American Journal of Epidemiology

Vol. 187, No. 12
DOI: 10.1093/aje/kwy173
Advance Access publication:
August 14, 2018

Practice of Epidemiology

Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies

Margarita Moreno-Betancur*, Katherine J. Lee, Finbarr P. Leacy, Ian R. White, Julie A. Simpson, and John B. Carlin

* Correspondence to Dr. Margarita Moreno-Betancur, Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, 50 Flemington Road, Parkville, VIC 3052, Australia (e-mail: margarita.moreno@mcri.edu.au).

Initially submitted January 2, 2018; accepted for publication August 3, 2018.

With incomplete data, the "missing at random" (MAR) assumption is widely understood to enable unbiased estimation with appropriate methods. While the need to assess the plausibility of MAR and to perform sensitivity analyses considering "missing not at random" (MNAR) scenarios has been emphasized, the practical difficulty of these tasks is rarely acknowledged. With multivariable missingness, what MAR means is difficult to grasp, and in many MNAR scenarios unbiased estimation is possible using methods commonly associated with MAR. Directed acyclic graphs (DAGs) have been proposed as an alternative framework for specifying practically accessible assumptions beyond the MAR-MNAR dichotomy. However, there is currently no general algorithm for deciding how to handle the missing data given a specific DAG. Here we construct "canonical" DAGs capturing typical missingness mechanisms in epidemiologic studies with incomplete data on exposure, outcome, and confounding factors. For each DAG, we determine whether common target parameters are "recoverable," meaning that they can be expressed as functions of the available data distribution and thus estimated consistently, or whether sensitivity analyses are necessary. We investigate the performance of available-case and multiple-imputation procedures. Using data from waves 1–3 of the Longitudinal Study of Australian Children (2004–2008), we illustrate how our findings can guide the treatment of missing data in point-exposure studies.

directed acyclic graphs; missing data; missing at random assumption; missing not at random assumption; multiple imputation; potential outcomes; recoverability; sensitivity analysis

Abbreviations: c-DAG, complete-data directed acyclic graph; DAG, directed acyclic graph; LSAC, Longitudinal Study of Australian Children; MAR, missing at random; m-DAG, missingness directed acyclic graph; MICE, multiple imputation by chained equations; MNAR, missing not at random; SDQ, Strengths and Difficulties Questionnaire.

Epidemiologic studies often suffer from missing data on multiple variables, including the exposure of interest, the outcome, and confounding factors. Methods such as multiple imputation (1, 2) allow unbiased estimation of all possible target parameters (e.g., mean values, regression-adjusted associations) if the "missing at random" (MAR) assumption holds. Investigators are thus urged to assess the plausibility of this assumption and encouraged to perform structured sensitivity analyses to examine the robustness of their conclusions to departures from MAR (3). Such analyses usually entail the elicitation of sensitivity parameters from subject-matter experts, requiring considerable effort.

However, there is a shortage of guidance on how investigators should assess the plausibility of MAR in practice. Seaman et al. (4) highlighted a lack of clarity about the definition of MAR in much of the missing-data literature, and they showed that a precisely defined condition called "everywhere MAR" (which is what we mean by MAR hereafter) is needed to guarantee valid frequentist inferences based on the likelihood—including methods such as multiple imputation. As Mealli and Rubin (5) noted, MAR is not an assumption about conditional independencies between variables, as is often thought, but about the nondependence of a function on one of its arguments. This is difficult to assess on the basis of substantive knowledge (6), which is crucial in the multivariable missingness setting (7–9). In particular, the stringency of MAR in general problems with multivariable missingness is poorly understood (4–6, 10, 11). Meanwhile, although MAR is sufficient for unbiased estimation, it is not necessary: In

problems with multivariable missingness, many "missing not at random" (MNAR) scenarios allow unbiased estimation of all or some parameters using standard implementations of multiple imputation or even a complete-case analysis. This has been noted in specific situations (7, 12), but there is little clarity as to how researchers can distinguish such MNAR settings from those in which sensitivity analyses are required.

Mohan and various colleagues (9, 13–15) proposed causal directed acyclic graphs (DAGs) as intuitive tools for depicting practically accessible and finer-grained missingness assumptions, beyond the MAR-MNAR dichotomy. This is a promising framework to guide the treatment of missing data. However, its practical applicability is currently impeded by the lack of a general algorithm with which to ascertain, given an arbitrary DAG, the nonparametric identifiability or "recoverability" (9, 13–15) of target parameters. That is, there is no general algorithm for determining from a DAG whether a given parameter can be estimated consistently using available data (and, if so, how) or whether sensitivity analyses are needed (16).

In this paper, we construct "canonical" DAGs capturing typical missingness mechanisms in the point-exposure study design with incomplete data on the exposure, outcome, and confounders. We derive recoverability results that can be used by epidemiologists to obtain or interpret their estimates given the canonical DAG(s) they consider plausible in their study. The article is organized as follows. First, we describe the canonical DAGs. Second, we determine the recoverability of parameters of major interest in each DAG. Third, we investigate the performance of available-case and multiple-imputation procedures. Finally, we use data from the Longitudinal Study of Australian Children (LSAC) to illustrate how our findings can be applied.

CANONICAL CAUSAL DIAGRAMS

We consider a general point-exposure study with incomplete exposure (X), incomplete outcome (Y), a set of complete confounders ($\boldsymbol{Z}_1$), and a set of incomplete confounders ($\boldsymbol{Z}_2$). All of these variables can be of any type (binary, continuous, etc.), and $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ can be univariate or multivariate.

Illustrative example

Our example examines the association between maternal mental illness and child behavior based on 4,882 children from the LSAC kindergarten cohort (17): children aged 4–5 years recruited in 2004 (wave 1 of LSAC; approved by the Australian Institute of Family Studies Ethics Committee), with five 2-yearly follow-up waves. The exposure variable (X) was a binary indicator of probable serious mental illness at wave 1 (yes/no; 15% missing), designated as affirmative if the mean value across the 6 items of the Kessler Psychological Distress Scale (18) was less than 4. The outcome variable (Y) was the child's score on the Strengths and Difficulties Questionnaire (SDQ) (range, 0–40; 23% missing) at wave 3. A higher score indicates increased behavioral difficulties.

