RESEARCH ARTICLE

Assessable and interpretable sensitivity analysis in the pattern graph framework for nonignorable missingness mechanisms

Alireza Zamanian[1,2] | Narges Ahmidi[2,3] | Mathias Drton[4]

[1]TUM School of Computation, Information and Technology, Department of Computer Science, Technical University of Munich, Munich, Germany

[2]Department of Reasoned AI Decisions, Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany

[3]Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany

[4]TUM School of Computation, Information and Technology, Department of Mathematics, Technical University of Munich, Munich, Germany

Correspondence
Alireza Zamanian, Department of Reasoned AI Decisions, Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany.
Email: alireza.zamanian@iks.fraunhofer.de

Funding information
Bavarian Ministry for Economic Affairs, Regional Development and Energy

The pattern graph framework solves a wide range of missing data problems with nonignorable mechanisms. However, it faces two challenges of assessability and interpretability, particularly important in safety-critical problems such as clinical diagnosis: (i) How can one assess the validity of the framework's a priori assumption and make necessary adjustments to accommodate known information about the problem? (ii) How can one interpret the process of exponential tilting used for sensitivity analysis in the pattern graph framework and choose the tilt perturbations based on meaningful real-world quantities? In this paper, we introduce Informed Sensitivity Analysis, an extension of the pattern graph framework that enables us to incorporate substantive knowledge about the missingness mechanism into the pattern graph framework. Our extension allows us to examine the validity of assumptions underlying pattern graphs and interpret sensitivity analysis results in terms of realistic problem characteristics. We apply our method to a prevalent nonignorable missing data scenario in clinical research. We validate and compare our method's results of our method with a number of widely-used missing data methods, including Unweighted CCA, KNN Imputer, MICE, and MissForest. The validation is done using both boot-strapped simulated experiments as well as real-world clinical observations in the MIMIC-III public dataset.

KEYWORDS
interpretability, nonignorable missing data, pattern graph, safety critical, sensitivity analysis

1 | INTRODUCTION

Pattern graphs[1] provide an effective framework for missing data problems with nonignorable mechanisms. The framework graphically models information about how patterns of missingness evolve across the timeline of observations. It makes the key a priori assumption that the "conditional odds of a [missingness] pattern against its parents depend only on the observed entries".[1] Associated with the framework is a sensitivity analysis scheme based on the concept of exponential

tilting, where an exponential factor captures departures from the aforementioned assumption. Subsequently, sensitivity analysis is done by varying the exponential factor in a chosen range.

The pattern graph framework assumption covers missingness problems with ignorable mechanisms as well as a wide range of ones with nonignorable mechanisms. However, like many other counterparts, it faces two challenges of assessability and interpretability: (i) How can one assess the validity of its key assumption in a given problem, and how should the framework's result be adjusted if some aspects of the problem are known to violate the assumption? (ii) How can one interpret the process of exponential tilting in sensitivity analysis, that is, how does *one unit of tilt* correspond to meaningful real-world quantities? Does the tilting perturbation comply with the realistic characteristics of the problem? These challenges become particularly important in safety-critical areas such as clinical diagnosis, where an unnoticed departure from assumptions can lead to catastrophic results, and sensitivity analyses need to be calibrated accordingly.

In this paper, we introduce an extension for the pattern graph framework which enables us to incorporate substantive available knowledge about the missingness mechanism, encoded in m-graph causal models for missing data. Our extension, which we refer to as *Informed Sensitivity Analysis* (ISA), allows us to examine the validity of the pattern graph's prior assumption, and if necessary, correct for potential biases. It makes a connection between the sensitivity analysis parameters and parameters of an assumed missingness model. As these parameters often have a clear interpretation in terms of prior knowledge, we achieve interpretability of the results in addition to their assessability. As a prerequisite, we modify the pattern graph assumption to a new form, which we name *edge-wise identifiability assumption*.

