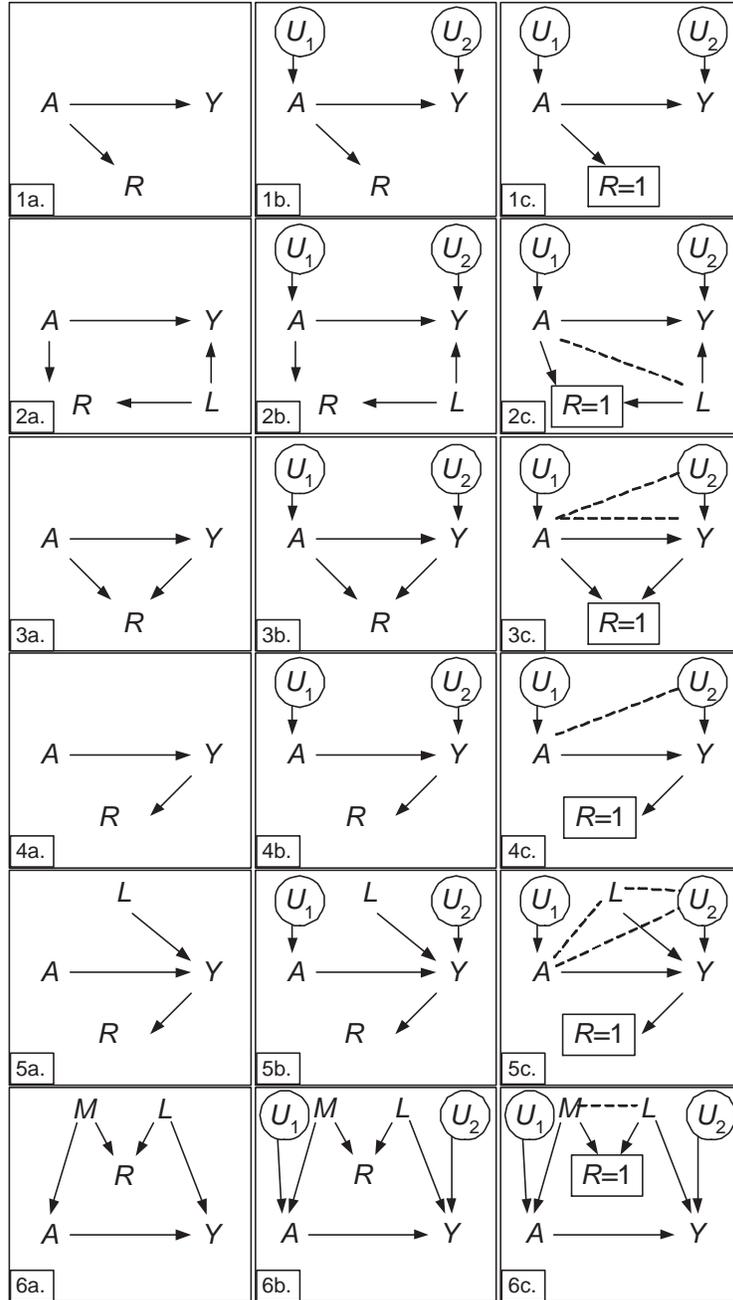
9. Rubin DB. Multiple imputations in sample surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 1978, pp. 20–34.
10. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**:473–489.
11. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
12. Rathouz PJ. Identifiability assumptions for missing covariate data in failure time regression models. *Biostatistics* 2007; **8**:345–356.
13. Giorgi R, Belot A, Gaudart J, Launoy G, and the French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. *Statistics in Medicine* 2008; **27**:6310–6331.
14. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Wiley, 2007.
15. Cornfield J. A method of estimating comparable rates from clinical data: applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* 1951; **11**:1269–1275.



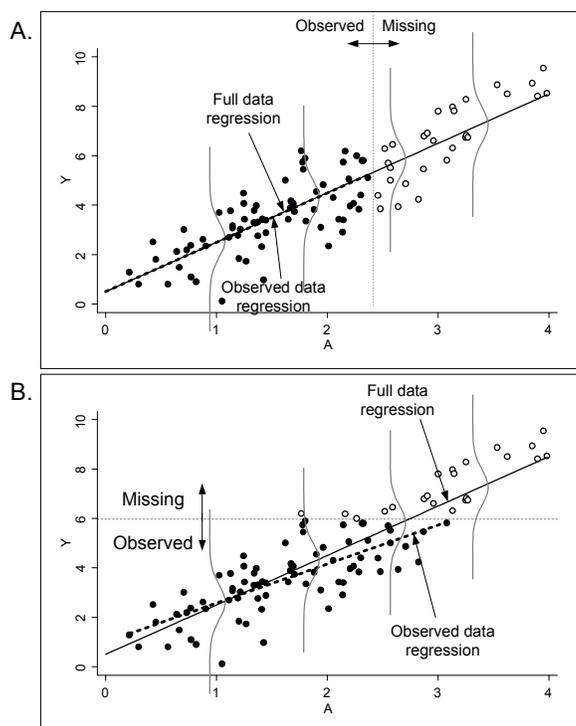
**Figure 1:** A possible causal diagram for the airline pilots study.



**Figure 2:** A modified diagram, derived from Fig. 1 after conditioning on  $R = 1$ .



**Figure 3:** The first column shows the causal diagrams associated with various causal missingness mechanisms. In the second column, 1b.–6b. are extended causal diagrams corresponding to each of 1a.–6a. In the third column, 1c.–6c. are modified diagrams corresponding to each of 1b.–6b., showing the effect of conditioning on  $R = 1$  on the relationship between other variables in the diagram.



**Figure 4:** A comparison of (i) the full data estimate and (ii) the complete records estimate of the causal effect of  $A$  on  $Y$  under two different missingness mechanisms: missingness depends on  $A$  in panel A. and missingness depends on  $Y$  in panel B.

