Marco Doretti, Sara Geneletti and Elena Stanghellini

Missing data: a unified taxonomy guided by conditional independence

Article (Accepted version) (Refereed)

Original citation:
Doretti, Marco and Geneletti, Sara and Stanghellini, Elena (2017) Missing data: a unified taxonomy guided by conditional independence. International Statistical Review. ISSN 0306-7734
DOI: 10.1111/insr.12242

Missing data: a unified taxonomy guided by conditional independence

Marco Doretti[1], Sara Geneletti[2], and Elena Stanghellini[3]

[1]Department of Political Science, University of Perugia, Perugia, 06123, Italy
Email: marco.doretti@unipg.it
[2]Department of Statistics, London School of Economics and Political Science,
London, WC2A 2AE, UK
Email: s.geneletti@lse.ac.uk
[3]Department of Economics, University of Perugia, Perugia, 06123, Italy
Email: elena.stanghellini@unipg.it

Abstract

Recent work (Seaman et al., 2013; Mealli & Rubin, 2015) attempts to clarify the not always well-understood difference between realised and everywhere definitions of missing at random (MAR) and missing completely at random (MCAR). Another branch of the literature (Mohan et al., 2013; Pearl & Mohan, 2013) exploits always-observed covariates to give variable-based definitions of MAR and MCAR. In this paper, we develop a unified taxonomy encompassing all approaches. In this taxonomy, the new concept of "complementary missing at random" (CMAR) is introduced, and its relationship with the concept of data observed at random (OAR) is discussed. All relationships amongst these definitions are analysed and represented graphically. Our paper covers both the univariate and the multivariate case, where attention is paid to monotone missingness and to the concept of sequential MAR (S-MAR). Specifically, for monotone missingness we propose an S-MAR definition which might be more appropriate than both everywhere and variable-based MAR to model dropout in certain contexts.

Key words: conditional independence; dropout; missing data; taxonomy.

1 Introduction

Almost every study involving human subjects suffers from some level of missingness, *i.e.* the data are not complete and there are individuals for whom some parts of the data are available but others are not. This has led to the development of a large number of statistical methods to adjust for missing data (Little & Rubin, 2002). These methods are often based on the assumptions of data *missing at random* (MAR) and *missing completely at random* (MCAR). However, three different groups of such definitions have been proposed over the years. The first two groups, also labelled *realised* and *everywhere* definitions in the literature (Seaman et al., 2013), are referred to as the *classical* definitions here. They were introduced to deal with the concept of *ignorability* of the missingness process, *i.e.* to state the conditions under which inference can be based on the observed data only. The third group of definitions, known as *variable-based* definitions, was introduced by a relatively novel branch of the literature to derive some results on recoverability of parameters (*i.e.* possibility of getting consistent estimates of such parameters from incomplete data) and testability of missingness processes (Mohan et al., 2013; Pearl & Mohan, 2013).

Differences and relationships amongst classical definitions were not fully understood for several years. The recent papers by Seaman et al. (2013) and Mealli & Rubin (2015) throw light on this topic. However, the distinction between classical and variable-based definitions has not yet been formally discussed. Clarifying this distinction is important as the aforementioned results about recoverability and testability rely on variable-based definitions and there is no guarantee that they hold when classical definitions are adopted. With this in mind, in this paper we build a unique taxonomy including classical and variable-based definitions as special cases.

For every group of definitions, the concept of *complementary missing at random* (CMAR) is introduced in a way such that the combination of MAR and CMAR is logically equivalent to MCAR. This scheme recalls the definition of data *observed at random* (OAR). Indeed, although the OAR definition has been rarely formalised in the literature, it is customary to affirm that data are MCAR when they are both MAR and OAR (Little & Rubin, 1987, p. 14; Heitjan, 1994; Heitjan & Basu, 1996). However, CMAR and OAR are generally different. Our approach relies on the formal language of conditional independence and makes explicit links between all definitions.

The paper is structured as follows. Section 2 sets up the necessary background including the literature review, notation and the basics of conditional independence. Section 3 defines the whole taxonomy for the univariate setting, with a particular focus on the special case of independence amongst sample units. Section 4 considers the multivariate setting, with a discussion on the case of monotone missingness and on the concept of sequential missing at random (S-MAR). In this section, we also define a novel missingness mechanism for the monotone missingness case, providing examples of applications where this might be more suitable than existing ones. In Section 5, some final remarks and conclusions are drawn.

2 Background

The missing data field was initiated by Rubin (1976), who first assigned also to missing data indicators the formal status of random variables. Thereafter, various definitions of MAR and MCAR have been proposed. Specifically, two groups of MAR/MCAR definitions were used for many years. Seaman et al. (2013) label these two groups *realised* and *everywhere* definitions respectively. The former are statements involving only *one* realisation of the missingness indicators' random vector, namely the sample realisation. Conversely, the latter are statements about *every* possible realisation of such random vector. In particular, Seaman et al. (2013) show that many authors have used everywhere MAR referring to realised MAR, which is the original definition proposed by Rubin (1976). Mealli & Rubin (2015) report the same definitions but adopt a different terminology. However, in this paper we adopt the terminology of Seaman et al. (2013) and denote realised and everywhere definitions by the prefixes R- and E- respectively.

These classical definitions are needed, together with additional conditions on parameters, to state sufficient conditions for *ignorability* of the missingness process across different inferential paradigms. Intuitively, if the missingness process is ignorable then valid inferential conclusions can be based only on a model for the observed data. Specifically, ignorability for Bayesian and direct likelihood inference requires the R-MAR definition, while for general frequentist inference the R-MCAR definition is needed. Further for frequentist likelihood inference ignorability can also be achieved by means of the E-MAR definition provided E-MCAR is assumed when the expected rather than the observed information matrix is used to calculate standard errors. For a detailed overview see Seaman et al. (2013).

Another recently developed branch of the literature considers a third group of definitions relying on the existence of always-observed auxiliary information (Pearl & Mohan, 2013; Mohan et al., 2013). These definitions are termed *variable-based* or *graph-based* as they enable the use of graphical tools like *missingness-graphs*, that are directed acyclic graphs including missingness indicators in their set of nodes (Mohan & Pearl, 2014a; Mohan et al., 2013). In this paper, we denote these definitions by adding the prefix VB-, which stands for "variable-based".