Our study is inspired by the *clinical sequential observations* problem, a prevalent and potentially nonignorable missing data scenario in clinical research. Examples can be found in almost all diagnostic routines, where physicians collect additional evidence gradually, to conclude a final diagnosis for patients. For its unique nature, this problem admits pattern graph modeling, but at the same time, bears more information that can be leveraged to modify and adjust the framework's solution; this is the objective of ISA. In Sections 1.1 and 4, we further introduce and study the problem in detail.

The main contributions of this paper are as follows:

1. We develop a new and more effective identifiability assumption for pattern graphs. Accordingly, We propose a modification to the pattern graph framework (our ISA method) and demonstrate how both assessability and interpretability are achieved within the ISA extension.
2. By accounting for structured missingness mechanisms, we demonstrate that ISA enables pattern graphs to cover more nonignorable scenarios, ones that otherwise would violate the framework's assumption.
3. We demonstrate how a prevalent clinical scenario that cannot be solved correctly by conventional methods may be formulated and solved by ISA method.
4. We compare and evaluate our method against widely-used missing data methods, using both simulated data as well as real-world clinical data.
5. As a more theoretical contribution to missing data analysis's body of work, ISA is the result of the first attempt (to the best of our knowledge) to study the confluence of two missing data models, namely pattern graphs and m-graphs.

The rest of the paper is organized as follows. The motivating medical problem for this paper is introduced in Section 1.1. Section 2 provides a brief overview of related works. We introduce our approach of Informed Sensitivity Analysis in Section 3. The results of simulation and empirical studies are presented in Section 4. Concluding remarks are made in Section 5.

1.1 | The clinical sequential observations process

Bickley and Szilagyi[2] describe the process of clinical semiology as a sequential step-by-step process of observing signs and symptoms. During this process, the physician first performs *primary tests* such as blood pressure or heart rate measurements. Based on the findings, more specific *secondary tests* such as MRI scanning or genetic tests are performed. These tests are secondary due to their potential harm, availability, or cost. The process, which we call *clinical sequential observations**, is highly selective with regard to a pool of possible medical tests. It often continues for several steps and results in sets of observations that differ in number and type of features. As we show later in the paper, the resulting missingness problem may be nonignorable. A formulation of this scenario, as well as the underlying assumptions, are presented in Section 4.

Clinical sequential observations is a unique process from a missing data analysis point of view; it often bears information about both the mechanism and pattern of missingness that can be leveraged for inference. In particular, the knowledge of the order of observations is attributed to medical diagnostics flowcharts and standards of clinical semiology.[4-6] Within these standards, the final missingness pattern gives us enough information to recreate (probabilistically) the underlying observation process. Likewise, the process bears information regarding the reasons of acquisition of a particular variable. The primary reason for acquisition is physicians' clinical judgments, supported again by diagnostics flowcharts: Based on the normal and abnormal values of primary tests, physicians decide to observe (or skip) later variables. In addition, unplanned reasons such as waiting queues, patients' reluctance to perform specific tests, insurance coverage policies, or missing hospital visits can induce more missingness in the data, all of which are often recorded in electronic health records. All together, these are considered elements in the *mechanism* model of missingness.

As an example of a clinical sequential observation process, Figure 1 presents a subpart of the *flowchart for thalassemia carrier identification* for differential diagnoses (DDx) of a set of related diagnoses such as β-thalassemia (and other thalassemia variants), iron deficiency, and Hbs (associated with Sickle cell disease), based on the following lab tests: Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Hb pattern, iron level **(IL)** and a number of genetic tests.[4] The first three tests are included in the *Hematology* category of the blood test. Based on the readings, clinicians then become suspicious of iron deficiency as a root cause and, therefore, decide to order Zinc protoporphyrin (ZnPP) or/and IL tests which are items of the *Chemistry* category of the blood test. Finally, low values of IL (or high values of ZnPP) convince clinicians to diagnose iron deficiency. In general, hematology category tests are ordered more frequently, while ZnPP and IL tests are more specific and decisive, and considered as the main diagnostic test for iron deficiency. As a conclusion, one can consider the IL test as the secondary test, and the set of MCV/MCH/Hemoglobin A2 (HbA2) items as the primary.

