Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
21(3) 243–256

DOI: 10.1177/0962280210394469
smm.sagepub.com
SAGE

Using causal diagrams to guide analysis in missing data problems

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

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

causal diagram, causal inference, missing data

1 Introduction

A key aim of medical and epidemiological research is to establish causal links between treatments, 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[1,2] can represent these and the accompanying theory is useful in informing the design and analysis of studies. Pearl[1] and Greenland et al.[2] 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.

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

Corresponding author:
Rhian M Daniel, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Email: rhian.daniel@lshtm.ac.uk

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.[3] 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 $\mathcal{Z}$, and consider estimating the causal effect of A on Y conditional on $\mathcal{Z}$. In this case, the definition of a complete record becomes a subject for whom A, Y and $\mathcal{Z}$ 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.[4] 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 National Health Service (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 article, 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 provide no information) 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.

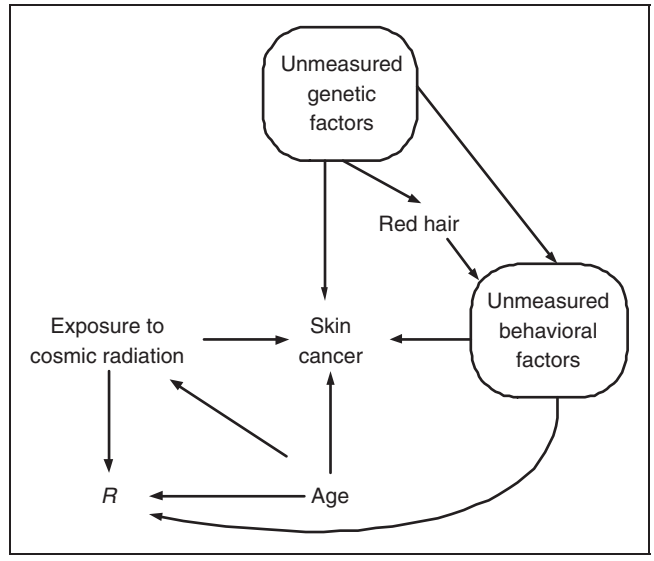

Figure 1. A possible causal diagram for the airline pilots study.

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 sporting 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.[2] 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.

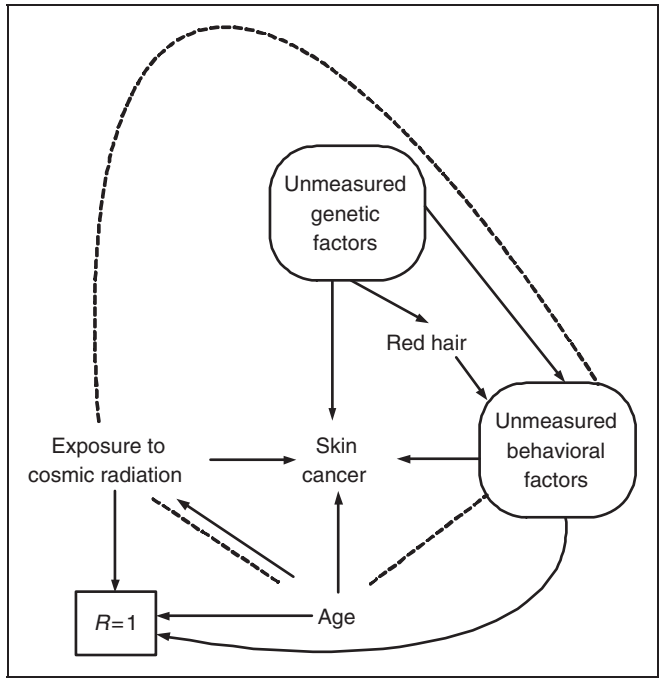

Figure 2. A modified diagram, derived from Figure 1 after conditioning on $R = 1$.

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.[2] 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* $W_1W_2 \ldots W_m$ from W_1 to W_m is a sequence of nodes such that, for each $k = 1, \ldots, m-1$, 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 $W_1W_2 \ldots W_m$ 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 a child of 'Exposure to cosmic radiation' and a descendant of 'Red hair' and 'Unmeasured genetic factors' an ancestor of R.