While realised and everywhere definitions fully address the concept of ignorability, variable-based definitions are concerned with recoverability and testability. Mohan et al. (2013) use missingness-graphs to explore recoverability of probabilistic queries. Mohan & Pearl (2014a) and Daniel et al. (2012) focus also on recoverability of causal relations. Potthoff et al. (2006) show that whether data are missing according to their variable-based definition of missing at random, which they term MAR+, can be tested when at least two variables have missing values. Finally, Mohan & Pearl (2014b) provide sufficient conditions for testability of different missingness models using missingness-graphs.

2.1 Notation

Although we use different symbols, we recover the notational scheme of Seaman et al. (2013) and extend it to include auxiliary information, which is essential to deal with variable-based definitions. Suppose that Y is the variable of interest and we plan to collect data on n sample units: then Y can be thought of as a $(n \times$

1) random column vector of *potentially observable* values, *i.e.* $Y = (\mathrm{Y}_1, \dots, \mathrm{Y}_n)'$ (notice that non-italic font is used to denote the components of Y so that Y_i is the random variable associated to the i-th sample unit). Realisations of random variables are always denoted by lower case letters. For instance, y and y^* represent two distinct realisations of Y while y_i represents a realisation of Y_i. We assign the symbol $\sim$ to *sample* realisations. Thus $\tilde{y} = (\tilde{\mathrm{y}}_1, \dots, \tilde{\mathrm{y}}_n)'$ is the vector the analyst would concretely observe if there were no missing data.

We denote by R an $(n \times 1)$ binary random vector with the i-th component taking value 1 if Y_i is observed and 0 otherwise. We also indicate by X an always-observed auxiliary covariate. Conventions for R and X are the same as those for Y. Furthermore, we introduce the function $o(Y, R)$ which returns the sub-vector of observed units, namely those Y_i such that $\mathrm{R}_i = 1$. Similarly, the function $m(Y, R)$ returns the sub-vector of missing units, that is those Y_i such that $\mathrm{R}_i = 0$. Thus, $(\tilde{x}, \tilde{r}, \tilde{y})$ are the *realised* sample data while $(\tilde{x}, \tilde{r}, o(\tilde{y}, \tilde{r}))$ are the *observed* sample data as Y is subject to missingness. In line with Seaman et al. (2013), we warn the reader not to confuse realised and observed values.

For the multivariate setting we use the bold symbol $\boldsymbol{Y}$ to denote the row vector $(Y_1, \dots, Y_T)$ or likewise the $(n \times T)$ matrix with elements Y_{it}, with $i = 1, \dots, n$ and $t = 1, \dots, T$. The notation $\bar{\boldsymbol{Y}}_t = (Y_1, \dots, Y_t)$ is used to represent partial collections up to the t-th variable. Again, $\boldsymbol{R}$ and $\boldsymbol{X}$ follow the same conventions, although $\boldsymbol{X}$ can have a general number of columns P. To be consistent with standard notation in this area, the rows of $\boldsymbol{Y}$ are partitioned into their observed $(\boldsymbol{Y}_i^o)$ and missing $(\boldsymbol{Y}_i^m)$ components, *i.e.* $\boldsymbol{Y}_i = (\mathrm{Y}_{i1}, \dots, \mathrm{Y}_{iT}) = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$. This notation is also adopted in the univariate case, where it will be

$$\begin{cases} \mathrm{Y}_i^o = \mathrm{Y}_i \text{ and } \mathrm{Y}_i^m = \emptyset & \text{if } \mathrm{R}_i = 1 \\ \mathrm{Y}_i^m = \mathrm{Y}_i \text{ and } \mathrm{Y}_i^o = \emptyset & \text{if } \mathrm{R}_i = 0. \end{cases}$$

It is important to bear in mind that $o(\boldsymbol{Y}, \boldsymbol{R})$ will be a sequence of column vectors, possibly of different lengths, while $o(\boldsymbol{Y}_i, \boldsymbol{R}_i)$ reduces to $\boldsymbol{Y}_i^o$.

Density functions will be denoted simply by f, with their lower case arguments clarifying without any ambiguity which variables they refer to as well as which variables they are conditioned on. Parameters of these densities are not explicitly reported and every statement is intended to hold for each value they can assume.

2.2 Conditional independence

Two random variables X and Y are independent conditional on another variable U when $f(x|u, y) = f(x|u)$ for every possible realisation x, y and u. In this case we write $X \perp\!\!\!\perp Y|U$ (Dawid, 1979). Another useful way of expressing this conditional independence (CI) is given by

$$f(x|u, y) = f(x|u, y^*) \quad \forall\, x, u, y, y^*.$$

The equation above says that the conditional density of X given U and Y is not a function of Y. Therefore, its equivalence with the original definition can be easily proved using the factorization criterion for conditional independence (Whittaker, 1990, p. 31; Dawid, 1979). When we are not conditioning on any variable U we talk about marginal independence of X and Y. By combining some properties of CI together (Dawid, 1979; Pearl, 2009, p. 11), it is possible to show that

$$X \perp\!\!\!\perp Y|U \quad \text{and} \quad X \perp\!\!\!\perp W|(Y,U) \iff X \perp\!\!\!\perp W,Y|U. \tag{2.1}$$

This property holds even marginally, *i.e.* when U is empty.

3 Univariate setting

3.1 General definitions of MAR/CMAR/MCAR

Armed with these preliminaries, we are able to provide a formal account of the aforementioned definitions. To this end, we define a generic 5-ple of realisations $z = (x, x^*, r, y, y^*)$ and denote by $\mathcal{Z}$ the set of all possible 5-ples. Then, we consider the following list of conditions that could be posed on $\mathcal{Z}$:

$$r = \tilde{r}; \tag{3.1}$$

$$x = x^*; \tag{3.2}$$

$$o(y,r) = o(y^*,r); \tag{3.3}$$

$$o(y,r) = o(y^*,r) = o(\tilde{y},r). \tag{3.4}$$

In words, Condition (3.1) forces to limit attention to the sample missingness pattern $\tilde{r}$ only. This condition will be used to state realised definitions. Condition (3.2) is satisfied only by those 5-ples such that x and x^* are equal. Finally, notice that Condition (3.3) is weaker than Condition (3.4) as it does require only that the observed parts of y and y^* be equal to each other, while Condition (3.4) also specifies that they must be equal to the observed part of the sample realisation $\tilde{y}$. Figure 1 represents the Venn diagram induced on $\mathcal{Z}$ by Conditions (3.1)-(3.4). Each condition holds in the part of $\mathcal{Z}$ where its respective number is depicted. For instance, (3.1) holds in the inner part of the small thick oval, while (3.2) holds on the left hand side of the double line. As not every 5-ple satisfying (3.3) will satisfy (3.4) as well, the subset of $\mathcal{Z}$ generated by (3.4) is contained into the one generated by (3.3).