This paper discusses why the pattern graph framework suits the clinical sequential observation process, why it is possible and necessary to adjust the framework, and how our proposed ISA method makes such adjustments. In Section 4, this flowchart will serve as a case study for ISA.
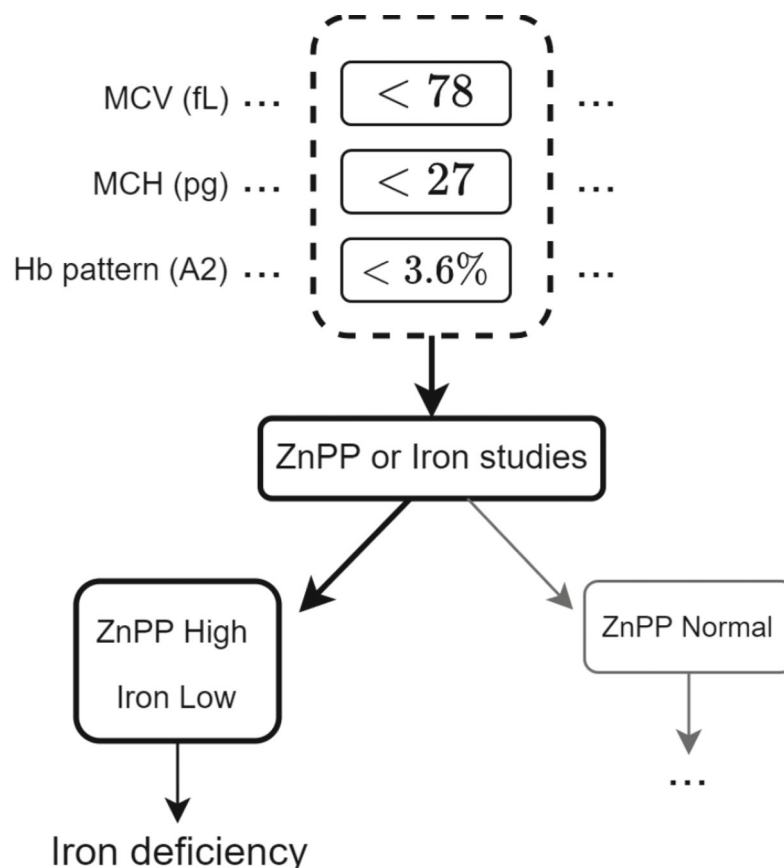

FIGURE 1   The iron deficiency branch in the thalassemia carrier DDx flowchart.[4]

2 | RELATED WORK

2.1 | Missingness problem formulation

Consider a $n \times d$ dataset comprised of n i.i.d. realizations of a d-dimensional random vector $L \in \mathbb{R}^d$ whose i-th component we denote by L_i. Define the missingness indicator $R \in \{0,1\}^d$ as a binary random vector such that L_i is missing if $R_i = 0$.[7] A realization of the missingness indicator, denoted by r, is called a missingness pattern, and the conditional probabilities of the patterns $p(r|l) \equiv p(R = r|L = l)$ describe the missingness mechanism. Given two patterns r and s, we write $s > r$ if the observed components under r are also observed under s, while there exists at least one component which is observed under s but missing under r. We denote the difference between two patterns by $s - r$. Finally, we use the notation $|r| = \sum_j r_j$ to denote the number of observed components in pattern r.

2.2 | Missingness mechanism assumptions

Missing data problems with nonignorable missingness mechanisms cannot be solved using methods that effectively ignore the mechanism. Therefore, further modeling assumptions are needed to proceed.[7] Examples of '*nonignorable models*' can be found in studies by Robins et al,[8] Robins,[9] Zhou et al[10] and Mohan et al.[11] In these works, assumptions are formulated as conditional independence statements about the variables and missingness indicators. Mohan et al,[11] Mohan and Pearl[12] and Shpitser et al[13] suggest that one can consider these assumptions as a factorization of a target distribution with respect to a causal directed acyclic graph (DAG). This way, identifiability in missing data problems is tackled as a causal identification problem. The associated causal DAG for the missing data problem is called m-graph. Nabi et al[14] further study the completeness of the identification methods for m-graphs.