Several potentially confounding covariates relating to the child, mother, and family were measured at wave 1. The completely observed covariates ($\boldsymbol{Z}_1$) were: sex of child; whether the child had siblings (yes/no); maternal completion of high school (yes/no); maternal age (years); consistent parenting score (range, 1–5); family financial hardship score (range, 0–6); and child's SDQ score at wave 1. The incomplete covariates ($\boldsymbol{Z}_2$) were: maternal current smoking status (yes/no; 16% missing); maternal risky alcohol drinking (>2 standard alcoholic drinks per day (yes/no); 18% missing); and child's physical functioning score (Pediatric Quality of Life Inventory (range, 0–100); 15% missing). Overall, 19% of the children had any covariate in $\boldsymbol{Z}_2$ missing, and 34% had any variable in (X, Y, $\boldsymbol{Z}_2$) missing.

Canonical complete-data DAG

A causal DAG depicts assumptions about causal relationships between variables using nodes connected by directed arrows (19, 20). The omission of variables or arrows encodes assumptions about the absence of relationships. With missing data, we first consider the DAG that would be assumed if the data were complete: the complete-data DAG (c-DAG).

Figure 1 shows the "canonical" c-DAG of a point-exposure study like the one in our LSAC example, where certain simplifications are used as a diagrammatic shorthand to encode, in a general way, the "no unmeasured or residual confounding" assumption usually underlying the primary analysis of such studies. To represent that the set of measured covariates is a sufficient set for confounding adjustment, a vector $\boldsymbol{U}$ representing all completely unmeasured common causes of the exposure and outcome is included. Further, the relationships between the measured covariates themselves are not depicted; rather, all covariates have been collected into 2 potentially vector-valued nodes representing the variables that are completely ($\boldsymbol{Z}_1$) or incompletely ($\boldsymbol{Z}_2$) observed.

Canonical missingness DAGs

Mohan et al. (9, 13–15) defined a "missingness graph," referred to here as a missingness DAG (m-DAG), as an extension of the c-DAG including the variable-specific missingness indicators to depict assumptions relating to missingness in each variable. This is more detailed than including the "complete case" indicator (21). The missingness indicator for X is defined as $M_X = 1$ if X is missing and $M_X = 0$ otherwise, and M_Y is defined similarly. We consider the incomplete confounders all together, with missingness indicator $M_{\boldsymbol{Z}_2} = 1$ if any of the components of $\boldsymbol{Z}_2$ is missing, and $M_{\boldsymbol{Z}_2} = 0$ otherwise. This simplification limits the number of possible DAGs while still capturing the essential detail for most practical purposes.

When constructing the m-DAG, the missingness indicators should be treated like any other variable, with all causal relationships depicted. To contain the complexity in constructing our canonical m-DAGs, we make 4 assumptions. The first 2 assumptions extend the "no unmeasured or residual confounding" assumption to the missingness setting:

- Assumption 1: There are no unmeasured common causes of a c-DAG variable and a missingness indicator. This precludes direct arrows from $\boldsymbol{U}$ to the missingness indicators and from $\boldsymbol{W}$, the vector of unmeasured common causes of the missingness indicators, to c-DAG variables.
- Assumption 2: There are no measured common causes of a c-DAG variable and a missingness indicator that are absent from the c-DAG. This precludes consideration of so-called "auxiliary

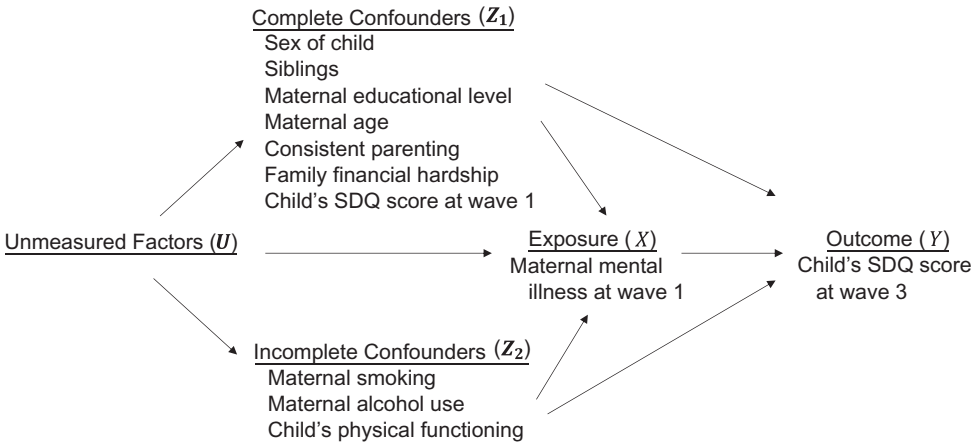

Figure 1. Canonical complete-data directed acyclic graph (c-DAG) for a general point-exposure study. For illustration, we provide under each node heading the variables involved in an example study of maternal mental illness and child behavior that used data from waves 1–3 of the Longitudinal Study of Australian Children (2004–2008). SDQ, Strengths and Difficulties Questionnaire.

variables" at the stage of making causal assumptions (see Discussion).

The next 2 assumptions are grounded in a truly causal interpretation of the arrows in a DAG, including the consideration that causality requires a cause to temporally precede an effect:

- Assumption 3: There are no direct arrows from missingness indicators to c-DAG variables. Such causal relationships would be plausible only in exceptional settings (e.g., when the data are used to determine a treatment).
- Assumption 4: There are no direct arrows between the missingness indicators. Such causal relationships would be rare in the point-exposure study, since, aside from the outcome, all variables and their missingness indicators are measured at the same time and so cannot cause one another. Similarly, the missingness of a variable at baseline would not truly cause missingness in the outcome a few years later.

Assumptions 1–3 imply that associations between a missingness indicator and a c-DAG variable can arise only in 2 ways: 1) a direct arrow from the substantive variable to the missingness indicator and 2) common causes of the substantive variable and the missingness indicator among c-DAG variables. By assumptions 1, 2, and 4, associations between missingness indicators can arise from either common causes among c-DAG variables or unmeasured common causes $\boldsymbol{W}$ distinct from $\boldsymbol{U}$. In the Discussion, we elaborate on the possibility of relaxing these assumptions.