Conditions (3.1)-(3.4) could be logically combined in many different ways. However, only a few among these combinations are of interest as they characterize the missing data definitions. More precisely, each definition requires the general equation

$$f(r|x,y) = f(r|x^*,y^*) \tag{3.5}$$

to be satisfied for a particular portion of $\mathcal{Z}$, and each portion represents a specific combination. The combinations associated to each definition are reported below.

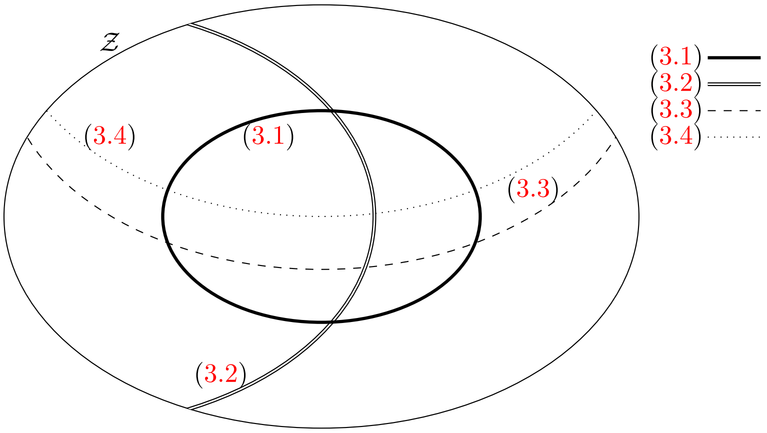


Figure 1: Venn diagram induced on $\mathcal{Z}$ by Conditions (3.1)-(3.4).

Specifically, (3.5) must hold $\forall\, z$:

R-MAR:	$r = \tilde{r} \,\wedge\, x = x^* \,\wedge\, o(y, r) = o(\tilde{y}, r) \,\wedge\, o(y^*, r) = o(\tilde{y}, r)$;
R-CMAR:	$r = \tilde{r} \,\wedge\, \big(x \neq x^* \,\vee\, o(y, r) \neq o(\tilde{y}, r) \,\vee\, o(y^*, r) \neq o(\tilde{y}, r)\big)$;
R-MCAR:	$r = \tilde{r}$;
E-MAR:	$x = x^* \,\wedge\, o(y, r) = o(y^*, r)$;
VB-MAR:	$x = x^*$;
E-CMAR:	$x \neq x^* \,\vee\, o(y, r) \neq o(y^*, r)$;
VB-CMAR:	$x \neq x^*$;
E-MCAR:	-

As mentioned above, realised definitions are, like the first definitions in Rubin (1976), always limited to the sample missingness pattern $\tilde{r}$. Conversely, E-MAR involves all possible patterns of R and therefore is stronger than R-MAR. Notice that a definition of VB-MCAR is not necessary; the reason will become clear in Subsection 3.2.1.

To illustrate what realised definitions mean in practice, we recall the numerical example of Seaman et al. (2013) and adapt it to our context. It considers a vector of four sample realisations $\tilde{y} = (10, 3, 4, 2)'$, the second of which is missing. Thus, we have $\tilde{r} = (1, 0, 1, 1)'$ and $o(\tilde{y}, \tilde{r}) = (10, 4, 2)'$. In this case R-MAR states that $f((1, 0, 1, 1)'|x, (10, a, 4, 2)') = f((1, 0, 1, 1)'|x, (10, b, 4, 2)')$ for every possible value of x, a and b. On the other hand, R-CMAR states that $f((1, 0, 1, 1)'|x, y) = f((1, 0, 1, 1)'|x^*, y^*)$ whenever x differs from x^* in at least one element or y and y^* are such that for at least one of them the first element is different from 10 or the third is different from 4 or the fourth is different from 2. To satisfy both R-MAR and R-CMAR, $f((1, 0, 1, 1)'|x, y) = f((1, 0, 1, 1)'|x^*, y^*)$ has to hold for every x, x^*, y and y^*. This corresponds to R-MCAR. Notice that the realised sample value $\tilde{x}$ is not explicitly involved. This representation with realised sample data does not shed light on other definitions as these involve generic realisations of X, R and Y.

In Figure 2, the region of $\mathcal{Z}$ for which (3.5) has to hold is highlighted in grey for each definition. From Figure 2 it is clear that, for every group of definitions, combining MAR and CMAR results in MCAR. The grey portion of a definition contains all the definitions that are *implied* by it: the larger the grey area, the stronger the assumption defining a missingness model. In particular, notice that E-CMAR does not imply R-CMAR, while R-MCAR does not imply E-MAR as remarked by Seaman et al. (2013). The implication diagram of Figure 3 provides a complete summary.

3.2 Alternative definitions

Definitions in Subsection 3.1 can be rewritten in several ways. Table 1 shows two possible alternatives, which are useful to highlight different noteworthy aspects. These alternative formulations are equivalent if, as we assume throughout the rest of the paper, sample units $(\mathrm{X}_i, \mathrm{R}_i, \mathrm{Y}_i)$ are independent, that is

