The ambiguity of Missing at Random: A study using staged trees & CEGs

Kate Gamper
jekaterina.gamper@warwick.ac.uk

Department of Statistics, University of Warwick, United Kingdom

Abstract

In real-world scenarios, most datasets have some missing values. Therefore, it is important to consider the missingness mechanism, since if missing values are not correctly handled, any inference will be biased. There are three main missingness mechanisms: Missing Not at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). MCAR occurs very rarely and MNAR complicates the analysis. On the contrary, MAR makes analysis easy because of the ignorable likelihood, which makes it a widespread assumption. However, two different definitions have been used interchangeably to define MAR – Realised MAR and Everywhere MAR. Two definitions came into existence from working in different frameworks. This report aims to review the differences in the definitions and discuss how they should be interpreted. Further, we provide an illustrative example of the research on a dataset of preterm infants. We use a probabilistic event-level graphical model – Chain Event Graphs, to explore the missingness. In addition, we show that a Chain Event Graph is a useful tool for identifying sparsity in the dataset and allows exploring patterns of missingness.

Abstract

Nowadays, research hugely relays on data analysis. However, it is extremely rare for the data set that is being investigated to be fully complete, i.e. to have no missing values. If the missingness assumptions are not properly accounted for, it can lead to incorrect conclusions. Therefore, it is important for a researcher to think about missing values and mechanisms that might have caused the missingness. There exist three main missingness mechanisms. However, there is considerable ambiguity in how some of the missingness mechanisms should be interpreted. This clearly causes a problem when a researcher is trying to establish the missingness mechanism in a practical problem. One of the contributing factors towards this ambiguity is that two different definitions have been used interchangeably. In this report, we review the different yet related definitions of the missingness mechanisms and discuss how these can be used in practice. Further, by using a clinical trials dataset of preterm infants, we illustrate the benefits of using Chain Event Graphs – a type of graphical model – in identifying the missingness mechanism as it applies to different subgroups within our dataset.

Contents

1 Introduction

Complete data sets without any missing values are extremely rare. For example, in a clinical trial, a person can drop out due to moving to another city and stop attending the appointments. Preceding Rubin's seminal work (Rubin, 1976) in which he characterises missing data, missingness was often treated as a nuisance which distracted away from the data analysis exercise. Moreover, this is complicated by the fact that many statistical procedures expect complete data and are unable to handle missingness. Statisticians had to either delete all inputs with missing entries or fill in those entries. However, if the underlying cause for missing data depends on the data itself it means the inference will be biased for the target population, since the important behaviour and trends cannot be spotted.

Rubin, 1976 has described the main mechanisms that drive missingness and raised the question of why it is important to consider them. Following Rubin's seminal paper, missingness has become an active field of research. There are three main defined types of missingness mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR), described further in Section 2. MCAR is the simplest form to deal with but it does not occur often in practice. Similarly, MAR makes analysis easy because of the ignorable likelihood (Rubin, 1976). Therefore, the main question for analyst is to distinguish between MAR and MNAR but the required data to do that is missing. MNAR complicates data analysis, since it requires to explicitly model the missingness processes. Thus, MAR is a convenient and common assumption, since the observed data can be directly used to perform inference (through the ignorable likelihood, see Section 2) without having to explicitly model the missingness mechanism. More recent research has brought to light that there are in fact two different but related definitions of MAR that have been used interchangeably. We present those definitions in Section 3 of the report. The confusion in MAR definitions can cause problem for practising data analyst. This report aims to discuss how different definitions of Missing at Random should be considered in different scenarios.

In order to study missingness, it is useful to have a visual representation of it. In particular, recent studies have found that the staged trees and the Chain Event Graphs (CEGs) framework have been useful in representing and analysing missing data (Barclay et al., 2014; Shenvi et al., 2019). A CEG is a probabilistic graphical model that captures asymmetries and context-specific conditional independencies of the underlying process (Smith & Anderson, 2008), further introduced in Section 4. A CEG is a transformation of its underlying event tree that describes the process being studied. In addition, a staged tree is an intermediate model in the transformation of an event tree to a CEG. The CEG framework provides a more compact representation of the process by leveraging the symmetries within the events of the process. Both staged trees and CEGs are event-level models which help us to study subgroup-based differences in the missingness.

Finally we use a CEG on a chosen data from a longitudinal cohort of infants of gestational age $\leq$ 32 weeks (Hutton et al., 1997) for illustration purposes. The data is presented in Section 5 and graphical model for the data in Section 6. Section 7 includes discussion on representation of missingness mechanisms and how the Missing at Random should be considered.

2 Missingness mechanisms

Missingness mechanism is the main cause of the missing data, which establishes the link between the present and absent information as well as any other measured covariate. In order to introduce the three types of missingness mechanisms as defined by Rubin, we first introduce the notation. Let $X = (X_1, X_2, \ldots, X_n)$ be a vector random variable that is completely observed such that it does not have any missing values and $Y = (Y_1, Y_2, \ldots, Y_n)$ is a vector random variable that is only partially observed (some entries are missing). Further suppose that Y is the response variable of interest and X is an explanatory covariate. Let parameter ϕ describe the $Y|X$ distribution. Let $R = (R_1, R_2, \ldots, R_n)$ be the missingness indicator random variable such that $R_i = 1$ if Y_i is observed and $R_i = 0$ if Y_i is missing. Let θ be a parameter that governs the missingness generating process conditional on Y, X. Assuming independence of ϕ and θ, the joint probability of Y and R decomposes as $P(Y, R|X, \theta, \phi) = P(R|Y, X, \theta)P(Y|X, \phi)$ by the chain rule.

Further we introduce three missingness mechanisms and provide a graphical representation for each of them in Figure 1.