Assumptions 1–4 limit the number of possible m-DAGs extending the c-DAG. Following a process summarized here and detailed in Web Appendix 1 (available at https://academic.oup.com/aje), we identified 10 m-DAGs that provide the most general forms of all essentially distinct extensions of m-DAG A in Figure 2 in terms of recoverability (see next section). This case forms the starting point, since it assumes the existence of arrows from completely observed confounders ($\boldsymbol{Z}_1$) to the missingness indicators (M_{Z_2}, M_X, M_Y), as is highly plausible in most epidemiologic studies. Briefly, we classified all extensions of m-DAG A into 16 categories according to whether there were arrows from 1) confounders and/or the exposure variable to missingness indicators of other variables, 2) confounders and/or the exposure variable to their own missingness indicators, 3) the outcome variable to missingness indicators of other variables, and 4) the outcome variable to its own missingness indicator. The m-DAG with the most arrows was selected as the "canonical" representative of each class, since it is the most general. Ten of the resulting 16 canonical m-DAGs, shown in Figure 2, were selected because they represent all distinct recoverability scenarios while having the most arrows.

RECOVERABILITY OF TARGET PARAMETERS

Definition of recoverability

Regardless of variable type (binary, continuous, etc.), 3 common target parameters in point-exposure studies are: 1) the expected value of the exposure (e.g., mean, proportion); 2) the expected value of the outcome; and 3) the exposure-outcome association adjusted for confounding through regression (e.g., regression-adjusted mean difference or odds ratio). Researchers would like to recover the estimates of these quantities that they would have obtained had there been no missing data. A parameter is recoverable if, based solely on causal assumptions, its value can be expressed as a function of the (large-sample) distribution of the available data (9, 13–15). In more formal statistical terms, recoverability is the same as nonparametric identifiability, meaning that it is possible to consistently estimate the parameter from the available data using an appropriate procedure. In Web Appendix 2, we provide a formal definition of recoverability by defining estimands in terms of potential outcomes.

One can similarly define the recoverability of target distributions. The 3 aforementioned target parameters are characteristics of the marginal exposure and outcome distributions and the conditional outcome distribution given X, $\boldsymbol{Z}_2$, and $\boldsymbol{Z}_1$, respectively. The recoverability of these distributions implies the recoverability

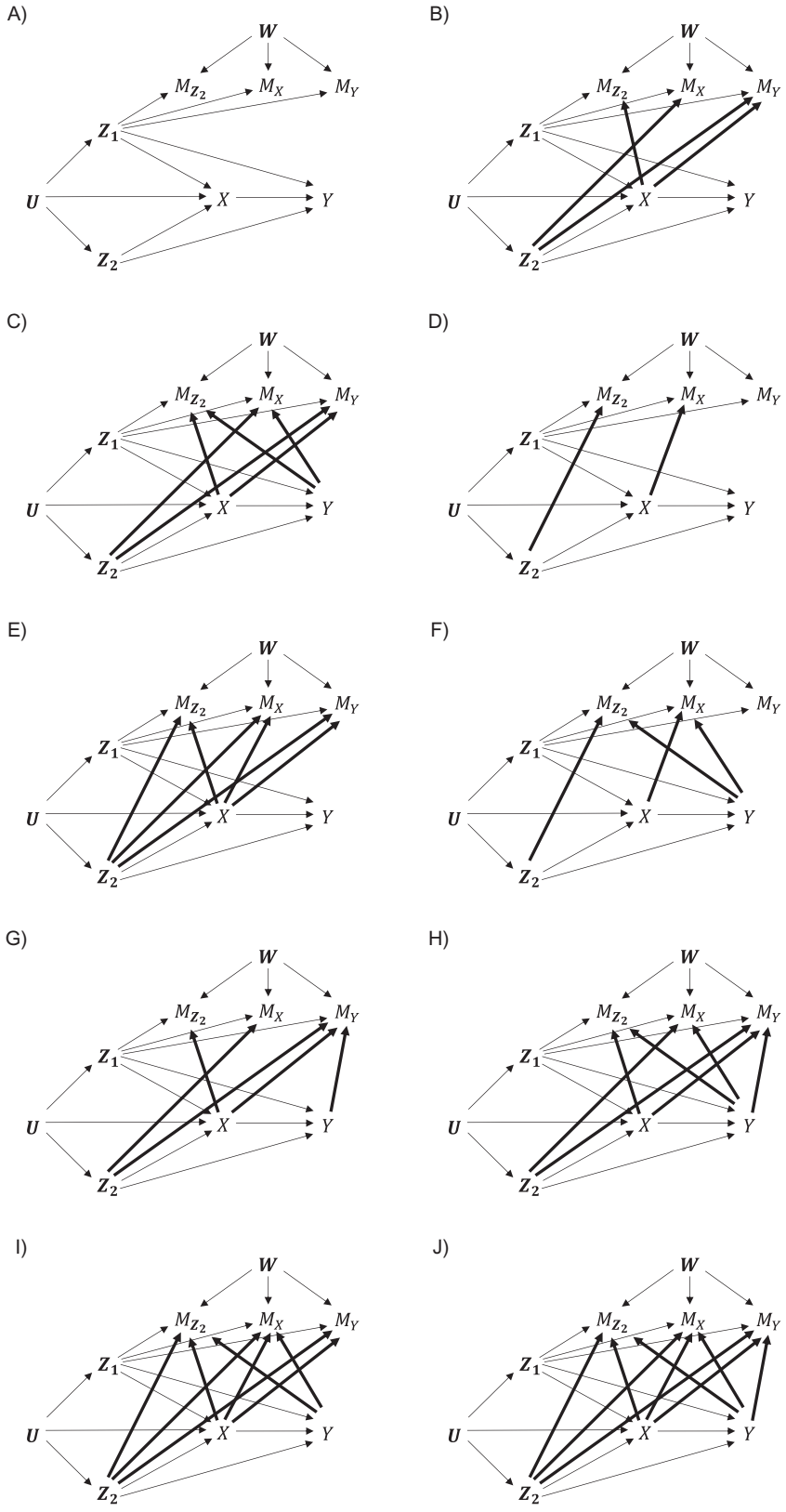

Figure 2. Canonical missingness directed acyclic graphs (m-DAGs) for a general point-exposure study. These 10 m-DAGs were identified as providing the most general forms of all essentially distinct extensions of the m-DAG shown in panel A (referred to as "m-DAG A") in terms of recoverability. To illustrate how each m-DAG extends m-DAG A, the additional arrows are indicated with a heavier line. In the text and tables, we refer to each m-DAG according to its figure locant (m-DAG A, m-DAG B, etc.).