$$f(r|x,y) = \prod_{i=1}^{n} f(\mathrm{r}_i|\mathrm{x}_i, \mathrm{y}_i).$$

3.2.1 Comments on definitions (a)

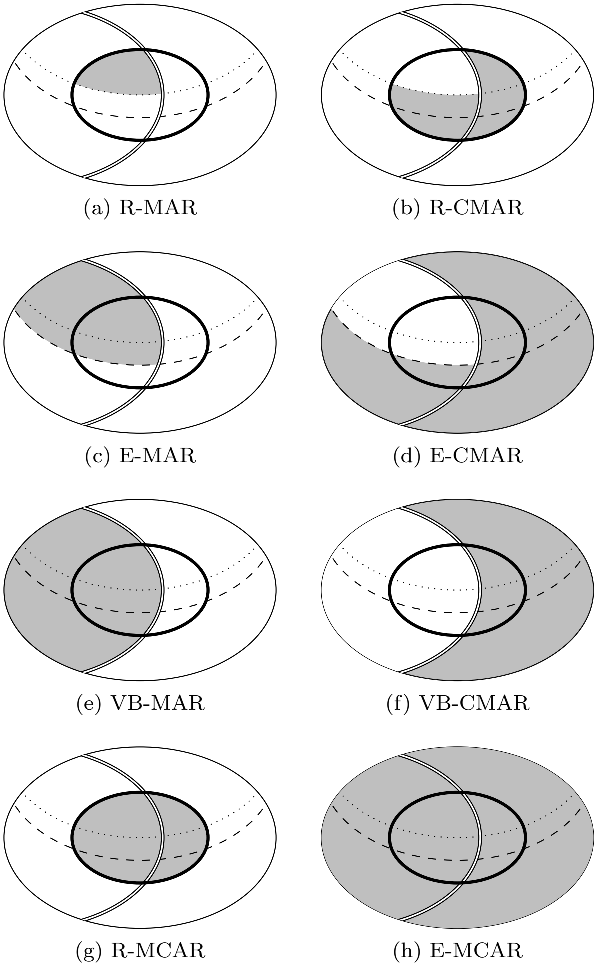
In Table 1a Conditions (3.1) and (3.2) are embedded, when possible, directly into Equation (3.5). Definitions for which (3.5) modifies in the same way are grouped together. For example, for both E-MAR and VB-MAR (3.5) becomes

$$f(r|x,y) = f(r|x,y^*),$$

which has to hold for all 4-ples (x, r, y, y^*) (VB-MAR), or only for those 4-ples such that $o(y,r) = o(y^*,r)$ (E-MAR). This formulation clarifies that, like E-MCAR, VB-MAR requires the related condition to hold for *every* possible realisation of *every* random variable involved in its expression. As a consequence, VB-MAR and E-MCAR can be encoded respectively by the two CI statements $R \perp\!\!\!\perp Y|X$ and $R \perp\!\!\!\perp (Y,X)$. Consequently, there is no need for a specific VB-MCAR definition as anticipated in Subsection 3.1. VB-MAR and E-MCAR both collapse into

$$f(r|y) = f(r|y^*) \;\forall\; r, y, y^*$$

(*i.e.* $R \perp\!\!\!\perp Y$) when no auxiliary information X is considered. The other classical definitions, including E-MAR, are statements “about functions being free of dependence on some of their arguments” (Mealli & Rubin, 2015) but not about conditional independence *on random variables*. Nevertheless, they could be interpreted as CI statements *on events*. For instance, R-MCAR is general with respect to the realisations of X and Y but is confined to the sample pattern $\tilde{r}$ so it can be associated to a CI statement on the event $\{R = \tilde{r}\}$ but not on the random variable R. Overall, it seems clear that questioning the plausibility of classical definitions in real data analysis is not straightforward. Conversely,



(a) R-MAR (b) R-CMAR

(c) E-MAR (d) E-CMAR

(e) VB-MAR (f) VB-CMAR

(g) R-MCAR (h) E-MCAR

Figure 2: Subsets of $\mathcal{Z}$ associated to every definition.

$$\begin{tikzcd}
 & \text{VB-CMAR} & \\
\text{E-CMAR} \arrow[ur, Rightarrow] \arrow[rr, Rightarrow, "\parallel" marking] & & \text{R-CMAR} \arrow[ul, Rightarrow, "/\!/" marking] \\
\text{E-MCAR} \arrow[u, Rightarrow] \arrow[rr, Rightarrow] \arrow[d, Rightarrow] & & \text{R-MCAR} \arrow[u, Rightarrow] \arrow[dl, Rightarrow, "\backslash\!\backslash" marking] \arrow[d, Rightarrow] \\
\text{VB-MAR} \arrow[r, Rightarrow] & \text{E-MAR} \arrow[r, Rightarrow] & \text{R-MAR}
\end{tikzcd}$$

Figure 3: The implication diagram: if definition a implies definition b, then there is a path of implication arrows going from a to b. The most relevant implications not holding are also shown.

variable-based definitions lend themselves to be more easily interpreted in applications: for example, analysts implicitly use them every time they wonder "is outcome actually independent of whether I observe it?".

Definitions (a) are also convenient as they show that "when all the data are observed, R-MAR cannot fail" (Seaman et al., 2013; Schafer & Graham, 2002). In this borderline situation $\tilde{r} = 1_n$, that is a $(n \times 1)$ vector of ones, and the $o(\cdot, \tilde{r})$ operator keeps y, y^* and $\tilde{y}$ unchanged. Therefore, R-MAR reduces to the trivially satisfied condition

$$f(1_n|x, y) = f(1_n|x, y^*)\, \forall\, x, y, y^* : \, y = \tilde{y} = y^*.$$

Finally, this formulation helps appreciate the difference between the concepts of CMAR and OAR. Ignoring auxiliary information X, Rubin (1976) indeed defined the R-OAR condition as

$$f(\tilde{r}|y) = f(\tilde{r}|y^*) \quad \forall\, y, y^* : \, m(y, \tilde{r}) = m(y^*, \tilde{r})$$

while Table 1a states that R-CMAR is (excluding X)

$$f(\tilde{r}|y) = f(\tilde{r}|y^*) \quad \forall\, y, y^* : \, o(y, \tilde{r}) \neq o(\tilde{y}, \tilde{r}) \,\vee\, o(y^*, \tilde{r}) \neq o(\tilde{y}, \tilde{r}).$$

To the best of our knowledge, formal definitions of OAR in the everywhere and in the variable-based framework have not been given, so a comparison with E-CMAR and VB-CMAR is not possible. However, we note that VB-CMAR does not correspond to the CI statement $R \perp\!\!\!\perp X$, which is what is needed, according to property (2.1), to obtain E-MCAR ($R \perp\!\!\!\perp (Y, X)$) from the weaker VB-MAR ($R \perp\!\!\!\perp Y|X$). We claim that a variable-based definition of OAR should correspond to such a CI statement.