Missing Completely at Random – MCAR Missing Completely at Random happens when the mechanism that governs whether or not data is missing does not depend on observed or unobserved data,

$$P(R|Y, X, \theta) = P(R|\theta).$$

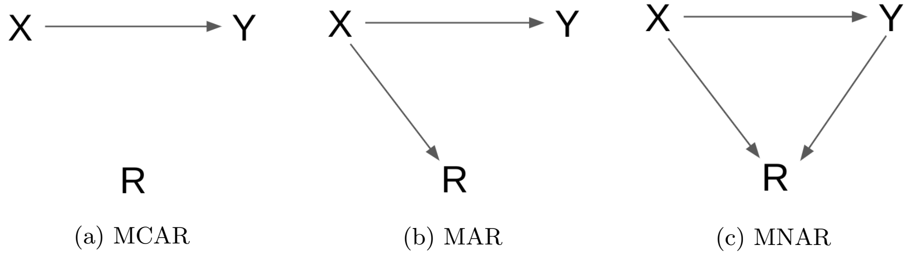
Missing at Random – MAR The mechanism that governs whether or not data is missing depends on the observed data, but not the unobserved data,

$$P(R|Y, X, \theta) = P(R|X, \theta).$$

Missing Not at Random – MNAR The missingness mechanism depends on some of or all the unobserved values as well as observed values,

$$P(R|Y, X, \theta) = P(R|Y, X, \theta).$$

Let us introduce a situation in order to give an example for the missingness mechanisms. Suppose, a data scientist is working on survey data based on a questionnaire about political freedoms. Questions were designed to determine the political freedom in the country. Suppose that X represent entries describing the age of the participant and Y is the the answer to the question “Do you feel free to express you political opinion?”. If some of the values of Y were accidentally left out due to a data entry error, the missingness mechanism would be MCAR. Otherwise, if some entries for people aged over 60 are missing because they are reluctant to answer the question but not because they don't feel free to express their political opinion – the missingness mechanism would be MAR. Finally, if people would not answer the main question because there is not enough political freedom in the country such that people don't feel comfortable expressing their opinion, the missingness mechanism would be MNAR.



(a) MCAR (b) MAR (c) MNAR

Figure 1: Graphical representations of the dependencies between observed, unobserved data and missingness mechanism for different type of missingness mechanisms

MCAR doesn't occur frequently and there exist several tests for it (Berrett & Samworth, 2022; Little, 1988). On the other hand, MNAR assumption complicates the analysis since it requires to model the mechanism in order to do any further inference. Therefore, it is very common for researcher to assume MAR instead for the dataset since it is very convenient assumption. MAR allows to simplify analysis by disregarding the missingness mechanism. It allows to assume the ignorable likelihood of parameter ϕ given X and observed values of Y. Several methods for handling missingness have been developed, often aimed for MAR data. For a review, see Appendix A.

3 Ambiguity in the use of different MAR definitions

Missing at Random is the most common assumption since it allows researcher to proceed with analysing the dataset without explicitly modelling the missingness mechanism. However, there exists two different but related MAR assumptions and it happened so that statisticians have been using both of them interchangeably without actually specifying the difference as pointed out by Seaman et al. (2013). In order to introduce MAR definitions, we extend notation introduced in Section 2. Let realised samples of Y and X be represented by y and x. In addition, let y^* and $\tilde{y}$ be theoretical repeated samples of Y. Realised sample for missingness indicator R is defined as r. Function $o(y, r)$ returns observed values of y where $r_i = 1$. Example of observed function for the sample is provided in Table 1. If y would be fully observed s.t. $r_i = 0 \, \forall \, i$, then $o(y, r) = y$.

We now introduce the two definitions of MAR: realised MAR and everywhere MAR (Doretti et al., 2018).

Realised MAR – RMAR Realised MAR interprets MAR condition as it holds exclusively for the realised dataset. RMAR allows only one missingness pattern $\tilde{r}$, which is the realised pattern of the sample. It does not consider other missingness patterns that could have been observed. Example of RMAR for 2 hypothetical datasets is presented in Table 2.

$$P(r|y,x) = P(\tilde{r}|\tilde{y},\tilde{x}) \,\forall\, \tilde{y}$$

$$\text{where } r = \tilde{r} \wedge o(y,r) = o(\tilde{y},r)$$

Everywhere MAR – EMAR Everywhere MAR interprets MAR as if it holds for several repeated samples of Y. Therefore EMAR allows for different missingness patterns to occur in different samples. EMAR generalised RMAR assumption as it concerns the process under study and not just the realised dataset.

$$P(r^*|y^*,x^*) = P(\tilde{r}|\tilde{y},x) \,\forall\, \tilde{y}, y^*$$

$$\text{where } o(y,r) = o(\tilde{y},\tilde{r})$$

y	r	$o(y,r)$
y_1	1	y_1
y_2	0	-
y_3	1	y_3
y_4	1	y_4
y_5	0	-

Table 1: Example of observed function for the specific sample missingness pattern, where realised y is of dimension 5 and the observed vector is $o(y,r) = (y_1, y_3, y_4)$.

y	$\tilde{y}$	$\tilde{r}$	$o(\tilde{y},\tilde{r})$
1	1	1	1
a	c	0	-
4	4	1	4
6	6	1	6
b	d	0	-

Table 2: Example of RMAR if there would exist two realised samples of Y: y and $\tilde{y}$. Missingness indicator for both samples should be the same, assuming RMAR.

We review here why this discrepancy might have happened. Seaman et al., 2013 describes that one reason might be due to the framework under which the analysis is conducted. Rubin in his work defined the RMAR definition. At that time Rubin thought about missingness when he was looking at direct likelihood so he was thinking about missingness in terms of only one data set. So the RMAR arose from thinking in the Bayesian framework. While EMAR assumes repeated sampling and was thought of in the frequentest framework. However, we can also think about EMAR when doing Bayesian inference. Since RMAR only concerns the realised dataset and not other datasets that can be sampled from the population, we argue that RMAR is more of a mathematical condition that leads to good statistical properties rather than a modelling assumption. In fact, to consider the missingness at a process-level, we would have to think in terms of EMAR. An exception to this might be where the sample and population are equivalent, i.e. RMAR and EMAR coincide.

4 Chain Event Graphs

4.1 Creating a chain event graph

A chain event graph (CEG) (Smith & Anderson, 2008) is an event-based probabilistic graphical model that allows to study categorical data. In order to construct a CEG it is necessary to create an event tree from the dataset. An event tree is a logical model that encapsulates the sequence of possible events and their probabilities. Event trees have been widely implemented to analyse the underlying process of the data (Shafer, 1996). However, an event tree can become quite big and hard to read. A CEG allows to represents the same process as an event tree in a more compact way by combining some of the events with the same probabilities (Shenvi & Smith, 2020). There are several steps to create a CEG out of a event tree: first an event tree has to be turned into a staged tree and after that a staged tree is reconstructed into a CEG.