## Web Appendix

# Using Causal Diagrams to Guide Analysis in Missing Data Problems

Rhian M. Daniel, Michael G. Kenward, Simon N. Cousens  
and Bianca L. De Stavola

Faculty of Epidemiology and Population Health,  
London School of Hygiene and Tropical Medicine

### Abstract

This Web Appendix gives the theoretical details behind the algorithm given in §3.2 of the main article, along with a discussion of other possible causal diagrams for the motivating example.

## 1 Introduction

Let  $\mathcal{G} = \{\mathcal{N}, \mathcal{A}\}$  be a causal diagram representing the relevant variables for a particular problem and the causal influences believed to exist between them.  $\mathcal{N}$  is the set of nodes, or variables, and  $\mathcal{A}$  is the set of ordered pairs, or directed arrows, that exist between members of  $\mathcal{N}$ . Let  $\mathcal{V}_0$  be the subset of  $\mathcal{N}$  which would have been observed on all subjects had there been no missing data.

Let  $R \in \mathcal{V}_0$  be a binary variable taking the value 1 if all variables in  $\mathcal{V}_0$  are observed, and 0 otherwise. According to the definition of a causal diagram (see Pearl, 1995<sup>1</sup>), any variable (whether observed or not) believed to be a common cause of two variables in  $\mathcal{N}$  should itself be in  $\mathcal{N}$ .

Let  $A \in \mathcal{V}_0$  be the exposure of interest, and  $Y \in \mathcal{V}_0$  be the outcome of interest.

Our aim is to determine from causal diagrams such as  $\mathcal{G}_1$  and  $\mathcal{G}_2$  shown in Figure 1 whether or not the causal effect of  $A$  on  $Y$  can be consistently estimated using only subjects with complete data on  $\mathcal{V}_0$ . In other words, can this causal effect be estimated using only variables in  $\mathcal{V}_0$  once we have conditioned on  $R = 1$ ?

Using the knowledge that conditioning on a common effect of two independent causes (*i.e.* a *collider* or a descendant of a collider) induces an association between them within strata of the conditioning variable, we might try to modify  $\mathcal{G}_1$  and  $\mathcal{G}_2$  as shown in Figure 2. Looking at the left-hand side of this figure, we see that conditioning on  $R = 1$  has induced an association between  $A$  and  $L$  not present in the full data. Thus, in the subset of the data represented by the complete records,  $A$  is associated with  $L$  and  $L$  has a causal effect on  $Y$ . Even if  $A$  has no causal effect on  $Y$ , the two variables would be associated among the complete records, by virtue of their association with  $L$ . This suggests that we should control for  $L$  in any complete records analysis.

On the other hand, conditioning on  $R = 1$  in  $\mathcal{G}_2$  appears not to have induced any new associations. However, if we extend  $\mathcal{G}_2$  to include the unmeasured residual error  $U$  in  $Y$ , then conditioning on  $R = 1$  induces an association between  $A$  and  $U$  (see Figure 3) which appears to be detrimental to our estimation of the causal effect of interest, since we cannot control for  $U$ .

Informally, we are modifying  $\mathcal{G}_1$  and  $\mathcal{G}_2$  to represent what happens when we condition on  $R$ . Then, we are looking for ‘backdoor paths’ from  $A$  to  $Y$  ( $A - -L \rightarrow Y$  and  $A - -U \rightarrow Y$ , respectively) and looking for variables on these paths that could be controlled in order to block these paths.

But there are several issues with this informal approach. First, the inclusion of  $U$  in  $\mathcal{G}_2$  seems essential to the subsequent argument, even though  $U$  is not required to be in the diagram according to the definition of a causal diagram. Second, we are attempting to apply the backdoor criterion to a modified diagram, containing associational lines as well as causal arrows. This is not well-defined. Put another way, we are applying the backdoor criterion including  $R$  in the conditioning set, when  $R$  is a consequence of the

exposure, something which is not permitted by the backdoor criterion as set out by Pearl<sup>1</sup> and Greenland et al<sup>2</sup>.

Our aim in this note is to define a new class of causal diagrams, which we call *extended* causal diagrams, together with the rules for their construction, so that they include all the additional nodes needed for missing data problems. Then, we will extend the backdoor criterion for use in modified extended diagrams, containing associational lines as well as causal arrows. Finally, we will prove that this extended backdoor criterion when applied to modified extended diagrams, in conjunction with one additional condition, is sufficient for determining when causal effects can be identified from incomplete data using only the complete records.

## 2 Preliminary definitions

**Definition 1.** A *causal diagram*  $\mathcal{G}$  consists of nodes (or points) denoting variables, and arrows between nodes denoting the assumed direction of causal influence. Any variable which is the common cause of two or more variables in  $\mathcal{G}$ , must itself be in  $\mathcal{G}$ .

We write  $\mathcal{G} = \{\mathcal{N}, \mathcal{A}\}$  where  $\mathcal{N}$  is the set of nodes, or variables, in  $\mathcal{G}$ ; and  $\mathcal{A}$  is the set of ordered pairs (*i.e.* directed arrows) that exist between members of  $\mathcal{N}$  in  $\mathcal{G}$ . Let  $\mathcal{V}_0$  be the subset of  $\mathcal{N}$  which would have been observed on all subjects had there been no missing data. Thus  $R$  takes the value 1 if all variables in  $\mathcal{V}_0$  are observed, and 0 otherwise. We assume that  $A, Y, R \in \mathcal{V}_0$ , that is we must include  $A, Y$  and  $R$  in our diagram.