3.2.2 Comments on definitions (b)

It is also useful to understand how these definitions relate to the generating process of each unit's missingness indicator R_i. In this spirit, we propose the unit-level definitions of Table 1b. Here some aspects concerning everywhere and variable-based definitions are noteworthy. To illustrate them, we recall from Subsection 2.1 that if Y_i is observed then $\mathrm{R}_i = 1$ and $\mathrm{Y}_i^o = \mathrm{Y}_i$, while when Y_i is missing we have $\mathrm{R}_i = 0$ and $\mathrm{Y}_i^o = \emptyset$. Thus, as R_i is a binary variable, looking at Table 1b it is clear that E-MAR can be rewritten as

$$\begin{cases} P(\mathrm{R}_i = 0|\mathrm{x}_i, \mathrm{y}_i) = P(\mathrm{R}_i = 0|\mathrm{x}_i) & \forall\, \mathrm{x}_i, \mathrm{y}_i \\ P(\mathrm{R}_i = 1|\mathrm{x}_i, \mathrm{y}_i) = P(\mathrm{R}_i = 1|\mathrm{x}_i, \mathrm{y}_i) & \forall\, \mathrm{x}_i, \mathrm{y}_i \end{cases} \tag{3.6}$$

or simply as

$$P(\mathrm{R}_i = 0|\mathrm{x}_i, \mathrm{y}_i) = P(\mathrm{R}_i = 0|\mathrm{x}_i) \qquad \forall\, \mathrm{x}_i, \mathrm{y}_i$$

because the second equality in (3.6) is a trivial identity. However, this also means that $P(\mathrm{R}_i = 1|\mathrm{x}_i, \mathrm{y}_i)$ does not depend on y_i as we have

$$P(\mathrm{R}_i = 1|\mathrm{x}_i, \mathrm{y}_i) = 1 - P(\mathrm{R}_i = 0|\mathrm{x}_i, \mathrm{y}_i) = 1 - P(\mathrm{R}_i = 0|\mathrm{x}_i) \qquad \forall\, \mathrm{x}_i, \mathrm{y}_i.$$

Therefore, we can conclude that in this case E-MAR is equivalent to

$$P(\mathrm{r}_i|\mathrm{x}_i, \mathrm{y}_i) = P(\mathrm{r}_i|\mathrm{x}_i) \quad \forall\, \mathrm{x}_i, \mathrm{r}_i, \mathrm{y}_i,$$

that is VB-MAR. The paper by Mealli & Rubin (2015) and its amendment show, with a different terminology, a related result for the multivariate setting stating that the everywhere and variable-based definitions of MAR are equivalent if sample units are independent and the components of the random vector $\boldsymbol{R}$ are mutually independent of each other conditionally on the random vector $\boldsymbol{Y}$. As the latter condition is trivial in the case of a singleton variable Y, our claim above is in line with their findings. Note that the requirement of independence across sample units is needed to achieve the equivalence of E-MAR and VB-MAR: Potthoff et al. (2006) state that E-MAR and VB-MAR are the same definition in the case of a singleton variable subject to missingness, but they do not explicitly mention the independence assumption.

4 Multivariate setting

We now turn our attention to the multivariate setting. In principle, definitions of Section 3 could apply to the $(nT \times 1)$ vectors obtained by stacking the columns of $\boldsymbol{Y}$ and $\boldsymbol{R}$ and to the $(nT \times P)$ matrix obtained by piling T identical copies of

$\boldsymbol{X}$. In any case, definitions of Section 3 remain valid for the multivariate case by simply replacing (Y, R, X) with $(\boldsymbol{Y}, \boldsymbol{R}, \boldsymbol{X})$.

Instead of reporting the entire list, we focus on those definitions which have a link with conditional independence. Though not explicitly mentioned, all definitions henceforth reported are intended to hold for every possible realisation of the random variables involved. Trivially, VB-MAR and E-MCAR translate into

$$(R_1, \ldots, R_T) \perp\!\!\!\perp (Y_1, \ldots, Y_T) | \boldsymbol{X}$$

and

$$(R_1, \ldots, R_T) \perp\!\!\!\perp (Y_1, \ldots, Y_T, \boldsymbol{X})$$

respectively, and they both reduce to

$$(R_1, \ldots, R_T) \perp\!\!\!\perp (Y_1, \ldots, Y_T) \tag{4.1}$$

in the absence of auxiliary information $\boldsymbol{X}$. These statements involve every component of the vectors $\boldsymbol{R}$ and $\boldsymbol{Y}$. However, other conditions are also useful: Pearl & Mohan (2013) state that the joint distribution $f(\boldsymbol{y})$ is recoverable in the presence of missing values if the data are sequentially missing at random (S-MAR), namely if an ordering $(Y_1, \ldots, Y_T)$ can be found that allows the existence of subsets $\boldsymbol{V}_{t-1} \subseteq \bar{\boldsymbol{Y}}_{t-1}$ such that

$$Y_t \perp\!\!\!\perp (R_t, \boldsymbol{R}_{v_{t-1}}) | \boldsymbol{V}_{t-1} \qquad t = 1, \ldots, T,$$

where $\boldsymbol{R}_{v_{t-1}}$ denotes the set of missingness indicators of variables in $\boldsymbol{V}_{t-1}$. Even in the most severe case, *i.e.* when $\boldsymbol{V}_{t-1}$ corresponds to $\bar{\boldsymbol{Y}}_{t-1}$ itself, S-MAR is weaker than VB-MAR as it assumes the form

$$Y_t \perp\!\!\!\perp \bar{\boldsymbol{R}}_t | \bar{\boldsymbol{Y}}_{t-1} \quad t = 1, \ldots, T. \tag{4.2}$$

Indeed, by iteratively applying property (2.1) it is easy to verify that (4.1) is equivalent to the combination of (4.2) and the sequence of conditional independences

$$R_t \perp\!\!\!\perp \bar{\boldsymbol{Y}}_{t-1} \quad t = 2, \ldots, T.$$

Notice that this definition of S-MAR is different from the one proposed by Hogan et al. (2004) and can obviously be extended in order to include auxiliary variables $\boldsymbol{X}$ in the conditioning set.

Unit-level definitions rely on the notation $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$. In particular, we have:

$$\text{E-MAR:} \quad f(\boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{y}_i) = f(\boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{y}_i^o); \tag{4.3}$$

$$\text{VB-MAR:} \quad f(\boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{y}_i) = f(\boldsymbol{r}_i | \boldsymbol{x}_i); \tag{4.4}$$

$$\text{E-MCAR:} \quad f(\boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{y}_i) = f(\boldsymbol{r}_i). \tag{4.5}$$

Condition (4.3) is often referred to simply as MAR (Little & Rubin, 2002, p. 18; Molenberghs et al., 2008). It means that for unit i the missing variables $\mathbf{Y}_i^m$ cannot probabilistically influence the missingness pattern $\boldsymbol{R}_i$ conditional on $\boldsymbol{X}_i$.