Steps to convert an event tree into a CEG

1. In the event tree colour non-leaf vertices with the same colour if their emanating edges represent the same succeeding event and also have same conditional probability. This creates a staged tree.

2. Vertices that have the same colour and their rooted sub-trees are the same[1] should be merged into a single vertex, which preserves the colour.

3. All the leaves are merged into a single vertex – sink.

Let us further introduce a staged tree and CEG formally. We denote a directed graph $\mathcal{G}$ which incorporates a finite vertex set V and a finite edge set $E \subseteq V \times V$ such that $\mathcal{G} = \{V, E\}$. Let e_{ij} be a directed edge from a vertex v_i to a vertex v_j where $i \neq j$. Then v_i is said to be a parent vertex of v_j and v_j is a child vertex. The vertex v_0 that has no parent is called root. Vertices with no existent children are called leaves. Define a vertex v with all its emanating

[1] Rooted sub-trees are considered the same when they have identical structure and all vertices preserve the colour.

edges $E(v)$ as a floret $F = (v, E(v))$. Let $\theta(e_{ij})$ denote the conditional transition parameter along the edge $e_{ij} \in E$. Finally, $\underline{\theta_v} = \{\theta(e), e \in E(v)\}$ is a vector of associated transition parameters for a floret of vertex v.

Definition 1 (Event tree) *An event tree $\mathcal{T}$ is a connected, directed graph $\mathcal{G}$ that has no cycles, where the sum of $\underline{\theta_v}$ entries is 1 for each vertex v in a tree such that $\sum_{e \in E(v_i)} \theta(e) = 1 \forall i$.*

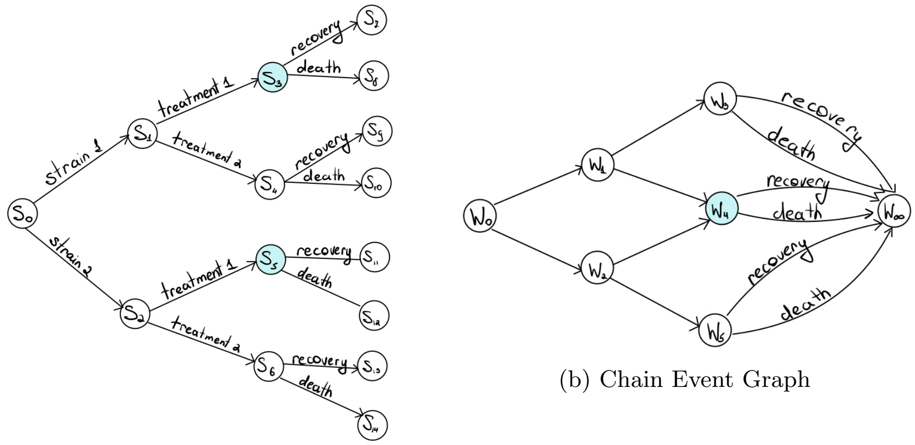
Definition 2 (Staged tree) *An event tree $\mathcal{T}$, where vertices v and v^* are in the same stage such that they are coloured by one colour whenever $\underline{\theta_v} = \underline{\theta_{v^*}}$ is defined as a staged tree $\mathcal{S}$.*

We say that vertices v and v^* of a staged tree $\mathcal{S}$ are in the same position k, when $\underline{\theta_v} = \underline{\theta_{v^*}}$ and vertices' subtrees have the same topology and colour such that $\mathcal{T}(v) = \mathcal{T}(v^*)$. Let $W_{\mathcal{S}}$ be a set of positions of a staged tree $\mathcal{S}$.

Definition 3 (Chain Event Graph) *Let $\mathcal{C}_{\mathcal{S}}$ be a chain event graph for an associated staged tree $\mathcal{S}$ such that $\mathcal{C}_{\mathcal{S}} = \{V(\mathcal{C}_{\mathcal{S}}), E(\mathcal{C}_{\mathcal{S}})\}$ where a vertex set $V(\mathcal{C}_{\mathcal{S}}) = W_{\mathcal{S}} \cup w_\infty$. Vertices in a set $V(\mathcal{C}_{\mathcal{S}})$ inherit their stage colour from $\mathcal{S}$. All leaves of $\mathcal{S}$ are merged into the single vertex w_∞, called sink. All emanating edges whose vertices where combined into the vertex w are merged into the same edge if they share the same label $\theta(e)$.*

Note on the non-coloured vertices For simplicity, when a vertex in a stage by itself, we do not display its colouring to minimise visual cluttering. If the subtrees have the same structure but their subsequent vertices are not coloured we treat them as different. To combine vertices subtrees have to have the same topology.

Example 1. Suppose that we are trying to analyse the probability of recovery for two different strains of illness when they are treated with two different treatments. Illustrative staged tree for this example is presented in Figure 2a. In this example vertices s_3 and s_5 are coloured with the same colour, which shows that those vertices are in the same stage. In terms of the example, this means no matter which strain is treated using treatment 1, possibility of recovery or death is the same for both strains. To convert this staged tree into a CEG we need to merge vertices s_3 and s_5 into one, since they are of the same colour and have same subtrees. In addition, all leaves $s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}$ have to be combined into one vertex called sink. The final CEG is shown in Figure 2b.



(b) Chain Event Graph

(a) Staged tree

Figure 2: Example of a staged tree and CEG for treating two different strains of illness with 2 different treatments

4.2 Structure Learning for CEGs

To learn a CEG from an event tree, we can use data, expert judgement or a combination of both. We now consider how a CEG can be learnt from data. In particular, here we consider the Bayesian Dirichlet metric for structure learning.

To find the score for CEG we assume that the parameters for each vertex are independent a priori. Therefore, we can treat each vertex separately, and only look at edges coming out of each vertex. For each vertex, we consider its prior and likelihood to find the posterior.

We consider CEG $\mathcal{C}$ with collection of stages $U = u_1, u_2, ..., u_m$. Stages are how many different colours we have. Denote by k_i the number of outgoing edges from each vertex in stage set u_i, $i = 1, 2, 3, .., m$

We let the conditional transition parameter vector for each vertex in stage u_i with k_i outgoing edges be denoted by $\underline{\theta_i} = (\theta_{i1}, \theta_{i2}, ..., \theta_{ik_i})$ where θ_{ij} is a probability that an individual in some situation $s \in u_i$ passes over along its outgoing edge j, for $j \in 1, 2, 3, ..., k_i$. Graphical representation is provided in Figure 3.