of the corresponding parameters. If the joint distribution of Y, X, $\mathbf{Z}_2$, and $\mathbf{Z}_1$ is recoverable, then all target distributions and parameters are recoverable.

Conditions for recoverability

In Web Appendix 2, we describe general conditions that are required for recoverability. These broadly mirror those required for causal effect estimation (22), but the multivariable missingness situation is considerably more complex. In particular, while theoretical results establish sufficient or necessary graphical criteria for recoverability in particular cases, especially for the joint distribution (9, 13–15, 23), currently no algorithm can definitively ascertain the recoverability of a specific parameter given an arbitrary m-DAG (16). Thus, recoverability needs to be ascertained mathematically on a case-by-case basis.

Recoverability in the point-exposure study

Using results from Mohan et al. (13, 15) and new derivations, we established the recoverability of the joint distribution and the 3 aforementioned target distributions and corresponding parameters in the 10 canonical m-DAGs (Tables 1 and 2). In brief, when no c-DAG variable causes its own missingness, the joint distribution and thus all parameters are recoverable. Otherwise, some quantities are recoverable but others are not. The expectation of a variable that causes its own missingness is nonrecoverable, and neither is the regression-adjusted association if the outcome causes its own missingness. Notably, if a parameter is recoverable in a canonical m-DAG, it is also recoverable in all m-DAGs in the class(es) it represents, since these are obtained by removing arrows (lemma 4 in the paper by Mohan et al. (13)). However, the converse does not hold.

ESTIMATION WITH COMMON MISSING-DATA METHODS

Estimation of recoverable parameters

By definition, recoverable parameters can be consistently estimated on the basis of available data only using an appropriate method. Theoretically, "maximally efficient" estimation requires semiparametric methods that have been investigated in specific settings (24–27), but we do not consider those here, as they require tailoring for each m-DAG and recoverable parameter and have not commonly been used. Instead, we consider the performance, from a theoretical standpoint, of 2 common approaches that are readily implementable for any m-DAG and parameter: "available-case analysis" and multiple imputation.

Available-case analysis consists of estimating the target parameter using only records with complete data on the variables involved (e.g., those with $M_X = 0$ for estimating the expectation of X). For the regression-adjusted exposure-outcome association, this approach coincides with "complete-case analysis" and is unbiased in m-DAGs A, B, D, and E because the conditional distribution is expressible as the conditional distribution among the complete cases (Table 2). For other recoverable parameters, the available-case analysis could be subject to selection bias.

Multivariate normal imputation (2) and multiple imputation by chained equations (MICE) (1) are the most common multiple imputation approaches for handling multivariable missingness. While the former imposes a normal assumption on the full joint distribution, MICE approximates the joint distribution by iteratively imputing all incomplete variables using conditional models that include all other variables as predictors. These approaches can potentially overcome the selection bias that is expected in some situations with available-case analysis. However, they involve parametric assumptions beyond those of the analysis model (e.g., the outcome regression), which may entail a gain in precision relative to available-case analysis but could also induce misspecification bias. The unbiasedness of multiple imputation largely depends on the quality of the parametric assumptions with respect to the target parameter, which is related to the notion of "congeniality" between the imputation and analysis models (28).

The comparative performance of the two approaches is therefore difficult to establish in general because they are subject to different sources of bias, depending on the causal (m-DAG) and parametric assumptions. Instead, simulation studies are required to assess bias in specific contexts.

Estimation of nonrecoverable parameters

Nonrecoverability arises from inestimable systematic differences between the observed and missing data. Such differences can induce an insuperable selection bias in both the available-case approach and standard implementations of multiple imputation, but bias magnitudes depend on the context (see next section). Consistent estimation of nonrecoverable parameters requires introducing external data, typically in the form of expert-elicited values for the unknown differences. Given the inherent uncertainty in elicitation, estimation can be framed as a sensitivity analysis, producing a range of estimates for the target parameter corresponding to expert-elicited ranges of values for a set of sensitivity parameters. Typically, the analysis includes the setting in which all sensitivity parameters are 0.

An extended discussion on sensitivity analyses is beyond the scope of this paper, but we note that in setting up such analyses, identification of the parameters that require elicitation can be guided by the m-DAG and recoverability results, since these highlight the problematic arrows that lead to nonrecoverability. As a simple example, nonrecoverability of the exposure expectation in m-DAG D arises from the arrow $X \rightarrow M_X$, which flags exposure distribution differences between persons with observed exposure data and those with missing exposure data. These differences cannot be estimated from available data and thus would require elicitation. Methods for integrating elicited values in the estimation have been proposed, particularly within the multiple-imputation framework (29–34).

Findings from simulation study

We conducted a simulation study to investigate bias magnitudes in available-case and MICE analyses across the canonical m-DAGs in a setting like LSAC (Web Appendix 3). Our data-generating mechanism assumed main-effects models for the missingness indicators and set values of inestimable parameters, relating to the effect of a variable on its missingness indicator, to be of a similar magnitude to other associations observed in LSAC. We used MICE with an approximately congenial procedure.