**Definition 2** (path). If  $W_1$  and  $W_m$  are disjoint nodes in  $\mathcal{G}$ , a *path*  $W_1W_2 \dots W_m$  from  $W_1$  to  $W_m$  is a sequence of nodes such that, for each  $k = 1, \dots, m - 1$ , there is either an arrow from  $W_k$  to  $W_{k+1}$  or from  $W_{k+1}$  to  $W_k$  in  $\mathcal{G}$ .

**Definition 3** (directed path). If  $W_1$  and  $W_m$  are disjoint nodes in  $\mathcal{G}$ , a *directed path*  $W_1W_2 \dots W_m$  from  $W_1$  to  $W_m$  is a sequence of nodes such that, for each  $k = 1, \dots, m - 1$ , there is an arrow from  $W_k$  to  $W_{k+1}$  in  $\mathcal{G}$ .

**Definition 4.** If there is an arrow from  $W_i$  to  $W_j$  in  $\mathcal{G}$ ,  $W_j$  is said to be a *child* of  $W_i$ , and  $W_i$  a *parent* of  $W_j$ . If there is a directed path from  $W_i$  to  $W_j$  in  $\mathcal{G}$ ,  $W_j$  is said to be a *descendant* of  $W_i$ , and  $W_i$  an *ancestor* of  $W_j$ .

Each child-parent family in  $\mathcal{G}$  (containing  $n$  nodes  $W_1, \dots, W_n$ ) corresponds to a deterministic function

$$W_i = f_i(pa(W_i), \varepsilon_i) \quad i = 1, \dots, n \quad (1)$$

from a nonparametric structural equations model, where  $\{\varepsilon_i : i = 1, \dots, n\}$  are independent unobserved random disturbances, the distributions of which remain unspecified, and  $pa(W_i)$  are the parents of  $W_i$  in  $\mathcal{G}$ . In other words, the value of  $W_i$  does not follow deterministically from the value of its parents, because of the unobserved random disturbance  $\varepsilon_i$ , but as a function of its parents and  $\varepsilon_i$ , the relationship is deterministic and given by the function  $f_i(\cdot)$ .

Pearl<sup>1</sup> argues that this formulation leads naturally to the notion of *intervening* on a particular variable in the model. For example, setting the value of  $W_j$  to  $w_j$  is equivalent to removing  $W_j = f_j(pa(W_j), \varepsilon_j)$  from (1) and replacing  $W_j$  with  $w_j$  in all the other equations.

**Definition 5** (do operator). The notation  $\check{w}_i$  is used to denote the act of *intervening* on  $W_i$  and setting its value to  $w_i$ .  $W_i = \check{w}_i$  is verbalised “do  $W_i$  equals  $w_i$ ” or “set  $W_i$  equal to  $w_i$ ”.

**Definition 6** (causal effect). For any  $l \neq k$ , the *causal effect* of  $W_l$  on  $W_k$ , denoted  $pr(w_k | \check{w}_l)$ , is a function from  $W_l$  to the space of probability distributions on  $W_k$ . For each realisation  $w_l$  of  $W_l$ ,  $pr(w_k | \check{w}_l)$  gives the probability of  $W_k = w_k$  induced by intervening on  $W_l$  and setting its value to  $w_l$ . This probability is calculated by removing  $W_l = f_l(pa(W_l), \varepsilon_l)$  from (1) and replacing  $W_l$  with  $w_l$  in all the other equations.

**Definition 7** ( $A$  and its descendants). Let  $d_{\mathcal{G}}(A)$  be the set consisting of  $A$  and its descendants in  $\mathcal{G}$ .

**Definition 8** (extended causal diagram). Let  $\mathcal{G}^+$  be the graph formed by adding to  $\mathcal{G}$  any node which is a parent of  $d_{\mathcal{G}}(A) \setminus \{R\}$  i.e. any node which is a parent of  $A$  or a parent of a descendant of  $A$  (except for parents of  $R$ ). This includes unobserved, and even unobservable, nodes, i.e. those which we might think of as ‘residual error’/‘random error’/ ‘chance’, so that these nodes that are descended from  $A$  (except for  $R$  if it is one of them) and  $A$  itself are no longer ‘probabilistic nodes’ in the diagram—their values are exactly determined by their (possibly unobservable) parents. We could include parents of  $R$  here too, but as they are not needed, it is simpler to omit them.

The extended causal diagrams  $\mathcal{G}_1^+$  and  $\mathcal{G}_2^+$  corresponding to  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are shown in Figure 4.

**Definition 9** (modified extended diagram). Take  $\mathcal{G}^+$  and join together— with a dashed line—any pair of variables that are both parents of  $R$ , or that share a child which is an ancestor of  $R$ . [Note that even if such a pair is already connected by an arrow (in either direction), a dashed line should be added as well.] The resulting diagram  $\mathcal{M}^+$  is known as the modified extended diagram.

The modified extended diagrams  $\mathcal{M}_1^+$  and  $\mathcal{M}_2^+$  corresponding to  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are shown in Figure 5.

**Definition 10** (generalised path). A generalised path  $p$  in a modified extended diagram  $\mathcal{M}^+$  is a sequence of variables  $X_1, \dots, X_n \in \mathcal{M}^+$  such that each pair  $(X_i, X_{i+1})$  is either connected by an arrow (in either direction) or a dashed line.

Note: A collider on a generalised path  $p$  is still defined as a variable  $X$  having converging arrows along  $p$ , so  $X_2$  is *not* a collider on the path  $X_1 - -X_2 \leftarrow X_3$ . Also, a descendant of  $X$  is still defined as a node  $W$  connected to  $X$  via a directed path from  $X$  to  $W$ , i.e.  $X_3$  is a descendant of  $X_1$  in the path  $X_1 \rightarrow X_2 \rightarrow X_3$ , but *not* in the path  $X_1 - -X_2 \rightarrow X_3$ .