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

$$W_i = f_i(pa(W_i), \varepsilon_i) \quad i = 1, \ldots, n \tag{1}$$

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

Definition 5: (do operator). $\breve{w}_j$ denotes the act of *intervening* on W_j and setting its value to w_j. $W_j = \breve{w}_j$ 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 $pr(w_k|\breve{w}_l)$, is a function from W_l to the space of probability distributions on W_k. For each w_l, $pr(w_k|\breve{w}_l)$ 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 $pr(w_k|\breve{w}_l)$, such as, for binary W_l, the causal mean difference

$$E(W_k|\breve{w}_l = 1) - E(W_k|\breve{w}_l = 0) = \sum_{w':pr(w_k=w')>0} w' pr(w_k = w'|\breve{w}_l = 1) - \sum_{w':pr(w_k=w')>0} w' pr(w_k = w'|\breve{w}_l = 0),$$

or, for binary W_l and W_k, the causal odds ratio

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

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

The probability $pr(w_k|\breve{w}_l)$ 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), $pr(Y = 1|\breve{a}) = pr(Y = 1)$. 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 $pr(w_k|w_l) \neq pr(w_k|\breve{w}_l)$ 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 $pr(w_k|\breve{w}_l, w_m)$, 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 w_l.

3.2 Algorithm for determining whether or not the causal effect of A on Y given $\mathcal{Z}$ 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

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).[5] 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.

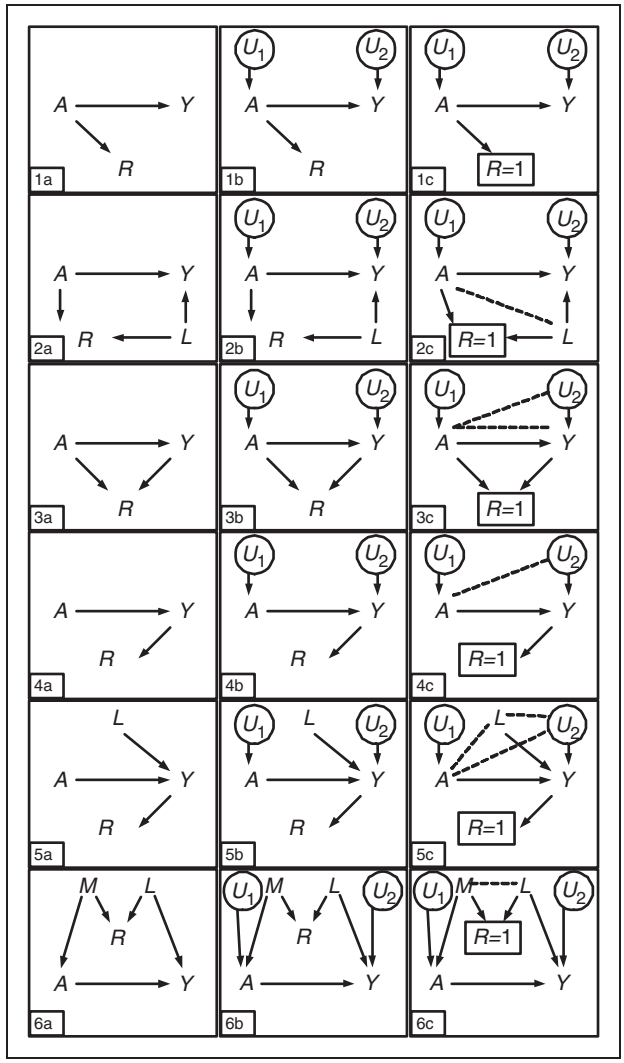

Figure 3. The first column shows the causal diagrams associated with various causal missingness mechanisms. In the second column, 1b–6b are extended causal diagrams corresponding to each of 1a–6a In the third column, 1c–6c are modified diagrams corresponding to each of 1b–6b, showing the effect of conditioning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

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 (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[6] and confirmed in simulation studies by Giorgi et al.[7] 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 of both 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