Table 1. Recoverability Results for the Missingness Directed Acyclic Graphs (m-DAGs) in Figure 2, Stating Whether Each Distribution and Parameter Was Found to Be Recoverable in Each m-DAG[a]

Missingness DAG	Joint Distribution of Y, X, Z_1, Z_2	Marginal Distribution of X		Marginal Distribution of Y		Conditional Distribution of Y	
		Entire Distribution	Expectation (e.g., Proportion Exposed)	Entire Distribution	Expectation (e.g., Mean of Y)	Entire Distribution	Expectation (If Yes, Also Holds for the Regression Coefficient)
A	Yes[b]	Yes[b]	Yes	Yes[b]	Yes	Yes[b]	Yes
B	Yes[c]	Yes[d]	Yes	Yes[d]	Yes	Yes[b]	Yes
C	Yes[c]	Yes[d]	Yes	Yes[d]	Yes	Yes[d]	Yes
D	No[e]	No[e]	No	Yes[b]	Yes	Yes[b]	Yes
E	No[e]	No[e]	No	Unable to establish	Conjecture no unless $M_Y \perp M_{\boldsymbol{Z}_2}, M_X \mid \boldsymbol{Z}_1, \boldsymbol{Z}_2, X$[b]	Yes[b]	Yes
F	No[e]	No[e]	No	Yes[b]	Yes	Unable to establish	Conjecture no[b]
G	No[e]	Yes[d]	Yes	No[e]	No	No[f]	No
H	No[e]	Unable to establish	Conjecture no unless $M_X \perp M_{\boldsymbol{Z}_2}, M_Y \mid \boldsymbol{Z}_1, \boldsymbol{Z}_2, Y$[b]	No[e]	No	No[f]	No
I	No[e]	No[e]	No	Unable to establish	Conjecture no unless $M_Y \perp M_{\boldsymbol{Z}_2}, M_X \mid \boldsymbol{Z}_1, \boldsymbol{Z}_2, X$[b]	Unable to establish	Conjecture no[b]
J	No[e]	No[e]	No	No[e]	No	No[f]	No

Abbreviation: DAG, directed acyclic graph.

[a] Expressions in terms of available data provided in Table 2 in case of recoverability.

[b] Proof is provided in Web Appendix 2.

[c] Result obtained by corollary 1 in the paper by Mohan and Pearl (15).

[d] By recoverability of joint distribution (possibly of a reduced graph).

[e] By theorem 3 in the paper by Mohan and Pearl (15).

[f] By corollary 2 in the paper by Mohan and Pearl (15).

Table 2. Recoverability Results for the Missingness Directed Acyclic Graphs in Figure 2, Providing for Each Recoverable Distribution Its Mathematical Expression in Terms of Available Data[a,b,c]

Missingness DAG	Joint Distribution	Marginal Distribution of X	Marginal Distribution of Y	Conditional Distribution of Y
A	$P(Y, X, \boldsymbol{Z}_2 \mid \boldsymbol{Z}_1, \boldsymbol{M} = \boldsymbol{0}) \times P(\boldsymbol{Z}_1)$	$\sum_{\boldsymbol{Z}_1} P(X \mid \boldsymbol{Z}_1, M_X = 0) \times P(\boldsymbol{Z}_1)$	$\sum_{\boldsymbol{Z}_1} P(Y \mid \boldsymbol{Z}_1, M_Y = 0) \times P(\boldsymbol{Z}_1)$	$P(Y \mid X, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{M} = \boldsymbol{0})$
B	$\frac{P(Y, X, \boldsymbol{Z}_2, \boldsymbol{Z}_1, \boldsymbol{M} = \boldsymbol{0})}{P(M_Y = 0 \mid \boldsymbol{Z}_1, \boldsymbol{Z}_2, X, M_{\boldsymbol{Z}_2} = 0, M_X = 0) \times P(M_{\boldsymbol{Z}_2} = 0 \mid \boldsymbol{Z}_1, X, M_X = 0) \times P(M_X = 0 \mid \boldsymbol{Z}_1, \boldsymbol{Z}_2, M_{\boldsymbol{Z}_2} = 0)}$	No simple expression	No simple expression	$P(Y \mid X, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{M} = \boldsymbol{0})$
C	$\frac{P(Y, X, \boldsymbol{Z}_2, \boldsymbol{Z}_1, \boldsymbol{M} = \boldsymbol{0})}{P(M_Y = 0 \mid \boldsymbol{Z}_1, \boldsymbol{Z}_2, X, M_{\boldsymbol{Z}_2} = 0, M_X = 0) \times P(M_{\boldsymbol{Z}_2} = 0 \mid \boldsymbol{Z}_1, X, Y, M_X = 0, M_Y = 0) \times P(M_X = 0 \mid \boldsymbol{Z}_1, \boldsymbol{Z}_2, Y, M_{\boldsymbol{Z}_2} = 0, M_Y = 0)}$	No simple expression	No simple expression	No simple expression
D			$\sum_{\boldsymbol{Z}_1} P(Y \mid \boldsymbol{Z}_1, M_Y = 0) \times P(\boldsymbol{Z}_1)$	$P(Y \mid X, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{M} = \boldsymbol{0})$
E				$P(Y \mid X, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{M} = \boldsymbol{0})$
F			$\sum_{\boldsymbol{Z}_1} P(Y \mid \boldsymbol{Z}_1, M_Y = 0) \times P(\boldsymbol{Z}_1)$	
G		No simple expression		
H, I, J				

Abbreviation: DAG, directed acyclic graph.
[a] Proofs are provided in Web Appendix 2.
[b] A blank space is left where the distribution is not recoverable or it has not been established as documented in Table 1.
[c] $\boldsymbol{M} = (M_Y, M_X, M_{\boldsymbol{Z}_2})$, $\boldsymbol{0} = (0, 0, 0)$.

Web Figure 1 provides an overview of the results, which we summarize here. For recoverable parameters, the available-case approach exhibited greater bias for mean values than for association parameters and was approximately unbiased for the latter where expected (m-DAGs A, B, D, and E). Multiple imputation yielded approximately unbiased mean and association estimates, with the latter being more precise than those obtained with the available-case approach. These findings are consistent with the common observation in practice that both approaches yield similar association estimates. For nonrecoverable parameters, both approaches generally exhibited nonnegligible bias for the mean value but limited bias for the association parameter, except for m-DAG H when the underlying associations were very strong, agreeing with observations in the literature (35).