**Definition 11** (blocking a generalised path). Let  $\mathcal{X}$ ,  $\mathcal{Y}$  and  $\mathcal{Z}$  be three disjoint subsets of nodes in a modified extended diagram  $\mathcal{M}^+$ . Let  $p$  be any generalised path from a node in  $\mathcal{X}$  to a node in  $\mathcal{Y}$ .  $\mathcal{Z}$  blocks  $p$  if there is a node  $W$  on  $p$  satisfying one of the following:

1.  $W$  is a collider on  $p$  and neither  $W$  nor any of its descendants are in  $\mathcal{Z}$ .
2.  $W$  is not a collider on  $p$  and  $W$  is in  $\mathcal{Z}$ .

**Definition 12** (generalised backdoor criterion). A set of variables  $\mathcal{Z}$  satisfies the generalised backdoor criterion relative to  $(A, Y)$  in a modified extended diagram  $\mathcal{M}^+$  if

1. No node in  $\mathcal{Z}$  is a descendant of  $A$ .

2.  $\mathcal{Z}$  blocks every generalised path between  $A$  and  $Y$  which does not pass through  $R$  and which either
  - (a) starts with an arrow into  $A$ , or
  - (b) contains a dashed line [Note: this sort of a backdoor path can start with an arrow *out* of  $A$ ].

### 3 The main result

**Theorem 1.** If a set of variables  $\mathcal{Z} \subset \mathcal{V}_0$  satisfies the following two conditions:

1.  $\mathcal{Z}$  satisfies the generalised back-door criterion relative to  $(A, Y)$  in the modified diagram  $\mathcal{M}^+$ , and
2. having removed all arrows *into*  $A$  in  $\mathcal{G}$ , all paths from  $R$  to  $Y$  are blocked by  $\mathcal{Z} \cup \{A\}$

then  $pr(y | \check{a}, z)$  is identifiable from the observed data on the complete records alone.

### 4 Proof of the main result

We prove the result using two lemmas. In the first lemma, we use the notation from Pearl<sup>1</sup> section 4:  $\mathcal{G}_{\underline{X}}$  is the graph obtained from  $\mathcal{G}$  by deleting any arrow emanating from  $X$ , and  $\mathcal{G}_{\overline{X}}$  is the graph obtained from  $\mathcal{G}$  by deleting any arrow pointing towards  $X$ .

**Lemma 2.** Let  $\mathcal{V}_0 \subset \mathcal{N}$  be the subset of  $\mathcal{N}$  on which data have been collected on at least some subjects, i.e. a subject with a complete record would have data on all of  $\mathcal{V}_0$  and none of  $\mathcal{N} \setminus \mathcal{V}_0$ . If  $\{R, \mathcal{Z}\} \subset \mathcal{V}_0$  satisfies the following:

1.  $(Y \perp\!\!\!\perp A | R, \mathcal{Z})_{\mathcal{G}_{\underline{A}}}$
2.  $(Y \perp\!\!\!\perp R | A, \mathcal{Z})_{\mathcal{G}_{\overline{A}}}$

then  $pr(y|\check{a}, z)$  can be identified from the observed data on the complete records alone.

*Proof.* This follows from rules 1 and 2 in Theorem 3 of Pearl 1995<sup>1</sup>. Pearl's rule 2 states that if  $(Y \perp\!\!\!\perp A | R, \mathcal{Z})_{\mathcal{G}_A}$ , then

$$pr(y|\check{a}, r, z) = pr(y|a, r, z)$$

In particular, if  $(Y \perp\!\!\!\perp A | R, \mathcal{Z})_{\mathcal{G}_A}$ , then

$$pr(y|\check{a}, R = 1, z) = pr(y|a, R = 1, z)$$

Pearl's rule 1 states that if  $(Y \perp\!\!\!\perp R | A, \mathcal{Z})_{\mathcal{G}_A}$ , then

$$pr(y|\check{a}, r, z) = pr(y|\check{a}, z)$$

In particular, if  $(Y \perp\!\!\!\perp R | A, \mathcal{Z})_{\mathcal{G}_A}$ , then

$$pr(y|\check{a}, R = 1, z) = pr(y|\check{a}, z)$$

Putting both of these together, if  $(Y \perp\!\!\!\perp A | R, \mathcal{Z})_{\mathcal{G}_A}$  and  $(Y \perp\!\!\!\perp R | A, \mathcal{Z})_{\mathcal{G}_A}$  then

$$pr(y|\check{a}, z) = pr(y|a, R = 1, z)$$

We have re-written  $pr(y|\check{a}, z)$  as a quantity,  $pr(y|a, R = 1, z)$ , which involves only an association between observed variables measured on subjects with complete records.  $\square$

**Lemma 3.** If  $\mathcal{Z}$  satisfies the generalised backdoor criterion relative to  $(A, Y)$  in  $\mathcal{M}^+$  then condition 1. of Lemma 2 holds.

*Proof.* Suppose that this statement is false. Then, there must exist a graph  $\mathcal{G}$  and a set of variables  $\mathcal{Z}$  such that  $\mathcal{G}_A$  contains a back-door path  $p$  from  $A$  to  $Y$  not blocked by  $\{\mathcal{Z}, R\}$ , but that all generalised back-door paths from  $A$  to  $Y$  in the corresponding  $\mathcal{M}^+$  are blocked by  $\mathcal{Z}$ .

Consider such a path  $p$  in  $\mathcal{G}_A$ . It must start with an arrow into  $A$  (since all arrows out of  $A$  have been deleted in forming  $\mathcal{G}_A$ );  $p$  also exists in  $\mathcal{G}$ , and

in  $\mathcal{G}^+$  and hence in  $\mathcal{M}^+$ , since, in moving from  $\mathcal{G}_{\underline{A}}$  to  $\mathcal{G}$  to  $\mathcal{G}^+$  to  $\mathcal{M}^+$ , nodes and arrows are added, but never taken away.