On the contrary, Condition (4.4) states that none of the variables which are missing for at least one unit (*i.e.* none of the variables in $\boldsymbol{Y}$) can influence $\boldsymbol{R}_i$ given auxiliary information (Mohan et al., 2013). (4.4) is MAR+ of Potthoff et al. (2006) and has also been named *covariate dependent* MCAR to distinguish it from (4.5) (Little, 1995; Hogan et al., 2004). When covariates are not involved, this distinction disappears and the term MCAR is used. It is worth remarking that in the multivariate setting we do not deal with the single binary indicator R_i anymore but we deal with the T-dimensional vector $\boldsymbol{R}_i$. Therefore, the same algebraic developments as in Subsection 3.2.2 are not possible and (4.3) cannot be further generalised as in principle $\boldsymbol{Y}_i^m$ (and consequently $\boldsymbol{R}_i$) differs for every unit. However, as we highlighted in Subsection 3.2.2, Mealli & Rubin (2015) provide conditions for the equivalence of (4.3) and (4.4) given additional assumptions.

4.1 Monotone missingness

In a multivariate cross-sectional analysis there is no temporal ordering of the components of $\boldsymbol{Y}$. In longitudinal studies missingness is termed *intermittent* when missing values can occur at any occasion and *monotone* when units can no longer be observed once they exited the study, *i.e.* $R_t = 0$ implies $R_s = 0$ for $s > t$. In the latter situation, also referred to as *dropout* here, it is customary to assume $R_1 = 1$, that is no units are missing at the baseline occasion. As a consequence, dropout allows us to work with a single random variable $D = \sum_{t=1}^{T} R_t$ taking values in $1, \ldots, T$ and indicating the time of last measurement (Little & Rubin, 2002, p. 17; Molenberghs et al., 1998). In this case definitions (4.3)-(4.5) change as follows:

$$\text{E-MAR:} \quad f(d|\boldsymbol{x}, \boldsymbol{y}) = f(d|\boldsymbol{x}, y_1, \ldots, y_d); \tag{4.6}$$

$$\text{VB-MAR:} \quad f(d|\boldsymbol{x}, \boldsymbol{y}) = f(d|\boldsymbol{x}); \tag{4.7}$$

$$\text{E-MCAR:} \quad f(d|\boldsymbol{x}, \boldsymbol{y}) = f(d); \tag{4.8}$$

for $d = 1, \ldots, T$. Again, it is straightforward to see that E-MCAR and VB-MAR can be encoded respectively by $D \perp\!\!\!\perp (\boldsymbol{Y}, \boldsymbol{X})$ and $D \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{X}$. Conversely, E-MAR (which is obviously implied by E-MCAR and VB-MAR) is *not* a complete conditional independence statement on the random variable D. However, it can be expressed as the collection of CI statements

$$Z_d \perp\!\!\!\perp (Y_{d+1}, \ldots, Y_T)|\bar{\boldsymbol{Y}}_d, \quad d = 1, \ldots, T,$$

with $Z_d = \mathbb{I}_{\{D=d\}}$ denoting the indicator variable of the event $\{D = d\}$. Notice that for $d = T$ both the CI statement above and E-MAR (4.6) become trivial.

In the context of monotone missingness, some authors showed that many missingness mechanisms usually expressed within the selection model framework have an equivalent pattern mixture model formulation (Molenberghs et al., 1998; Kenward et al., 2003). Specifically, Molenberghs et al. (1998) define MAR for selection models as our (4.6) (which, again, is E-MAR), proving its equivalence to the pattern mixture model condition

$$\forall\, t \geq 2, \forall\, j < t \quad f(y_t|\boldsymbol{x}, \bar{\boldsymbol{y}}_{t-1}, D = j) = f(y_t|\boldsymbol{x}, \bar{\boldsymbol{y}}_{t-1}, D \geq t). \tag{4.9}$$

They term (4.9) *available case missing value* (ACMV).

We now introduce a novel mechanism for the particular case of monotone missingness. This mechanism is encoded by the sequences of CI statements

$$D \perp\!\!\!\perp (Y_{t+1}, \ldots, Y_T)|(\boldsymbol{X}, \bar{\boldsymbol{Y}}_t) \qquad t = 1, \ldots, T-1 \tag{4.10}$$

and

$$D \perp\!\!\!\perp Y_t|(\boldsymbol{X}, \bar{\boldsymbol{Y}}_{t-1}) \qquad t = 2, \ldots, T, \tag{4.11}$$

which are equivalent. To prove this statement we consider, without loss of generality, the special case reported in Table 2 ($T = 4$). Looking at this table it is immediate to evince that (4.10) implies (4.11) as trivially (a) implies (b), (c) implies (d) while (e) and (f) are the same statement. Proving the converse requires property (2.1), which ensures that (c) holds if (d) and (f) do. Similarly, (a) is guaranteed by the combination of (f) and

$$D \perp\!\!\!\perp (Y_2, Y_3)|(\boldsymbol{X}, Y_1),$$

which in turn is implied by (b) and (d). Once again, (e) is trivially satisfied by (f) as they are identical. Notice that statement (a) of Table 2 implies (c) and (e) (and thus the entire sequence (4.10)) because of property (2.1).

An iterative application of property (2.1) also highlights that the statement

$$D \perp\!\!\!\perp Y_1|\boldsymbol{X} \tag{4.12}$$

is what (4.10) and (4.11) lack in order to be equivalent to VB-MAR (4.7). Therefore, this novel mechanism can be interpreted as the monotone missingness version of sequential MAR (4.2). However, it is also possible to show that (4.10) (and consequently (4.11)) are stronger than E-MAR. Therefore, such mechanism is "intermediate" with respect to the everywhere and the variable-based definition of MAR.