APPLICATION TO LSAC

In applying our findings to guide the treatment of missing data in the LSAC analysis, we first assess which m-DAG(s) are plausible using substantive knowledge (7–9). In LSAC, data on the exposure (X) and the incomplete confounders ($\boldsymbol{Z}_2$) were collected through questionnaires completed by parents and returned by mail (36). Incompleteness was due to the form not being returned (76% of missing cases for $\boldsymbol{Z}_2$ and 96% for X) or the question not being answered (remaining cases). Data on the outcome (SDQ score) were collected via a self-completed questionnaire administered during an in-person interview, and parents who failed to complete the questionnaire were asked to return the form by mail (36). Missing SDQ data were due to either attrition in LSAC (54%), failure to return the form (45%), or not answering the specific question (1%). Major reasons for attrition in LSAC are participants opting out or moving away and loss of contact (37).

It is plausible that some complete confounders are causes of missingness in all variables (e.g., low maternal education seems to be a cause of not returning forms and attrition (38, 39)), justifying the choice of m-DAG A as the base scenario. In Figure 3, we document the evidence regarding the possible presence of arrows from each of the incomplete variables to each of the missingness indicators.

Both m-DAG E and m-DAG J appear plausible. In both cases, the proportion exposed is nonrecoverable and sensitivity analyses would be required, and similarly for mean SDQ score, unless we adopt m-DAG E and the additional assumption that missingness in SDQ score is independent of missingness in other variables given the exposure and confounders (Table 1). We do not consider this plausible, since failure to return forms in waves 1 and 3 could have common causes (e.g., behavioral traits) that are not captured by the exposure and confounders. With m-DAG E, the regression-adjusted exposure-outcome association can be unbiasedly estimated using common methods, but sensitivity analyses would be required with m-DAG J.

These remarks shed light on the estimates obtained using available-case analysis and standard MICE with an approximately congenial procedure (Table 3). Both approaches yielded qualitatively similar results for all parameters, and we would expect both methods to be biased for the proportion with maternal mental illness and the mean SDQ score. If m-DAG E is adopted, we would expect both methods to be unbiased for the regression-adjusted association, but with m-DAG J, both methods could be biased.

		Arrow To:		
		M_{Z_2}	M_X	M_Y
Arrow From:	Z_2 (Maternal Alcohol Drinking and Smoking and Child's Physical Functioning at Wave 1)	Likely Failure to answer smoking and drinking questions can be due to stigma attached to these, and failure to return form can be due to maternal drinking and child physical functioning problems (42–44).	Uncertain Failure to return form can be due to maternal drinking and child physical functioning problems (42–44).	Uncertain Failure to return form and attrition by wave 3 can be due to maternal drinking and child physical functioning problems (42–44).
	X (Maternal Mental Illness at Wave 1)	Likely Failure to return form and non-response to specific questions can be due to mental health issues (43).	Likely Failure to return form can be due to mental health issues (43).	Likely Failure to return form and attrition by wave 3 can be due to mental health issues (39, 43).
	Y (Child's SDQ Score at Wave 3)	Not Likely Missingness in confounders preceded outcome by around 4 years.	Not Likely Missingness in exposure preceded outcome by around 4 years.	Uncertain Failure to return form and opt out from study at wave 3 can be due to increased current child difficulties (45).

Figure 3. Assessment of the existence of an arrow from each incomplete variable to each missingness indicator in the example from the Longitudinal Study of Australian Children (2004–2008), drawing from evidence in the literature (39, 42–45). SDQ, Strengths and Difficulties Questionnaire.

DISCUSSION

We investigated the use of m-DAGs in an epidemiologic setting; these were proposed by Mohan et al. (9, 13–15) as a new paradigm with which to frame clearer and finer-grained assumptions about missing data than the classical MAR-MNAR framework. Specifically, we constructed a series of "canonical" m-DAGs, providing results that can guide the analysis of point-exposure studies affected by missing data. The study of canonical structures facilitates the use of m-DAGs in practice, which is currently impeded by the complexity of determining the recoverability of parameters.

In addition to providing intuitive tools to depict detailed assumptions, the m-DAG paradigm reveals that it is crucial to draw a distinction between recoverable and nonrecoverable parameters and between missingness assumptions (m-DAG) and estimation procedures (with their potential parametric assumptions). For recoverable parameters, available-case and multiple-imputation procedures are subject to different sources of bias depending on the assumptions made, but both can be approximately unbiased in certain settings, as seen in our simulation study. Meanwhile, estimation of nonrecoverable parameters warrants sensitivity analysis using externally specified parameters. Uncertainty around the m-DAG, and thus recoverability, can have a bigger impact in terms of bias than the choice between estimation approaches. Thus, in settings with more than one plausible m-DAG, the likely recoverability of the target parameter (e.g., across most/almost none of them) should be considered in judging the reliability of estimates derived from different approaches.

Our recoverability results are useful for determining when sensitivity analyses are needed. This is important since these analyses are far from straightforward, requiring access to cooperative experts and elicitation and consensus methods that need to be tailored to each problem (40). Ultimately, the pertinence of undertaking a sensitivity analysis for a parameter hinges on recognizing it as nonrecoverable and assessing the potential magnitude of selection bias, in addition to its relevance for the study. The canonical m-DAGs can also guide sensitivity analyses when they are deemed necessary, which we plan to investigate further in future work.