In general, m-graphs model the mechanism of missing data and view the missingness patterns only as realizations of the model; in other words, under m-graphs, a missingness pattern (eg, $R = 1101$) is viewed as a realized set of missingness indicators, and is studied by factorization of $(R_1, \ldots, R_i)$ distribution with respect to the m-graph. Under this view, the order of the acquisitions is ignored. In Section 3, we study in details, the m-graph assumptions and properties which are related to the current paper.

2.3 | Pattern graphs

Chen[1] introduces the pattern graph framework for missing data problems. A pattern graph is a DAG whose nodes are the missingness patterns seen in the data, and whose directed edges represent "possible hidden scenarios that generate a response pattern".[1] Unlike m-graphs, pattern graphs directly model the emerging missingness patterns via the *transition probabilities* from one pattern to another. These probabilities are assigned to edges. Rather than adopting the m-graph's set-view on patterns, pattern graphs leverage the prior knowledge about the chronology of observations of the data components, for example, if a component has been missed/observed systematically after missing/observing another component.

Pattern graphs allow for an imputation algorithm, applied row-wise on incomplete data to impute the missing values, through the following steps (The related algorithms are presented in Appendix H).

(Step 1): A pattern graph is generated for a given dataset using the algorithm from Chen.[1]
(Step 2): The following identifiability assumption is made: The conditional extrapolation density given the observed data in a missingness pattern is identical to that in the parents of the pattern, that is,

$$\frac{p(l_{\bar{r}}|l_r, R = r)}{p\left(l_{\bar{r}}|l_r, R \in \mathrm{pa}_{\mathrm{PG}}(r)\right)} = 1, \tag{1}$$

where $\mathrm{pa}_{\mathrm{PG}}(r)$ denotes the set of parents of the pattern r in the pattern graph PG, while $l_{\bar{r}}$ and l_r denote the unobserved and observed components of the realization l of the complete random vector L under the pattern r, respectively.

(Step 3): Let r be the pattern of the to-be-imputed row. First, we select a parent $s \in \text{pa}_{\text{PG}}(r)$ based on the occurrence frequency of the parents in the data. Then we impute the components that are different between s and r, that is, we impute the values of L_{s-r}, by sampling from the distribution $p(l_{s-r}|l_r, R = s)$ (Algorithm 1). We repeat this step in the same row for the new pattern s, and recursively thereafter across the parent patterns until no missing value is left (Algorithm 2).

2.4 | Exponential tilting sensitivity analysis

Kim and Yu[15] consider unidentifiable missingness mechanisms as departures from an identified mechanism. They suggest to model this departure based on the concept of exponential tilting introduced by Scharfstein et al.[16] Specifically, they investigate the unidentifiable self-censoring non-response missingness, where in a data model (L_1, L_2), variable L_1 is fully observed and the response variable L_2 is subjected to missingness. Based on their formulation, the observed conditional distribution $p(l_2|l_1, R_2 = 1)$ deviates from the unobserved version $p(l_2|l_1, R_2 = 0)$ by a factor of

$$\frac{p(R_2 = 0|l_1, l_2)/p(R_2 = 1|l_1, l_2)}{p(R_2 = 0|l_1)/p(R_2 = 1|l_1)}. \tag{2}$$

In case of logistic modeling of the missingness indicator, (2) reduces to

$$\frac{1}{C(L_1)} \exp(-\phi L_2), \tag{3}$$

where ϕ parameterizes the effect of self-censoring, and $C(L_1)$ is independent of L_2. The expression in (3) is interpreted as an exponential tilt of the extrapolation density $p(l_2|l_1, R_2 = 0)$ from the observed distribution $p(l_2|l_1, R_2 = 1)$. Franks et al[17] demonstrate that, such logistic modeling results in an interpretable connection between the substantive knowledge about the missingness mechanism and estimation of a parameter of interest. Carpenter et al[18] introduce a re-weighted multiple-imputation (MI) method to solve a bivariate self-censoring missingness problem. In their work, they use the exponential tilt as the importance sampling's ratio in sampling step of MI. The authors suggest that one can verify the value of tilting parameters by domain experts.