If  $p$  is open (after conditioning on  $\{\mathcal{Z}, R\}$ ) in  $\mathcal{G}_{\underline{A}}$  but blocked by  $\mathcal{Z}$  in  $\mathcal{M}^+$ , then we must be in one of the following two scenarios:

1. A variable on  $p$  is a member of  $\mathcal{Z}$  (and hence blocks  $p$  in  $\mathcal{M}^+$ ) but not of  $\{\mathcal{Z}, R\}$  (and hence does not block  $p$  in  $\mathcal{G}_{\underline{A}}$ ). This is clearly absurd since  $\mathcal{Z} \subset \{\mathcal{Z}, R\}$ .
2.  $p$  contains a collider  $C$  such that either  $C$  belongs to  $\{\mathcal{Z}, R\}$  or has descendants in  $\{\mathcal{Z}, R\}$  (so that  $p$  is open in  $\mathcal{G}_{\underline{A}}$ ), but  $C$  does not belong to  $\mathcal{Z}$  and does not have descendants in  $\mathcal{Z}$  (so that  $p$  is blocked in  $\mathcal{M}^+$ ). This means that  $p$  must contain a collider  $C$  with either  $C = R$  or  $C$  being an ancestor of  $R$ . Either way,  $p$  corresponds to a generalised back-door path  $p'$  in  $\mathcal{M}^+$ , which is identical to  $p$  except that  $C$  is not on  $p'$  and the parents of  $C$  on  $p$  are connected with a dashed line on  $p'$ . But if  $p$  is open in  $\mathcal{G}_{\underline{A}}$ , then  $p'$  must be open in  $\mathcal{M}^+$ , since none of the variables (except for  $C$ ) on  $p$  is in  $\{\mathcal{Z}, R\}$  and thus none of the variables on  $p'$  is in  $\mathcal{Z}$ .

Thus we have shown that if a path  $p$  is open (after conditioning on  $\{\mathcal{Z}, R\}$ ) in  $\mathcal{G}_{\underline{A}}$  then  $p$  is an open generalised back-door path (after conditioning on  $\mathcal{Z}$ ) in  $\mathcal{M}^+$ , which proves the lemma.  $\square$

The proof of the main result follows automatically from these two lemmas.

This implies the following algorithm for determining whether or not a set of variables  $\mathcal{Z}$  is sufficient such that the conditional causal effect of  $A$  on  $Y$  given  $\mathcal{Z}$  can be identified from the complete records alone.

1. Draw a causal diagram ( $\mathcal{G}$ ) for the problem, making sure to include the exposure of interest ( $A$ ), the outcome of interest ( $Y$ ), and the missingness indicator ( $R$ ). Any common parent of two or more variables in  $\mathcal{G}$  must itself be included in  $\mathcal{G}$ , irrespective of whether or not data on this common cause have been collected.

2. Extend this causal diagram by adding any variable which is a parent of  $A$  or a parent of a descendant of  $A$  (except for parents of  $R$ ). This includes unobserved, and even unobservable, variables, *i.e.* those which we might think of as residual error or random error or ‘chance’. Call this extended causal diagram  $\mathcal{G}^+$ .
3. Take  $\mathcal{G}^+$  and join together—with a dashed line—any pair of variables that are both parents of  $R$ , or that share a child which is an ancestor of  $R$ . [Note that even if such a pair is already connected by an arrow (in either direction), a dashed line should be added as well.] The resulting diagram  $\mathcal{M}^+$  is known as the modified extended diagram.
4. Join together—with a dashed line—any pair of variables that are both parents of a variable in  $\mathcal{Z}$ , or that share a child which is an ancestor of a variable in  $\mathcal{Z}$ .
5. Look for a generalised path (where a generalised path can consist of arrows in any direction and dashed lines) from  $A$  to  $Y$ , not passing through  $R$ , that either (i) starts with an arrow *into*  $A$ , or (ii) contains a dashed line [Note: such a generalised path can start with an arrow *out* of  $A$ ]. We call such a generalised path a *generalised back-door path*. Does it contain a collider, and/or pass through a member of  $\mathcal{Z}$ ? If the answer to this question is ‘yes’ for every generalised back-door path, we say that **condition 1. is satisfied**. We also call this condition the *generalised back-door criterion*.
6. Return to the original causal diagram,  $\mathcal{G}$ . Remove all arrows *into*  $A$ .
7. Join together—with a dashed line—any pair of variables that are both parents of a variable in  $\mathcal{Z}$ , or that share a child which is an ancestor of a variable in  $\mathcal{Z}$ .
8. Look for a path from  $R$  to  $Y$ , not passing through  $A$ . Does it contain a collider, and/or pass through a member of  $\mathcal{Z}$ ? If the answer to this question is ‘yes’ for every such path, we say that **condition 2. is satisfied**.

Intuitively, the first condition ensures that any association seen between  $A$  and  $Y$  (possibly conditional on other variables) is causal. Suppose this

association is estimated from a generalised linear model, then the coefficient of  $A$  can be given a causal interpretation if the first condition holds. The second condition can intuitively be regarded as concerning the ‘intercept’. In order to identify  $pr(y|\check{a}, z)$  from the observed data, we must additionally be able to estimate the distribution of  $Y$  under the intervention  $\check{a} = 0$ , and for this to be identifiable from the complete records, the second condition is needed.