Our construction of the canonical m-DAGs relied on the treatment of missingness in the confounders taken together, which led to simplified structures. In future work, more detailed DAGs could be considered. Further, the assumptions we made in constructing the canonical m-DAGs may not be appropriate in all contexts. Assumption 1 could be easily violated, similarly to the common occurrence of unmeasured confounding in observational studies. We expect more parameters to be nonrecoverable when there is an unmeasured common cause of a variable and its

Table 3. Estimates of 3 Target Parameters Using 2 Approaches to Handle Missing Data in the Example Study of Maternal Mental Illness and Child Behavior, Longitudinal Study of Australian Children (Waves 1–3), 2004–2008

Parameter	Estimate (SE)	95% CI	Is Estimate Reliable[a] if We Adopt:	
			m-DAG E?	m-DAG J?
Proportion of mentally ill mothers at wave 1				
Available-case analysis	0.21 (0.01)	0.20, 0.22	No	No
MICE	0.21 (0.01)	0.20, 0.23	No	No
Mean SDQ score[b] of children at wave 3				
Available-case analysis	7.48 (0.09)	7.31, 7.65	No	No
MICE	7.74 (0.09)	7.57, 7.90	No	No
Regression-adjusted difference in mean SDQ score[c]				
Available-case analysis	0.59 (0.20)	0.20, 0.98	Yes	No
MICE	0.64 (0.21)	0.23, 1.06	Yes	No

Abbreviations: CI, confidence interval; m-DAG, missingness directed acyclic graph; MICE, multiple imputation by chained equations; SDQ, Strengths and Difficulties Questionnaire; SE, standard error.

[a] This indicates whether the estimate can be considered reliable according to which m-DAG from Figure 2 is adopted, based on the recoverability of the parameter in that m-DAG.

[b] Range, 0–40. A higher score indicates increased behavioral difficulties.

[c] Comparing mentally ill mothers with non–mentally ill mothers.

missingness indicator, since this is similar, from a recoverability point of view, to the situation where there are direct arrows between these.

It would be possible to relax assumption 2, constructing m-DAGs that include auxiliary variables, which are often available in studies such as LSAC. This is a pragmatically driven limitation of our proposal. Given that auxiliary variables usually have missing data themselves, incorporating them into the m-DAGs not only would imply a considerable increase in the number of scenarios but also would require researchers to consider assumptions about missingness in these variables of secondary importance. We are investigating feasible avenues for incorporating auxiliary variables in the "assumptions" step, but in the meantime we suggest that these continue to be used in the "estimation" step in multiple-imputation procedures, as they are usually beneficial for precision (41).

Relaxing assumptions 3 and 4 would not only lead to an increase in the number of scenarios but also require further theoretical work. Assumption 3 underlies all the results of Mohan et al., and the setting in which assumption 4 is relaxed is treated separately by these authors and is substantially more complex mathematically (9, 13–15). Fortunately, these assumptions appear reasonable in the point-exposure design.

Mohan et al. (9, 13–15) proposed DAG-based definitions of "missing completely at random" and "MAR," which in our study correspond to the "trivial" m-DAG with no arrows from c-DAG variables to missingness indicators (see Web Appendix 1) and m-DAG A, respectively. The connections between the graph-based and classical (4, 5, 11) definitions have been explored (6, 9, 13). The graph-based "MAR" is stronger than (everywhere) MAR, although the two are equivalent under additional conditions: 1) independent records and 2) independence of missingness indicators given substantive variables (i.e., absence of $\boldsymbol{W}$ in our m-DAGs) (5, 10, 11). Then, MAR implies that missingness can depend only on fully observed variables, becoming more stringent as the number of incomplete variables grows. Understanding these connections is of interest, since the classical definitions underlie the theoretical results that underpin common missing-data methods. However, the full potential of the m-DAG approach is realized when focus is shifted from the MAR-MNAR classification and directed towards the specification of detailed mechanisms, a substantial number of which allow the possibility of unbiased estimation without needing to specify unidentifiable sensitivity parameters.

A limitation of DAGs is the impossibility of portraying interactions. Our m-DAGs thus encode structural assumptions about the missing-data mechanism, that is, main-effects models. These are reflected in our simulation study, which constitutes an initial investigation, under particular conditions, of the performance of common estimation approaches in conjunction with the proposed causal modeling framework. Further simulations are needed to assess biases in more general settings, considering missingness models with interactions, and also comparing performance with semiparametric estimators built, for example, using the recoverability formulae provided in Table 2.

In conclusion, our findings can be used to guide the treatment of missing data in point-exposure studies, and they provide avenues for future work on refining DAGs, estimation, and sensitivity analysis procedures.

ACKNOWLEDGMENTS

Author affiliations: Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Melbourne, Victoria, Australia (Margarita Moreno-Betancur, Katherine J. Lee, John B. Carlin); Centre for Epidemiology and

Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia (Margarita Moreno-Betancur, Julie A. Simpson, John B. Carlin); Department of Paediatrics, Melbourne Medical School, University of Melbourne, Melbourne, Victoria, Australia (Katherine J. Lee); Data Science Centre, Royal College of Surgeons in Ireland, Dublin, Ireland (Finbarr P. Leacy); and MRC Clinical Trials Unit, London, United Kingdom (Ian R. White).

This work was funded by 2 grants from the Australian National Health and Medical Research Council (NHMRC), a Project Grant (grant 1102468) and a Centre of Research Excellence grant awarded to the Victorian Centre for Biostatistics (grant 1035261). K.J.L. was supported by a fellowship from the NHMRC (grant 1120571). I.R.W. was supported by the Medical Research Council Unit Programme (grant MC_UU_12023/21). J.A.S. was supported by a Senior Research Fellowship from the NHMRC (grant 1104975). The Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

This paper was presented at the 38th Annual Conference of the International Society for Clinical Biostatistics, Vigo, Spain, July 9–13, 2017.

Conflict of interest: none declared.