To demonstrate that (4.10) implies E-MAR (4.6), we can rely, without loss of generality, on the particular case $T = 4$ of Table 2. Here equation (4.6) becomes:

$$\begin{aligned}
P(D = 1|\boldsymbol{x}, \bar{\boldsymbol{y}}_4) &= P(D = 1|\boldsymbol{x}, y_1);\\
P(D = 2|\boldsymbol{x}, \bar{\boldsymbol{y}}_4) &= P(D = 2|\boldsymbol{x}, \bar{\boldsymbol{y}}_2);\\
P(D = 3|\boldsymbol{x}, \bar{\boldsymbol{y}}_4) &= P(D = 3|\boldsymbol{x}, \bar{\boldsymbol{y}}_3);\\
P(D = 4|\boldsymbol{x}, \bar{\boldsymbol{y}}_4) &= P(D = 4|\boldsymbol{x}, \bar{\boldsymbol{y}}_4).
\end{aligned}$$

The first three equalities are implied respectively by statements (a), (c) and (e) of Table 2 while the fourth is trivially verified as already mentioned above. The same example immediately shows that the converse is not true: many instances

$$\begin{tikzcd}
 & \text{E-MCAR (4.8)} \arrow[d, Rightarrow] & \\
 & \text{VB-MAR (4.7)} \arrow[dl, Rightarrow] \arrow[dr, Rightarrow] & \\
\text{S-MAR (4.10)} \arrow[rr, Leftrightarrow] \arrow[d, Rightarrow] & & \text{S-MAR (4.11)} \arrow[d, Rightarrow] \\
\text{E-MAR (4.6)} \arrow[rr, Leftrightarrow] & & \text{ACMV (4.9)}
\end{tikzcd}$$

Figure 4: The implication diagram for monotone missingness.

can be found which are required for (4.10) to hold but are not met by (4.6). For example, condition (a) of Table 2 also requires that

$$P(D = j|\boldsymbol{x}, \bar{\boldsymbol{y}}_4) = P(D = j|\boldsymbol{x}, y_1)$$

hold for $j = 2, 3, 4$. As a confirmation, it is also easy to show that (4.11) is stronger than ACMV (4.9). It is sufficient to notice that (4.11) can be written like

$$\forall\, t \geq 2,\ \forall\, j \quad f(y_t|\boldsymbol{x}, \bar{\boldsymbol{y}}_{t-1}, D = j) = f(y_t|\boldsymbol{x}, \bar{\boldsymbol{y}}_{t-1})$$

while Lemma 1 of Molenberghs et al. (1998) states that ACMV (4.9) is equivalent to the clearly weaker condition

$$\forall\, t \geq 2,\ \forall\, j < t \quad f(y_t|\boldsymbol{x}, \bar{\boldsymbol{y}}_{t-1}, D = j) = f(y_t|\boldsymbol{x}, \bar{\boldsymbol{y}}_{t-1}).$$

In the implication diagram in Figure 4, the relationships existing among all the mechanisms for monotone missingness considered in this section are represented.

From a practical standpoint, the model described by (4.10) or (4.11) can be helpful in a number of contexts, especially in the medical field. Specifically, if one thinks of $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$ as a vector of repeated measurements of a certain feature (for instance, body mass index of individuals), then there might be studies or clinical protocols where decision makers (*i.e.* doctors/nutritionists) at the first occasion schedule the total number of visits for each individual (and thus determine their value of D) based on some covariates $\boldsymbol{X}$ and on the value of the first measurement Y_1. In this case, the CI statement (4.12) does not hold by design so VB-MAR (4.7) is not suitable. However, also E-MAR is inappropriate since successive measurements are independent of the number of visits conditional on covariates and on the first measurement, that is

$$D \perp\!\!\!\perp (Y_2, \ldots, Y_T) | (\boldsymbol{X}, Y_1).$$

The conditional independence above indeed holds assuming (4.10) or (4.11) but not E-MAR (4.6). This example, like the one in Subsection 3.2.1, shows once again that in applications definitions involving CI statements on random variables are more tailored to interpretation than everywhere definitions.

5 Summary

Three types of definitions of missing at random (MAR) and missing completely at random (MCAR) can be found in the literature: realised, everywhere and variable-based definitions. Everywhere and realised definitions address ignorability of the missingness process, which is a concept that varies with the inferential framework adopted. Variable-based definitions, typically easier to interpret in applications, are often adopted when testability of missingness mechanisms or recoverability of probabilistic/causal queries in the presence of missing data are investigated. The differences existing among these groups of definitions do not seem to be entirely understood yet.

In this paper, we developed a unified taxonomy including all these definitions as special cases. As a by-product, the new concept of complementary missing at random (CMAR) was introduced. For every group of definitions, MCAR always corresponds to the combination of MAR and CMAR. We discussed the relationship between CMAR and observed at random (OAR). Links with the language of conditional independence were also explored. Moreover, the relationships and implications among all these definitions were discussed and represented in a number of graphical forms. As pointed out by other authors (Seaman et al., 2013), we remarked that everywhere MCAR is itself a conditional independence statement involving random variables, so a variable-based definition of MCAR is pointless. Our approach covers the univariate as well as the multivariate case, where a definition of sequential MAR (S-MAR) in line with Pearl & Mohan (2013) was provided. The special case of monotone missingness (dropout) was also discussed. Within this discussion, we presented a new mechanism which is "intermediate" between everywhere and variable-based MAR. We argue that such a mechanism can be interpreted as a monotone missingness specific version of sequential MAR. This novel mechanism demonstrates not only that everywhere and variable-based MAR differ, but also that, at least for monotone missingness, "something in between", potentially more suitable in certain applications, exists.

References

Daniel, R. M., M. G. Kenward, S. N. Cousens, & B. L. De Stavola (2012). Using causal diagrams to guide analysis in missing data problems. *Stat. Methods*

Med. Res. 21(3), 243–256.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol. 41*(1), 1–31.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika 81*(4), 701–708.

Heitjan, D. F. & S. Basu (1996). Distinguishing "missing at random" and "missing completely at random". *Amer. Statist. 50*(3), 207–213.

Hogan, J. W., J. Roy, & C. Korkontzelou (2004). Tutorial in Biostatistics. Handling drop-out in longitudinal studies. *Stat. Med. 23*(9), 1455–1497.

Kenward, M. G., G. Molenberghs, & H. Thijs (2003). Pattern-mixture models with proper time dependence. *Biometrika 90*(1), 53–71.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Amer. Statist. Assoc. 90*(431), 1112–1121.

Little, R. J. & D. B. Rubin (1987). *Statistical Analysis with Missing Data.* Wiley.

Little, R. J. & D. B. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). Wiley.

Mealli, F. & D. B. Rubin (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika 102*(4), 995–1000. With amendment (2016), *103*(2), 491.

Mohan, K. & J. Pearl (2014a). Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems*, pp. 1520–1528.

Mohan, K. & J. Pearl (2014b). On the testability of models with missing data. In *AISTATS*, pp. 643–650.

Mohan, K., J. Pearl, & J. Tian (2013). Graphical models for inference with missing data. In *Advances in neural information processing systems*, pp. 1277–1285.

Molenberghs, G., C. Beunckens, C. Sotto, & M. G. Kenward (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 70*(2), 371–388.