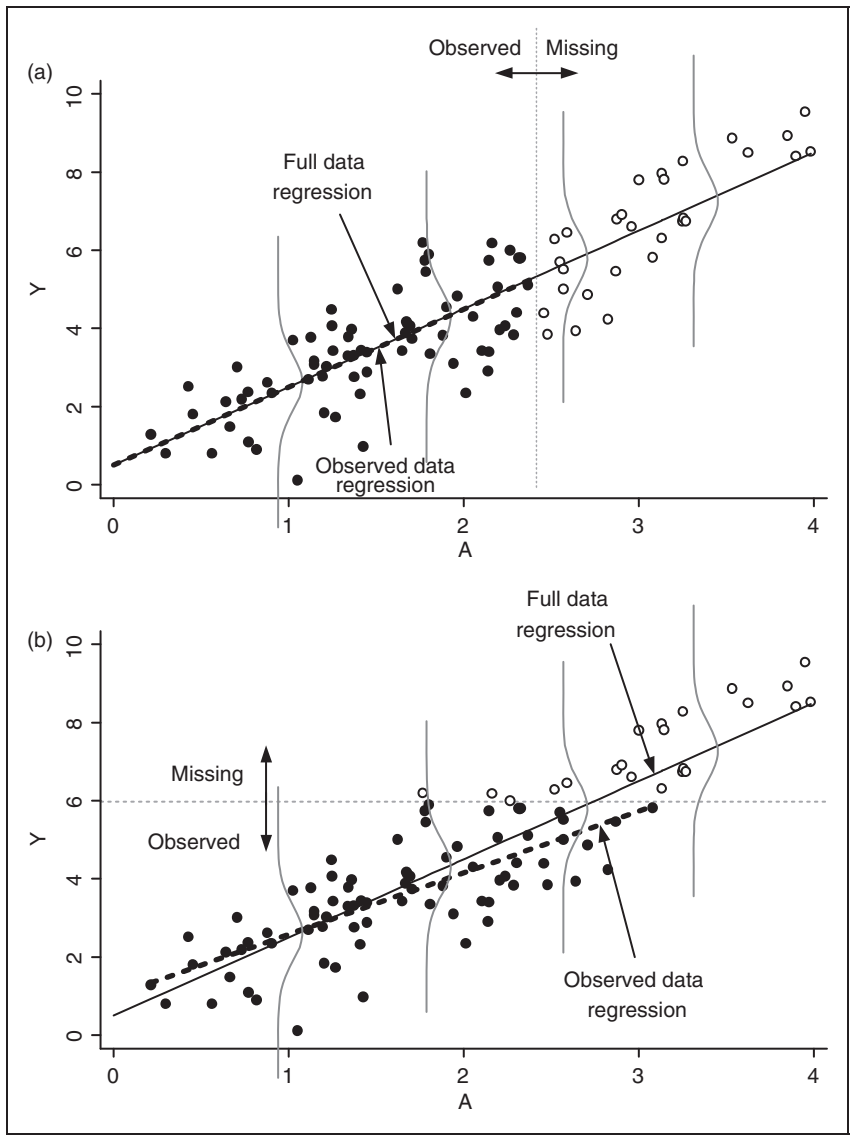

Figure 4. a comparison of (i) the full data estimate and (ii) the complete records estimate of the causal effect of *A* on *Y* under two different missingness mechanisms: missingness depends on *A* in panel a. and missingness depends on *Y* in panel b.

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 (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,[8] part V).

4.5 Missingness driven only by outcome

When missingness depends only on Y (Figure 3(4a)), conditioning on $R=1$ (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 included in a complete records analysis under this mechanism, whereas extremely low Y-values for a given A are included. 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 incomplete data, such as direct likelihood,[9] multiple imputation,[10,11] or inverse probability weighting[12] might be considered. However, if the outcome is incomplete (and hence the mechanism is MNAR), 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.[13] 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 ($pr(y|\breve{a})$) 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 $pr(y|\breve{a}) = pr(y)$ cannot be estimated without bias from the complete records. This agrees with the intuition given in §3.2.

Figure 3(5a) shows 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 accordance 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) shows 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.[2] Conditioning on $R = 1$ (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 affected (*cf*. age in Figure 1). When the data are incomplete, our algorithm encompasses the original back-door criterion found in Pearl[1] and Greenland et al.[2]

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 article, 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 and Robins JM. Causal diagrams for epidemiological research. *Epidemiology* 1999; **10**: 37–48.
3. Hernán MA, Hernández-Díaz S and 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 and dos Santos Silva I. Lifestyle of UK commercial aircrews relative to air traffic controllers and the general population. *Aviat Space Environ Med* 2008; **79**: 964–974.
5. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–592.
6. Rathouz PJ. Identifiability assumptions for missing covariate data in failure time regression models. *Biostatistics* 2007; **8**: 345–356.
7. Giorgi R, Belot A, Gaudart J and Launoy G. French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. *Stat Med* 2008; **27**: 6310–6331.
8. Molenberghs G and Kenward MG. *Missing data in clinical studies*. Chichester: Wiley, 2007.
9. Little RJA and Rubin DB. *Statistical analysis with missing data*. New York: Wiley, 2002.
10. Rubin DB. Multiple imputations in sample surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 1978, pp.20–34.
11. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; **91**: 473–489.
12. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; **90**: 106–121.
13. Cornfield J. A method of estimating comparable rates from clinical data: applications to cancer of the lung, breast and cervix. *J Nat Cancer Inst* 1951; **11**: 1269–1275.