REFERENCES

1. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1–67.
2. Schafer JL. *Analysis of Incomplete Multivariate Data*. London, United Kingdom: Chapman & Hall Ltd.; 1997.
3. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
4. Seaman S, Galati J, Jackson D, et al. What is meant by "missing at random"? *Stat Sci*. 2013;28(2):257–268.
5. Mealli F, Rubin DB. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*. 2015;102(4):995–1000.
6. Doretti M, Geneletti S, Stanghellini E. Missing data: a unified taxonomy guided by conditional independence. *Int Stat Rev*. 2018;86(2):189–204.
7. Bartlett JW, Harel O, Carpenter JR. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *Am J Epidemiol*. 2015;182(8):730–736.
8. Mohan K, Pearl J. On the testability of models with missing data. In: Kaski S, Corrander J, eds. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. (Proceedings of Machine Learning Research, vol. 33). Brookline, MA: JMLR, Inc. and Microtome Publishing; 2014:643–650.
9. Mohan K, Pearl J. Graphical models for processing missing data. *arXiv*. 2018. (doi: arXiv:1801.03583v1). Accessed April 1, 2018.
10. Mealli F, Rubin DB. 'Clarifying missing at random and related definitions, and implications when coupled with exchangeability' [erratum]. *Biometrika*. 2016;103(2):491.
11. Potthoff RF, Tudor GE, Pieper KS, et al. Can one assess whether missing data are missing at random in medical studies? *Stat Methods Med Res*. 2006;15(3):213–234.
12. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med*. 2010;29(28):2920–2931.
13. Mohan K, Pearl J, Tian J. Graphical models for inference with missing data. In: Burges C, Bottou L, Welling M, et al., eds. *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Red Hook, NY: Curran Associates, Inc.; 2013:1277–1285.
14. Thoemmes F, Mohan K. Graphical representation of missing data problems. *Struct Equ Modeling*. 2015;22(4):631–642.
15. Mohan K, Pearl J. Graphical models for recovering probabilistic and causal queries from missing data. In: Ghahramani Z, Welling M, Cortes C, et al., eds. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Red Hook, NY: Curran Associates, Inc.; 2014:1520–1528.
16. Shpitser I, Robins JM. Towards a complete identification algorithm for missing data problems. Presented at the "What-If?" workshop of the 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 5–10, 2016. http://www.homepages.ucl.ac.uk/~ucgtrbd/whatif/Paper10.pdf. Accessed August 31, 2017.
17. Sanson A, Nicholson J, Ungerer J, et al. *Introducing the Longitudinal Study of Australian Children*. (LSAC Discussion Paper no. 1). Melbourne, Victoria, Australia: Australian Institute of Family Studies; 2002.
18. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychol Med*. 2002;32(6):959–976.
19. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37–48.
20. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669–688.
21. Daniel RM, Kenward MG, Cousens SN, et al. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res*. 2012;21(3):243–256.
22. Hernan MA, Robins JM. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC. In press. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. Accessed July 10, 2017.
23. Shpitser I, Mohan K, Pearl J. Missing data as a causal and probabilistic problem. In: Meila M, Heskes T, eds. *Uncertainty in Artificial Intelligence: Proceedings of the Thirty-First Conference (2015)*. Corvallis, OR: AUAI Press; 2015:802–811.
24. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962–973.
25. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846–866.
26. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Stat Assoc*. 1999;94(448):1096–1120.
27. Díaz I, van der Laan MJ. Doubly robust inference for targeted minimum loss-based estimation in randomized trials with missing outcome data. *Stat Med*. 2017;36(24):3807–3819.
28. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci*. 1994;9(4):538–558.
29. Moreno-Betancur M, Chavance M. Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: methodology and application in a clinical trial with drop-outs. *Stat Methods Med Res*. 2016;25(4):1471–1489.
30. Moreno-Betancur M, Rey G, Latouche A. Direct likelihood inference and sensitivity analysis for competing risks regression with missing causes of failure. *Biometrics*. 2015;71(2):498–507.

31. Leacy FP, Floyd S, Yates TA, et al. Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *Am J Epidemiol*. 2017;185(4):304–315.
32. White IR, Carpenter J, Evans S, et al. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clin Trials*. 2007;4(2):125–139.
33. Leacy FP. *Multiple Imputation Under Missing Not at Random Assumptions via Fully Conditional Specification* [dissertation]. Cambridge, United Kingdom: University of Cambridge; 2016.
34. Tompsett DM, Leacy F, Moreno-Betancur M, et al. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Stat Med*. 2018;37(15):2338–2353.
35. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14(3):300–306.
36. Australian Institute of Family Studies. *Growing Up in Australia: The Longitudinal Study of Australian Children. An Australian Government Initiative. Data User Guide—November 2015*. Melbourne, Victoria, Australia: Australian Institute of Family Studies; 2015.
37. Norton A, Monahan K. *Growing Up in Australia: The Longitudinal Study of Australian Children (LSAC). Wave 6 Weighting and Non-Response*. (LSAC Technical Paper no. 15). Canberra, Australian Capital Territory, Australia: Australian Bureau of Statistics; 2015. http://www.growingupinaustralia.gov.au/pubs/technical/tp15.pdf. Accessed October 1, 2017.
38. Soloff C, Lawrence D, Misson S, et al. *Growing Up in Australia. The Longitudinal Study of Australian Children: An Australian Government Initiative. Wave 1 Weighting and Non-Response*. (LSAC Technical Paper no. 3). Canberra, Australian Capital Territory, Australia: Australian Bureau of Statistics; 2006. http://www.growingupinaustralia.gov.au/pubs/technical/tp3.pdf. Accessed October 1, 2017.
39. Sipthorp M, Misson S. *Growing Up in Australia. The Longitudinal Study of Australian Children: Wave 3 Weighting and Non-Response*. (LSAC Technical Paper no. 6). Canberra, Australian Capital Territory, Australia: Australian Bureau of Statistics; 2009. http://www.growingupinaustralia.gov.au/pubs/technical/tp6.pdf. Accessed October 1, 2017.
40. White I. The elicitation and use of expert opinion. In: Molenberghs G, Fitzmaurice G, Kenward M, et al., eds. *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman & Hall/CRC; 2014:471–490.
41. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330–351.
42. Gorman E, Leyland AH, McCartney G, et al. Assessing the representativeness of population-sampled health surveys through linkage to administrative data on alcohol-related outcomes. *Am J Epidemiol*. 2014;180(9):941–948.
43. Cheung KL, Ten Klooster PM, Smit C, et al. The impact of non-response bias due to sampling in public health studies: a comparison of voluntary versus mandatory recruitment in a Dutch national survey on adolescent health. *BMC Public Health*. 2017;17:276.
44. Lemmens PH, Tan ES, Knibbe RA. Bias due to non-response in a Dutch survey on alcohol consumption. *Br J Addict*. 1988; 83(9):1069–1077.
45. La Greca AM, Silverman WK. Parent reports of child behavior problems: bias in participation. *J Abnorm Child Psychol*. 1993; 21(1):89–101.