Molenberghs, G., B. Michiels, M. G. Kenward, & P. J. Diggle (1998). Monotone missing data and pattern-mixture models. *Stat. Neerl. 52*(2), 153–161.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2 ed.). New York, NY, USA: Cambridge University Press.

Pearl, J. & K. Mohan (2013). Recoverability and testability of missing data: Introduction and summary of results. Technical Report R-417, University of California, Los Angeles. Available at http://eprints.cdlib.org/uc/item/4c4996s0.

Potthoff, R. F., G. E. Tudor, K. S. Pieper, & V. Hasselblad (2006). Can one assess whether missing data are missing at random in medical studies? *Stat. Methods Med. Res. 15*(3), 213–234.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Schafer, J. L. & J. W. Graham (2002). Missing data: our view of the state of the art. *Psychol Methods 7*(2), 147.

Seaman, S., J. Galati, D. Jackson, & J. Carlin (2013). What is meant by "missing at random"? *Statist. Sci. 28*(2), 257–268.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

R-MAR	$f(\tilde{r}\|x,y) = f(\tilde{r}\|x,y^*) \quad \forall\, x,y,y^*:$	$o(y,\tilde{r}) = o(\tilde{y},\tilde{r}) \wedge o(y^*,\tilde{r}) = o(\tilde{y},\tilde{r})$
R-CMAR	$f(\tilde{r}\|x,y) = f(\tilde{r}\|x^*,y^*) \quad \forall\, x,x^*,y,y^*:$	$x \neq x^* \vee o(y,\tilde{r}) \neq o(\tilde{y},\tilde{r}) \vee o(y^*,\tilde{r}) \neq o(\tilde{y},\tilde{r})$
R-MCAR		-
E-MAR	$f(r\|x,y) = f(r\|x,y^*) \quad \forall\, x,r,y,y^*:$	$o(y,r) = o(y^*,r)$
VB-MAR		-
E-CMAR	$f(r\|x,y) = f(r\|x^*,y^*) \quad \forall\, z:$	$x \neq x^* \vee o(y,r) \neq o(y^*,r)$
VB-CMAR		$x \neq x^*$
E-MCAR		-

(a)

R-MAR	$f(\tilde{\mathrm{r}}_i\|\mathrm{x}_i,\mathrm{y}_i) = f(\tilde{\mathrm{r}}_i\|\mathrm{x}_i,\tilde{\mathrm{y}}_i) \quad \forall\, \mathrm{x}_i,\mathrm{y}_i : o(\mathrm{y}_i,\tilde{\mathrm{r}}_i) = o(\tilde{\mathrm{y}}_i,\tilde{\mathrm{r}}_i)$
R-CMAR	$f(\tilde{\mathrm{r}}_i\|\mathrm{x}_i,\mathrm{y}_i) = f(\tilde{\mathrm{r}}_i\|\mathrm{x}_i^*,\mathrm{y}_i^*) \quad \forall\, \mathrm{x}_i,\mathrm{x}_i^*,\mathrm{y}_i,\mathrm{y}_i^* : \mathrm{x}_i \neq \mathrm{x}_i^* \vee o(\mathrm{y}_i,\tilde{\mathrm{r}}_i) \neq o(\tilde{\mathrm{y}}_i,\tilde{\mathrm{r}}_i) \vee o(\mathrm{y}_i^*,\tilde{\mathrm{r}}_i) \neq o(\tilde{\mathrm{y}}_i,\tilde{\mathrm{r}}_i)$
R-MCAR	$f(\tilde{\mathrm{r}}_i\|\mathrm{x}_i,\mathrm{y}_i) = f(\tilde{\mathrm{r}}_i) \qquad \forall\, \mathrm{x}_i,\mathrm{y}_i$
E-MAR	$f(\mathrm{r}_i\|\mathrm{x}_i,\mathrm{y}_i) = f(\mathrm{r}_i\|\mathrm{x}_i,\mathrm{y}_i^o) \quad \forall\, \mathrm{x}_i,\mathrm{r}_i,\mathrm{y}_i$
VB-MAR	$f(\mathrm{r}_i\|\mathrm{x}_i,\mathrm{y}_i) = f(\mathrm{r}_i\|\mathrm{x}_i) \quad \forall\, \mathrm{x}_i,\mathrm{r}_i,\mathrm{y}_i$
E-CMAR	$f(\mathrm{r}_i\|\mathrm{x}_i,\mathrm{y}_i) = f(\mathrm{r}_i\|\mathrm{x}_i^*,\mathrm{y}_i^*) \quad \forall\, \mathrm{x}_i,\mathrm{x}_i^*,\mathrm{r}_i,\mathrm{y}_i,\mathrm{y}_i^* : \mathrm{x}_i \neq \mathrm{x}_i^* \vee o(\mathrm{y}_i,\mathrm{r}_i) \neq o(\mathrm{y}_i^*,\mathrm{r}_i)$
VB-CMAR	$f(\mathrm{r}_i\|\mathrm{x}_i,\mathrm{y}_i) = f(\mathrm{r}_i\|\mathrm{x}_i^*,\mathrm{y}_i^*) \quad \forall\, \mathrm{x}_i,\mathrm{x}_i^*,\mathrm{r}_i,\mathrm{y}_i,\mathrm{y}_i^* : \mathrm{x}_i \neq \mathrm{x}_i^*$
E-MCAR	$f(\mathrm{r}_i\|\mathrm{x}_i,\mathrm{y}_i) = f(\mathrm{r}_i) \quad \forall\, \mathrm{x}_i,\mathrm{r}_i,\mathrm{y}_i$

(b)

Table 1: Missing data definitions: two alternatives.

Condition (4.10)			Condition (4.11)
$D \perp\!\!\!\perp (Y_2, Y_3, Y_4)\vert(\boldsymbol{X}, Y_1)$	(a)	(b)	$D \perp\!\!\!\perp Y_2\vert(\boldsymbol{X}, Y_1)$
$D \perp\!\!\!\perp (Y_3, Y_4)\vert(\boldsymbol{X}, Y_1, Y_2)$	(c)	(d)	$D \perp\!\!\!\perp Y_3\vert(\boldsymbol{X}, Y_1, Y_2)$
$D \perp\!\!\!\perp Y_4\vert(\boldsymbol{X}, Y_1, Y_2, Y_3)$	(e)	(f)	$D \perp\!\!\!\perp Y_4\vert(\boldsymbol{X}, Y_1, Y_2, Y_3)$

Table 2: Conditions (4.10) and (4.11) for $T = 4$.