## 5 A note on our choice of Theorem 1 over Lemma 2 as a basis for our algorithm

It may seem counter-intuitive that we choose the first condition given in Theorem 1 as a basis for our algorithm (given in §3.2 of the main manuscript) rather than working directly with  $\mathcal{G}$  and the first condition given in Lemma 2 (conditions 2 are identical in both). Our reasons for doing so are two-fold. First, as highlighted in §2.2 (also in the main manuscript), the condition given in Theorem 1 is more closely related to the informal use of causal diagrams in missing data problems that we have encountered in practice, and thus stating which extra nodes are required for this approach to be valid, and what exactly constitutes a back-door path in this setting via our generalised back-door criterion, achieves the goal of formalising what is already informally done. Second, the generalised back-door criterion allows us to see why and how the estimator of the causal effect may be biased. For example, in Figure 2c in the main manuscript, we see that  $L$  acts as a confounder once we condition on  $R = 1$ . We do not believe that such an intuitive illustration of how biases arise can be obtained from using condition 1. of Lemma 2 instead.

## 6 Additional possible causal diagrams for the Airline Pilots and ATCOs study

Figures 6–8 show three alternative candidate diagrams to the one shown in Figure 1 of the main manuscript. Conditioning on  $R = 1$  results in the graphs shown in Figures 9–11.

First, we informally treat Figures 9–11 as if they were causal diagrams and apply the *back-door criterion* to see if the causal effect of exposure to cosmic radiation on skin cancer can be estimated without bias using only the complete records.

In Figure 9, there is a back-door path from exposure to outcome via age, suggesting that we should condition on age in the analysis.

In Figure 10, there is a back-door path from exposure to outcome via occupation and unmeasured behavioural factors, suggesting that we should condition on one or other (or both) of these in the analysis. Since the latter is unmeasured, this would suggest conditioning on the former. Finally, in Figure 11, the one back-door path cannot be blocked by measured variables.

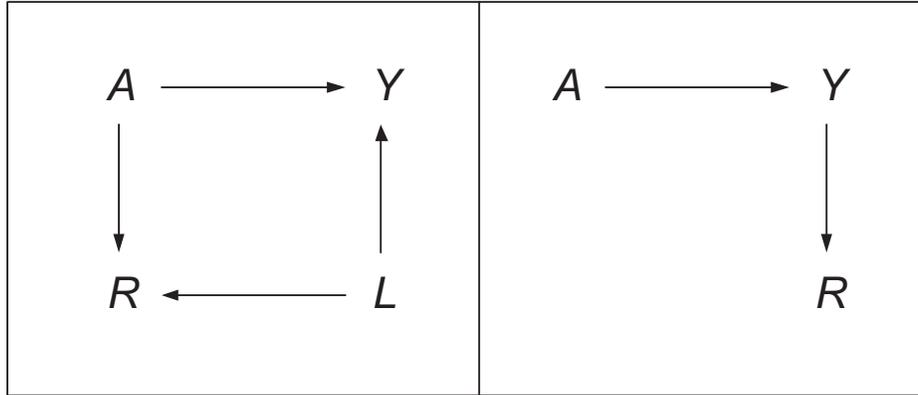
In the light of the result given above, we now return to Figures 6–8 and attempt a more formal interpretation of these diagrams. First, it is necessary to include two additional nodes to each diagram: one representing all causes of ‘Exposure to cosmic radiation’ and the other representing all other causes of ‘Skin cancer’ (not already in the diagram). Assuming that ‘Unmeasured genetic & behavioural factors’ includes all these for ‘Skin cancer’ in Figure 8, it transpires that these additional nodes have no bearing on any subsequent argument relating to these particular diagrams, and thus we have not shown them. Now, in addition to checking for unblocked backdoor paths, we have learnt that we must also check one additional condition: that there be no unblocked paths from ‘Skin cancer’ to  $R$  except through ‘Exposure to cosmic radiation’.

In Figure 6, assuming that we adjust for age, this additional condition is satisfied. However, in Figure 7, the path from  $R$  to ‘Skin cancer’ through ‘Unmeasured behavioural factors’ cannot be blocked and thus the conditions of Theorem 1 are not satisfied, even though there are no unblocked generalised back-door paths from the exposure to the outcome here. This is an example where condition 2. of Theorem 1 is required, and our original informal argument did not suffice.

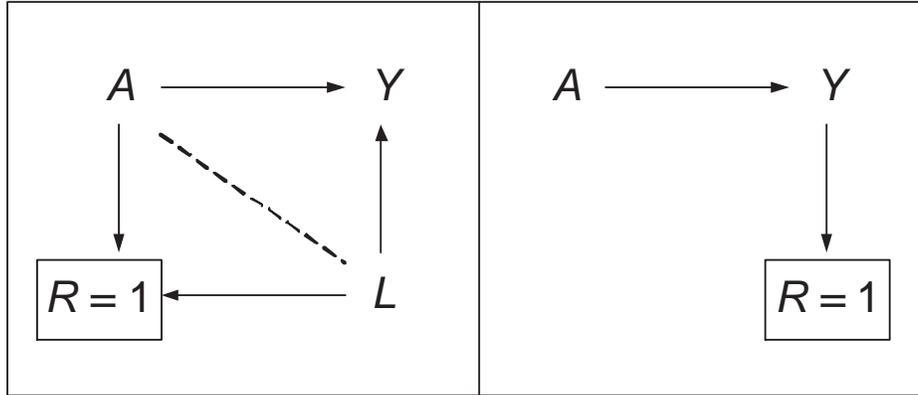
Turning to Figure 8, we had already noted that the first condition of Theorem 1 could not hold in this setting. As for the additional condition, this clearly does not hold either, by virtue of the direct arrow from ‘Skin cancer’ to  $R$ .