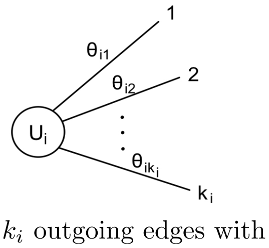


Figure 3: Vertex u_i with k_i outgoing edges with event probabilities θ_{ij}

The vector $\underline{\theta_i}$ represents the event probability parameters of a Multinomial distribution. Therefore, $\sum_{j=1}^{k_i} \theta_{ij} = 1$ and $\theta_{ij} > 0$.

Assuming we have a complete dataset such that $\underline{y} = (\underline{y_1}, \underline{y_2}, ..., \underline{y_k})$. The vector $\underline{y_i} = (y_{i1}, y_{i2}, ..., y_{ik_i})$ summarises the number of individuals $y_{ij}, j \in 1, 2, 3, ..., k_i$ that start in some situations $s \in u_i$ and traverse along its jth edge. Assuming a random sample, individuals are independent. Therefore, vertex likelihood:

$$p(\underline{y_i}|\underline{\theta_i}, \mathcal{C}) = \sum_{j=1}^{k_i} \theta_{ij}^{y_{ij}}$$

Let's assume that each parameter $\underline{\theta_i}$ has a Dirichlet prior distribution with parameter vector $\underline{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ik_i})$ s.t. $\alpha_{ij} > 0$ for $j \in 1, 2, ..., k_i$.

$$\text{Dirichlet prior for a situation } s\text{: } p(\underline{\theta_i}|\mathcal{C}) = \frac{\Gamma(\overline{\alpha_{ij}})}{\prod_{j=1}^{k_i} \Gamma(\alpha_{ij})} \prod_{j=1}^{k_i} \theta_{ij}^{\alpha_{ij}-1}$$

$$\text{where } \overline{\alpha_{ij}} = \sum_{j=1}^{k_i} \alpha_{ij}$$

Therefore, posterior of a situation s is also a Dirichlet distribution:

$$p(\underline{\theta_i}|\underline{y_i}, \mathcal{C}) \propto p(\underline{\theta_i}|\mathcal{C}) p(\underline{y_i}|\underline{\theta_i}, \mathcal{C})$$

$$\propto \prod_{j=1}^{k_i} \theta_{ij}^{\alpha_{ij}-1} \theta_{ij}^{y_{ij}}$$

$$p(\underline{\theta_i}|\underline{y_i}, \mathcal{C}) = \frac{\Gamma(\overline{\beta_i})}{\prod_{j=1}^{k_i} \Gamma(\beta_{ij})} \prod_{j=1}^{k_i} \theta_{ij}^{\beta_{ij}-1}$$

$$\text{where } \beta_{ij} = \alpha_{ij} + y_{ij} \text{ and } \overline{\beta_{ij}} = \sum_{j=1}^{k_i} (\alpha_{ij} + y_{ij})$$

In order to decide whether any two vertices should be in the same stage, we compare the log likelihood of two CEG models which are identical except that one of them has these two vertices in separate stages and the other has them in the same stage. Then we can calculate the marginal likelihood of the new structure and the old structure. Thus, we can reject or accept the new structure by evaluating the Bayes Factor given by the ration of their marginal likelihood.

Bayesian Agglomerative Hierarchical Clustering (AHC) algorithm is a greedy algorithm that at every iteration chooses two best situations to merge, which would improve the Bayes Factor. It proceeds doing so as long as there no more vertices to merge. AHC implements steps described in this section (Freeman & Smith, 2011). The algorithm was used to create CEGs for this report.

5 An illustrative Example

5.1 Description of the Dataset

The data set was used from the study of preterm infants born in 1980-1981 aimed to investigate the effects of preterm birth and being small for gestational age on a child's cognitive and motor abilities (Hutton et al., 1997). The infants chosen for the study where all born at ≤ 32 weeks gestational age and weighed ≤ 2000g. The cohort consisted of 180 infants. All mothers were residents in Cheshire and Merseyside counties. At the age of eight/nine years old children were assessed according to the Wechsler Intelligence Scale for Children, the Neale analysis of reading ability, and the Stott-Moyes-Henderson test of motor impairment. Fifteen children with diagnosed severe disabilities were excluded from the IQ assessment as well as nine children who refused to participate or immigrated. The study encompassed 31 variable in total to describe each child. For the purpose of illustration, we focus on five variables in this report:

- Birth weight (BW) – child's weight at birth. Fully observed variable.
- Gestational age (GA) – number of weeks since the first day of the woman's last menstrual cycle to birth. Fully observed variable.
- Birth weight ratio (BWR) – a ratio of observed birth weight to the expected birth weight for a given gestational age. Fully observed variable.
- Age when mother left education (AMLE) – this information was collected from the questionnaire and since not every person provided the answer the variable has missing entries.
- Intelligence quotient (IQ) – IQ was assessed using the Wechsler Intelligence Scale for Children (WISC) 1974 revision (British version). The variable was only partially observed due to 24 entries missing (children with severe disability and refused).

The data set is used for the illustrative purposes only. The simplified version was chosen such that only one variable has missingness. The report follows the procedure presented by (Barclay et al., 2014), such that response variable is fully observed and missingness appears only in one of the covariates. IQ was chosen as the response variable. Therefore, missing data was removed from IQ variable such that it would be fully observed. AMLE is the only covariate with missing entries. Hence, the dataset used in the remainder of the report consisted of 158 entries out of 180.

5.2 Data categorisation

In order to analyse the data set using CEG, the data had to be categorised.

The World Health organisation defines following subgroup for birth weight: extremely low BW (less than 1000g), very low BW (less than 1500g), moderately low BW (less than 2500g), and pre-term gestational age is classified into:

extremely preterm (less than 28 weeks), very preterm (28 to 32 weeks), moderate to late preterm (32 to 37 weeks). However, our dataset consists of only 158 entries and categorising a variable into three subgroups results in sparsity. Thus, the classification was changed into binary variables for illustration purposes. To turn birth weight into binary variable several classification were attempted:

1. extremely low BW and very low BW groups were combined s.t. two classes became : 'extremely & very low BW' and 'moderately low BW'
2. very low BW and moderately low BW groups were combined s.t. two classes became : 'extremely low B' and 'very & moderately low BW'
3. birth weight was divided according to the median such that two classes became: 'below median' and 'above median'

The gestational age variable was classified into two groups according to the median: 'below median', 'above median'.

Variable age when mother left education was categorised into 3 groups: left education 'before 16', left education 'after 16' (including 16), age is 'missing'. Age of 16 was chosen as a threshold due to the fact that 16 was established as minimal school leaving age in 1972 ("Education Act", 1968).

The variable IQ was categorised into 2 groups: IQ 'below 100', IQ 'above 100' (including 100).

6 Staged trees and CEGs for the data set

The staged trees and CEGs for this report were produced using *cegpy* python package. *cegpy* package uses Agglomerative Hierarchical Clustering (AHC) algorithm to find a CEG given the dataframe. In this research, priors were set to the default as defined by the package.

6.1 Sparsity representation using staged trees and CEGs

Figure 4 shows the staged tree for the data set when extremely and very low BW are combined into one class. Staged tree allows us to see the sparsity of the subgroup which includes infants of moderately low BW and who are born before 29 weeks of gestational age. This subgroup is represented by the subtree $s_0 \Rightarrow s_2 \Rightarrow s_5$. Figure 5 depicts the CEG for this staged tree and sparse subgroup is faded out. Figure 7 depicts the staged tree when very and moderately low BW are combined into one class. We can see a similar case of sparse subgroup when a baby is born with extremely low birth weight and over 29 weeks gestational age. The branch $s_0 \Rightarrow s_1 \Rightarrow s_4$ represents this subgroup. Figure 6 depicts the CEG for this staged tree and sparse subgroup is faded out.

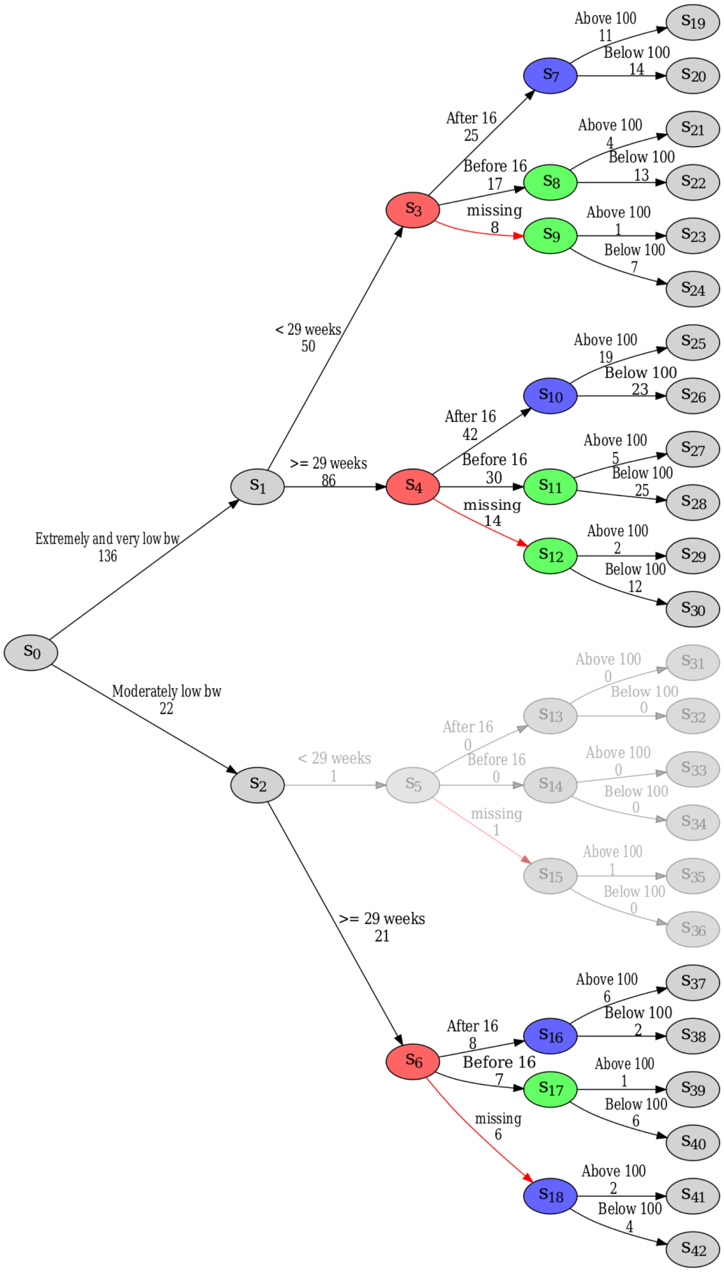


Figure 4: Staged tree for the data set where birth weight is categorised into 2 groups: 1) extremely and very low BW (< 1.5kg), 2) moderately low BW (≥ 1.5kg)

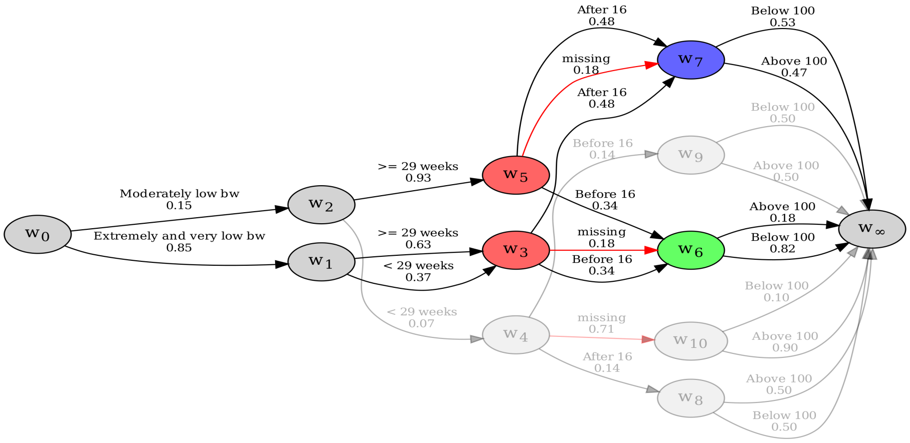


Figure 5: CEG for the data set where birth weight is categorised into 2 groups: 1) extremely and very low BW (< 1.5kg), 2) moderately low BW (≥ 1.5kg)