## References

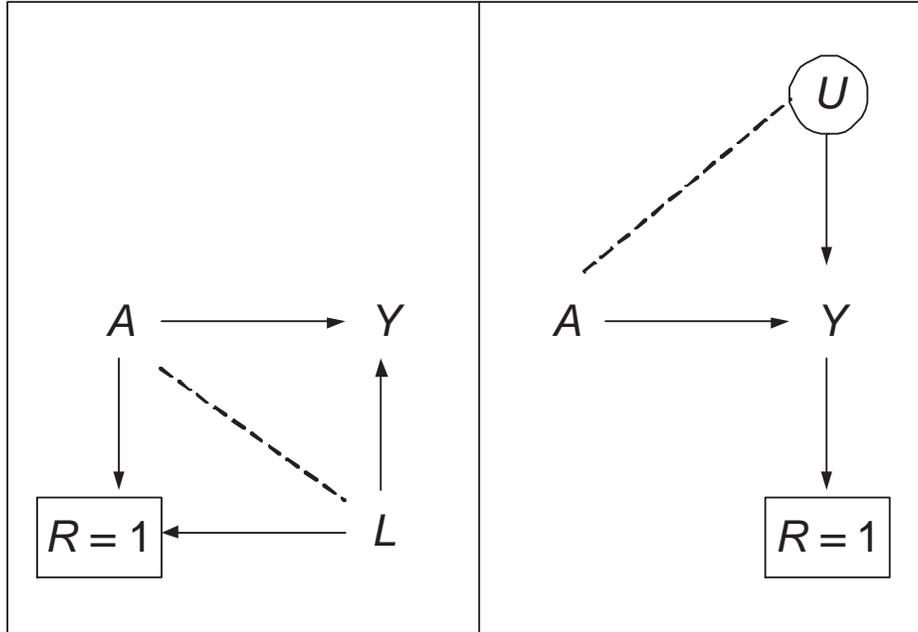
1. Pearl J. Causal diagrams for empirical research. *Biometrika* 1995; **82**:669–709.
2. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiological research. *Epidemiology* 1999; **10**:37–48.



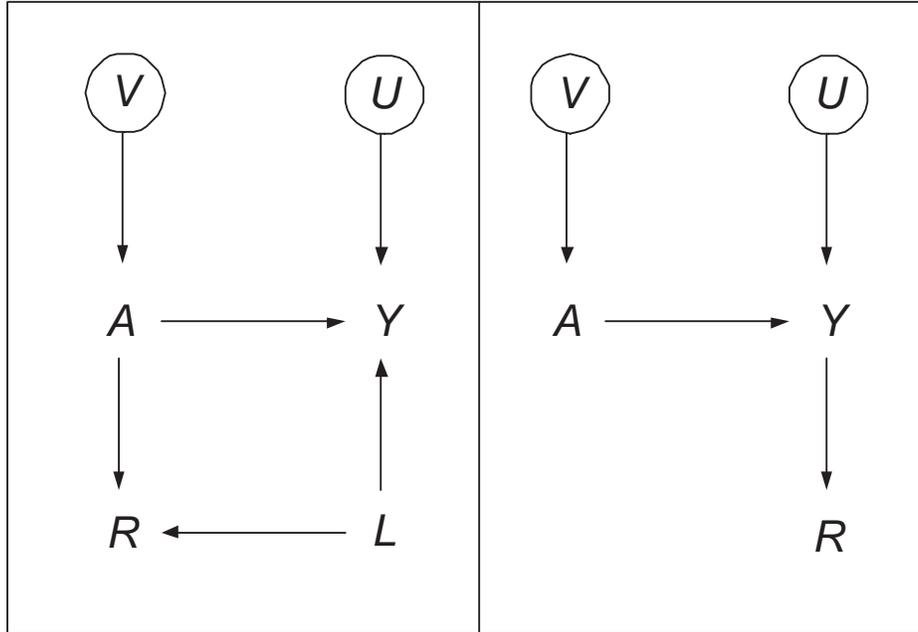
**Figure 1:** Two simple causal diagrams,  $\mathcal{G}_1$  (left) and  $\mathcal{G}_2$  (right).



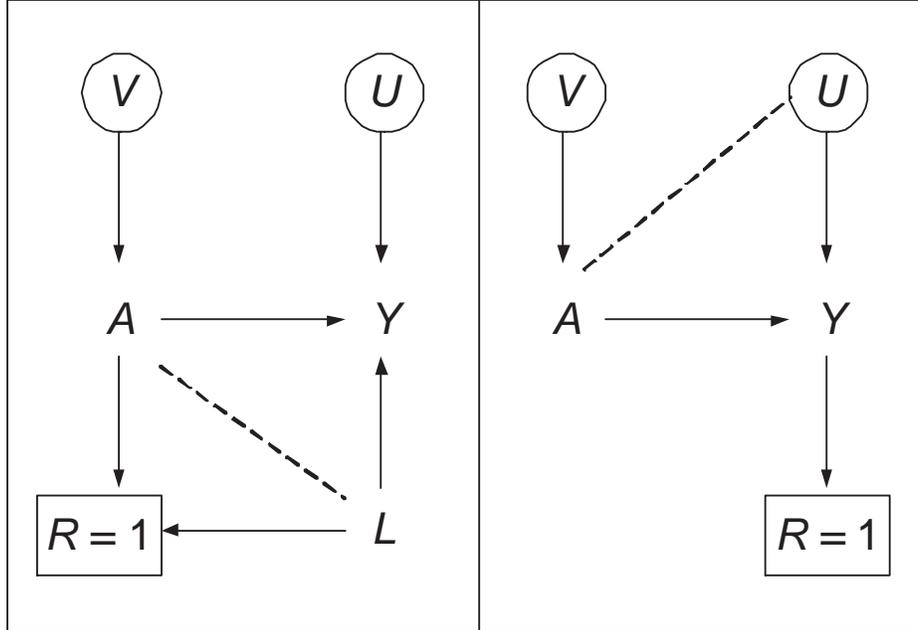
**Figure 2:** The same two simple diagrams, conditioning on  $R = 1$ .