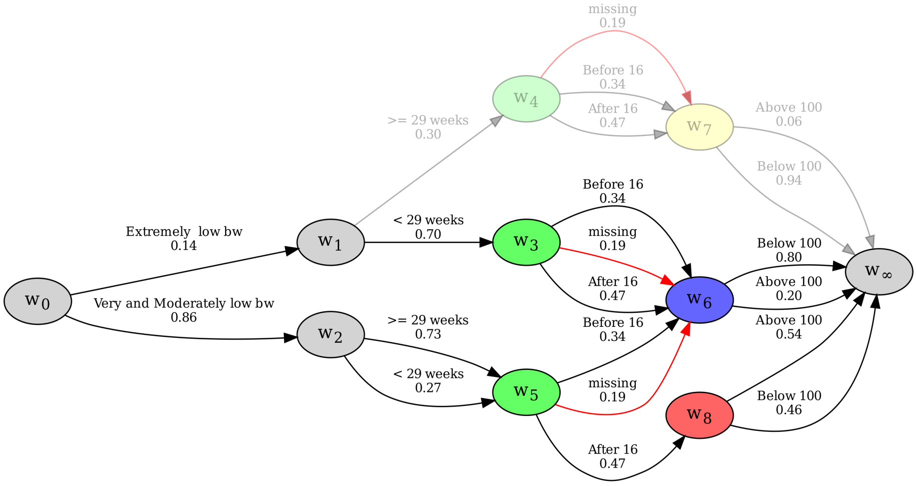


Figure 6: CEG for the data set where birth weight is categorised into 2 groups: Extremely low birth weight (< 1kg), Very low and Moderately low birth weight (≥ 1kg)

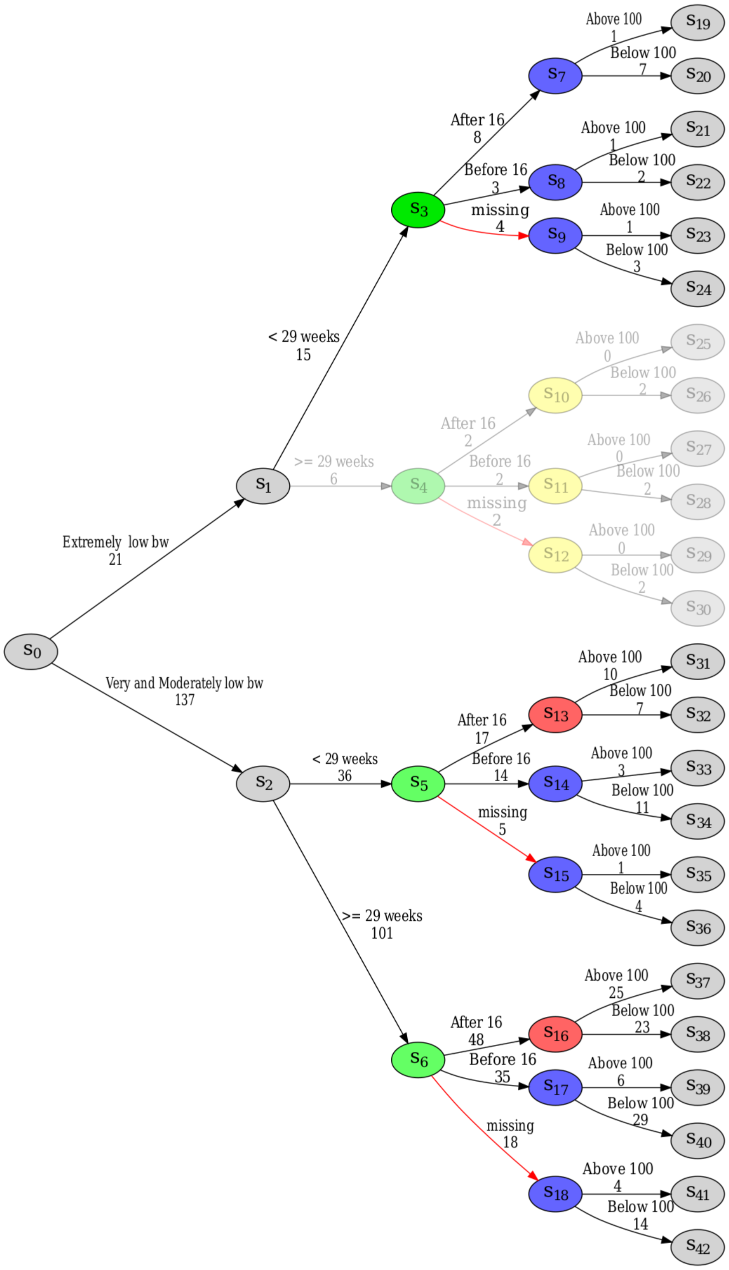


Figure 7: Stage tree for the data set were birth weight is categorised into 2 groups: 1) extremely low birth weight(< 1kg), 2) very low and Moderately low birth weight (≥ 1kg)

We can see that such graphical representation allows data analyst to assess the sparsity of the data set and examine whether this sparsity is acceptable for representation of the target population. In the case of this data set, the sparsity is acceptable since when a baby is born with extremely low BW it is much less likely to be over 29 weeks gestational age. Similarly, it is unlikely for an infant to have gestational age of less than 29 weeks and have moderate BW. Sometimes, as in this case, sparsity is unavoidable for certain subgroups but staged trees make it clear which subgroups we can confidently talk about, based on the amount of data we have for them. However, this sparsity would not be highlighted by other classical methods such as linear regression, as was done in (Hutton et al., 1997).

6.2 Missingness representation in staged trees and CEGs

Despite the sparsity, the staged trees and CEGs still allow to analyse missingness in other subgroups that have sufficient counts[2]. In both figures 5 and 6, 'missing' class is grouped together with the class when mother left education 'before 16' for at least one of the subgroups. This suggests that missingness can be informative, such that the missingness mechanism is MNAR. In both of those examples, the number of children with IQ less than 100 is much bigger than the number of children with the IQ over 100. On the contrary, when mother left education after 16, the number of children whose IQ is below 100 is either only slightly bigger or smaller than the number of children with IQ above 100.