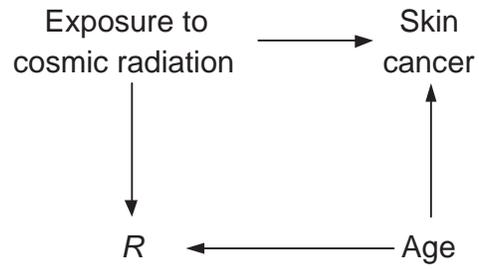
**Figure 3:** The same two simple diagrams, conditioning on  $R = 1$ , after extending  $\mathcal{G}_2$  to include the unmeasured  $U$ .



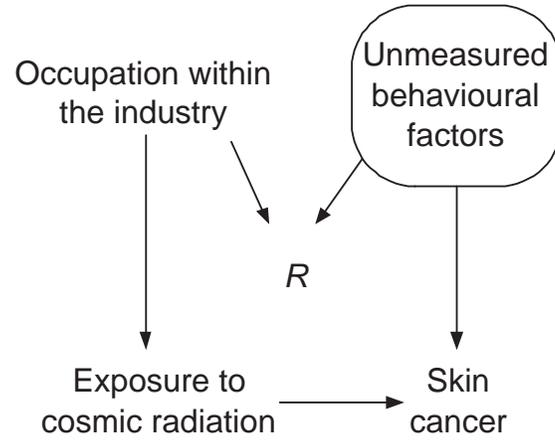
**Figure 4:** The extended causal diagrams,  $\mathcal{G}_1^+$  and  $\mathcal{G}_2^+$  corresponding to  $\mathcal{G}_1$  and  $\mathcal{G}_2$ .



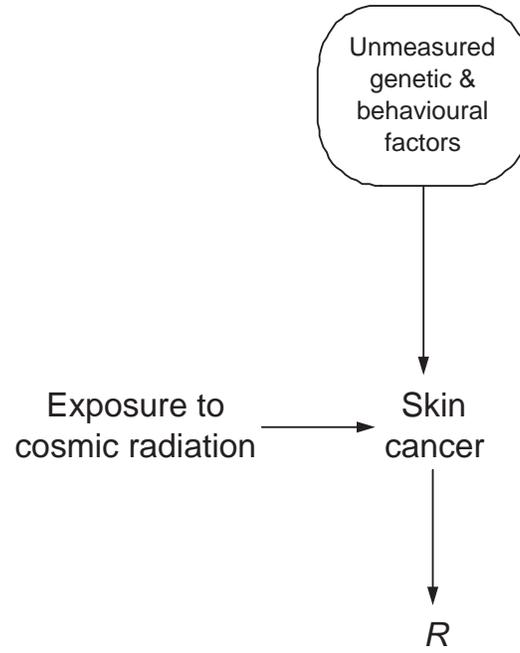
**Figure 5:** The modified extended diagrams,  $\mathcal{M}_1^+$  and  $\mathcal{M}_2^+$  corresponding to  $\mathcal{G}_1$  and  $\mathcal{G}_2$ .



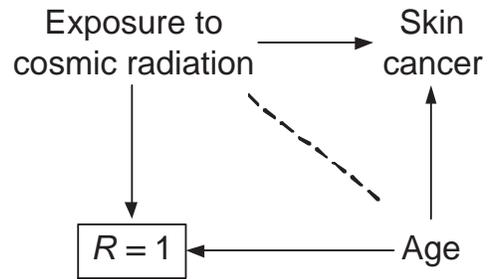
**Figure 6:** An alternative causal diagram for the airline pilots study.



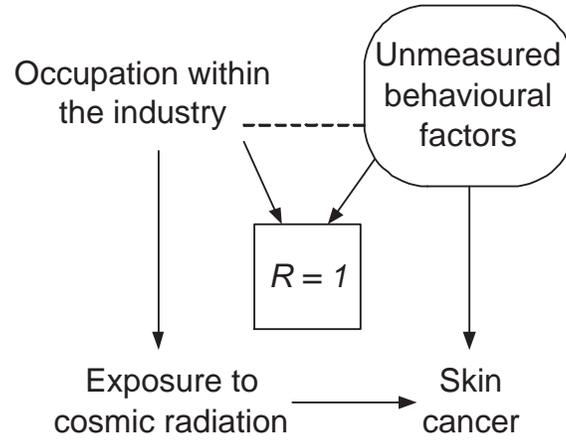
**Figure 7:** Another alternative causal diagram for the airline pilots study.



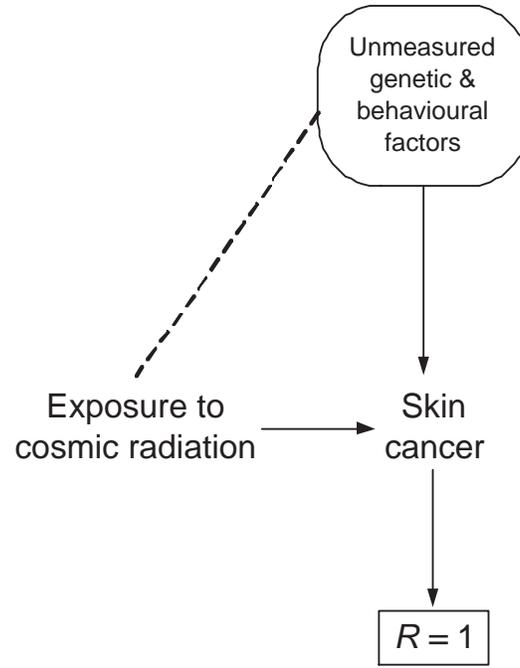
**Figure 8:** A fourth possible causal diagram for the airline pilots study.



**Figure 9:** A modified diagram, corresponding to Fig. 6, after conditioning on  $R = 1$ .



**Figure 10:** A modified diagram, corresponding to Fig. 7, after conditioning on  $R = 1$ .



**Figure 11:** A modified diagram, corresponding to Fig. 8, after conditioning on  $R = 1$ .