For illustration purposes the BW is divided into two groups according to the BW median. We have taken this approach to try to mitigate the problems associated with having sparse subtrees. The resultant CEG and staged tree are presented in Figure 8 and Figure 9 respectively. However, the sparse sub-tree $s_0 \Rightarrow s_1 \Rightarrow s_3$ is still present. The missingness also appears to be informative, since missingness has the same colour as 'before 16' vertex. In the CEG we can see that for subgroups that are not sparse, edge representing class 'after 16' is directed to the same vertex w_7.

[2] Whether subgroup has sufficient count depends on what is the aim of the modelling and what the domain is

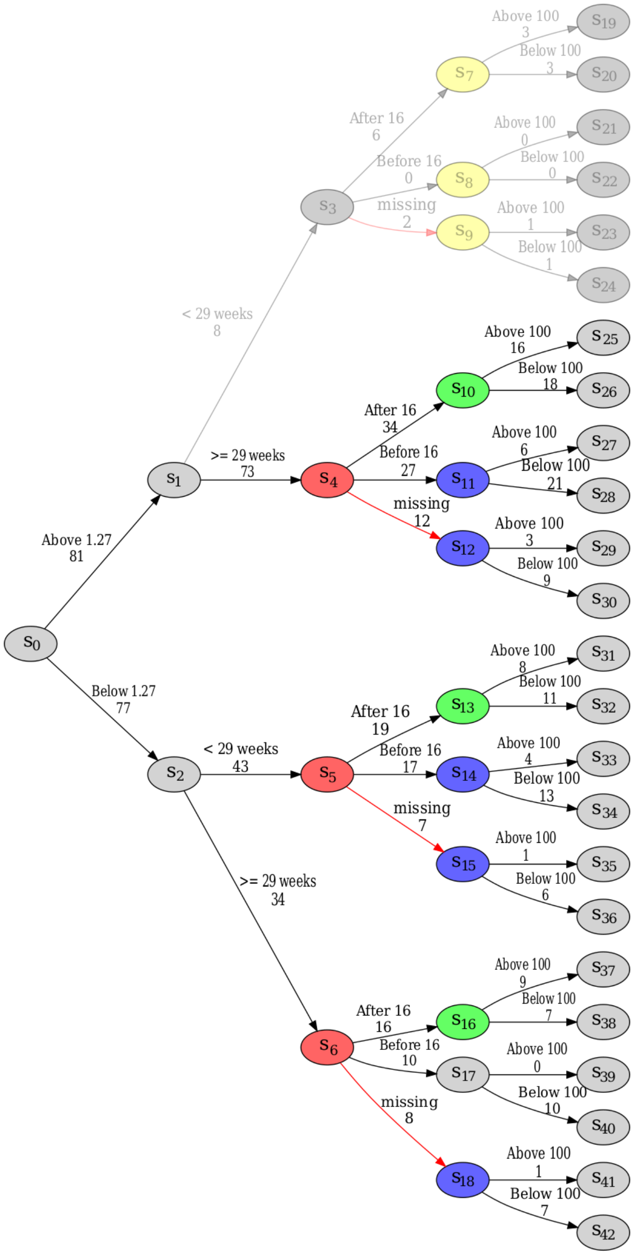


Figure 8: Stage tree for the data set were birth weight is categorised into 2 groups: birth weight below median of the data set(< 1.27kg), birth weight above median of the data set(≥ 1.27kg)

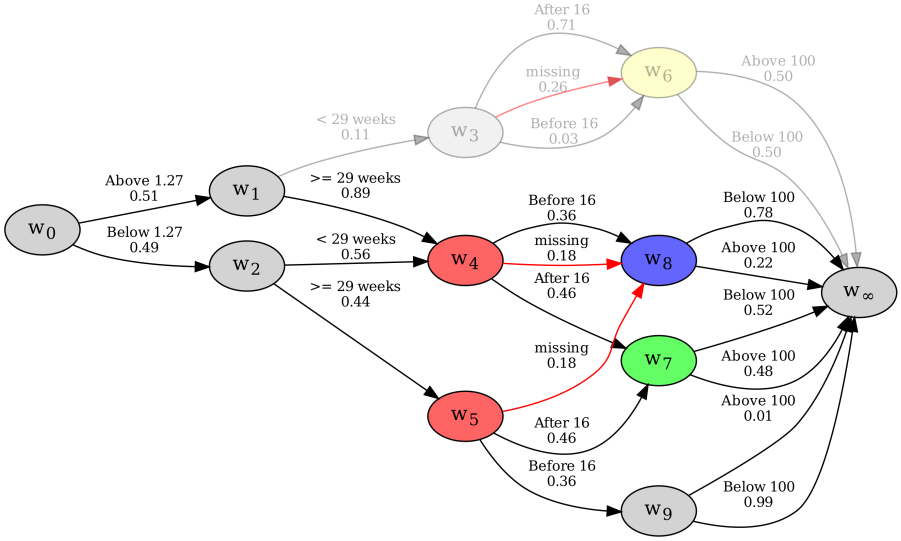


Figure 9: CEG for the data set were birth weight is categorised into 2 groups: birth weight below median of the data set($<$ 1.27kg), birth weight above median of the data set($\geq$ 1.27kg)

However, classifying the BW variable by median still did not allow us to get rid of sparsity. This is because as discussed above, some of the subgroups are naturally sparse due to the combinations of BW and GA. Thus, BW and GA were changed to BWR for illustrative purposes. Figure 10 present CEG for variable BWR, AMLE and IQ. In this CEG 'missing' class and 'before 16' are combined into one vertex w_3. This behaviour is similar to all the examples that were examined before.

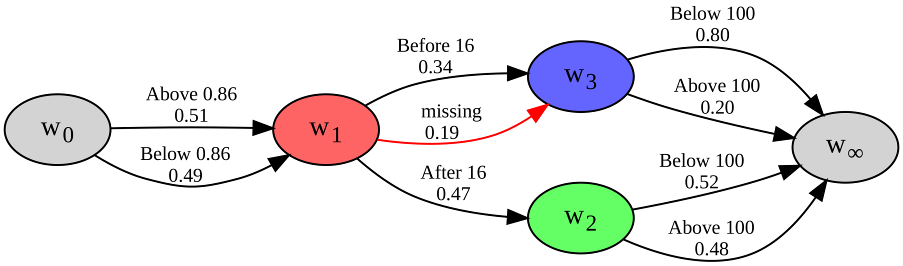


Figure 10: CEG for the data set were following 3 variables are taken: birth weight ratio, age when mother left education, IQ. Birth weight ratio was categorised into 2 groups: below median and above median

7 Discussion on the Realised and Everywhere MAR

The missingness in the AMLE variable was likely to be informative as showed in the previous section. Therefore, data scientist would have to assume MNAR and first model the missingness mechanism before proceeding with further analysis.

However, let's assume that this is not the case and we see different scenario in our staged trees and CEGs. The expected posterior conditional transition probabilities for the edges when IQ is above 100 and below 100, for the infants whose AMLE is missing, would be the average of when AMLE is not missing. For example, let's look at the case of when BW is divided according to the median in Figure 8. In particular, consider the subgroup of infants whose BW is below median, and GA below 29 weeks and whose AMLE is missing ($s_0 \Rightarrow s_2 \Rightarrow s_5 \Rightarrow s_{15}$). Thus, expected conditional transition probability $\theta_{15,1}$ for the edge 'above 100' would be the average of $\theta_{14,1}$ and $\theta_{13,1}$. Similarly, $\theta_{15,2}$ for the edge 'below 100' would be the average of $\theta_{14,2}$ and $\theta_{13,2}$. Then, as suggested by (Barclay et al., 2014) it would indicate that data is MAR. Let's refer to this hypothetical MAR dataset as hyp-MAR dataset. What would it mean for a data analyst depending on which MAR definition they are using?

RMAR Thinking in terms of Realised MAR would mean that the fact that we can see MAR in hyp-MAR dataset is enough and we don't need to think about processes that caused missingness. However, in this example it would lead to the wrong assumption. It might have been an accident that only in this particular sample it does not matter whether mother provided her age of leaving education or not for child's IQ.

EMAR Thinking in terms of Everywhere MAR would force as to think what would happen if we would sample again from the population and whether we would get MAR again. Therefore, EMAR encourages data analyst to think about the mechanisms that drive missingness and evaluate whether they are plausible. In case of pre-term infants study, we can think that it might be possible that women are more reluctant to provide the age at which they left education if they left education early. Then we would investigate how it can affect child's IQ s.t.: if mother left education earlier she might not be familiar with opportunities available to her child to help with the development.

We believe that RMAR is more of a descriptive statement than a modelling assumption, since it does not require to think about the missingness mechanism. When researcher wants to generalise their work or think about the overall target population, they would always end up thinking about the process of data collection systematically and would be forced to consider EMAR.

Acknowledgements

This project was supported by the Heilbronn Institute for Mathematical Research and was conducted at the Applied Statistics & Risk Unit in the Statistics Department at Warwick University.

I would like to express my gratitude to my supervisor Dr Aditi Shenvi for introducing me to such an interesting and important topic as well as supporting me throughout the project. Thank you!

I am also thankful to Prof Jane Hutton for her invaluable suggestions and comments on my work.

A Methods to deal with missing data

Due to the fact that most of the data analysis techniques are designed for complete data sets, we need to either delete all inputs with missing entries or fill in those entries. The appendix provides the small list of some of the techniques that are used to deal with missing entries.

A.1 Complete Case Analysis CC

Delete the whole data entry row if at least one of the entries is missing. The main advantage of such technique is simplicity. However, it requires to throw away potentially useful information, for which resources were used to collect.

A.2 Weighting

If we assume that data is not MCAR, we can weight to attempt to adjust for bias.

Example when it is used:If we think that respondents to a survey that fit a certain category are more or less likely to respond to certain questions in the survey, then we can estimate the probability of how likely from the information that is already available. Later, it is possible to weigh to make sure this group of responders is not under-represented.

A.3 Available Case Analysis

We use all the observations for a category where observation is recorded for that category, even if that response has missing values for other categories. However, this method has drawbacks of having different sample sizes for each category. It might not be more efficient than CC analysis if the number of entries is missing is small comparing to the whole data set. It might be more computationally demanding to account for different sample sizes while the gain would not be very big.

A.4 Single Imputation Methods

Imputation is the process of replacing the missing data with substituted values. There are different imputation methods.

Unconditional Mean Imputation Missing data is replaced with the mean of the other entries for that category.

Regression Imputation Uses regression to predict what a missing value would be, using the observed values to inform the regression equation.

Stochastic Regression Imputation Input missing values that deviate from regression line in a random pattern.

Hot Deck Imputation Fill in missing values by using recorded responses from similar respondents in a survey/dataset.

Bootstrap Resample with replacement several times from the observed data and then perform imputation on each resample data set.

Multiple Imputation Chained Equation MICE This method is for multivariate missing data. In this method we do imputation and then update missing values for $t+1$ step by drawing them from a sequence of predictive distribution, given imputed values in step t.

References

Barclay, L. M., Hutton, J. L., & Smith, J. Q. (2014). Chain event graphs for informed missingness. *Bayesian Analysis*, *9*(1), 53–76.

Berrett, T. B., & Samworth, R. J. (2022). Optimal nonparametric testing of missing completely at random, and its connections to compatibility. *arXiv preprint arXiv:2205.08627*.

Doretti, M., Geneletti, S., & Stanghellini, E. (2018). Missing data: A unified taxonomy guided by conditional independence. *International statistical review*, *86*(2), 189–204.

Education act. (1968). https://www.legislation.gov.uk/ukpga/1968/17/contents

Freeman, G., & Smith, J. Q. (2011). Bayesian MAP model selection of chain event graphs. *Journal of multivariate analysis*, *102*(7), 1152–1165.

Hutton, J. L., Pharoah, P. O., Cooke, R. W., & Stevenson, R. C. (1997). Differential effects of preterm birth and small gestational age on cognitive and motor development. *Archives of disease in childhood. Fetal and neonatal edition*, *76*(2), 75–81.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, *83*(404), 1198–1202.

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

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

Shafer, G. (1996). *The art of causal conjecture*. MIT Press.

Shenvi, A., & Smith, J. Q. (2020). Constructing a chain event graph from a staged tree. In M. Jaeger & T. D. Nielsen (Eds.), *Proceedings of the 10th international conference on probabilistic graphical models* (pp. 437–448). PMLR.

Shenvi, A., Smith, J. Q., Walton, R., & Eldridge, S. (2019). Modelling with non-stratified chain event graphs. *Bayesian Statistics and New Generations*, 155–163.

Smith, J. Q., & Anderson, P. E. (2008). Conditional independence and chain event graphs. *Artificial intelligence*, *172*(1), 42–68.