Chen[1] implements the concept of exponential tilting as described in (3) and develops a sensitivity analysis scheme for pattern graphs. In this scheme, an exponential factor models departures from the assumption made in Equation (1) in the following form:

$$\frac{p(l_{\bar{r}}|l_r, R = r)}{p(l_{\bar{r}}|l_r, R \in \text{pa}_{\text{PG}}(r))} = \exp(\omega^\top l_{\bar{r}}), \tag{4}$$

where ω is the tilting parameter vector of the size $|\bar{r}|$ for each pattern r and controls the amount of deviation. It is used to correct the imputation sampling procedure via rejection sampling with the probability of

$$\frac{\exp(\omega^\top l_{\bar{r}}^\dagger)}{\exp(\omega^\top \mathcal{U})}, \tag{5}$$

where $l_{\bar{r}}^\dagger$ is a candidate sampled imputation, $\mathcal{U}$ is an upper bound on $L_{\bar{r}}$, and the exponential coefficient ω parameterizes the tilt. The rejection step is applied at the end of Algorithm 2) to correct the sampling distribution. A range of variation of $[-1, +1]^{|\bar{r}|}$ for all ω is suggested by Chen.[1] The modified algorithm for sensitivity analysis which replaces Algorithm 2 is presented by Algorithm 3. In this paper, we call Algorithm 3, the *default* sensitivity analysis approach.

3 | METHODS

The validity of the pattern graph's results depends on the validity of its assumption. Also, the interpretability and practicality of the framework's sensitivity analysis rely on the possibility of providing a real-world interpretation of the tilting parameters. We claim that pattern graphs address both if we incorporate available prior (causal) knowledge about the missing data problem into the framework. Hence, we introduce the *Informed Sensitivity Analysis* (ISA) method to achieve this goal. In ISA, we utilize the causal information of m-graphs and missingness models to identify possible structural
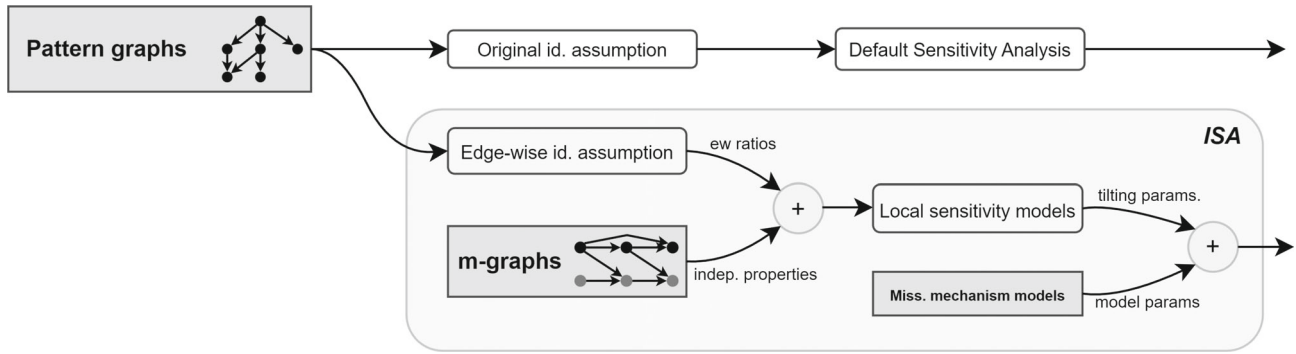
FIGURE 2   The relation between pattern graph framework, m-graphs, missingness mechanism models, and ISA method: Applying the causal information encoded in the m-graphs and the edge-wise assumption ratios, results in local sensitivity models. The tilting parameters of these models will be connected to explainable missingness model parameters to allow for an informed sensitivity analysis. ew, edge-wise; id, identifiability; indep, independence; params, parameters.

departures from the pattern graph assumption. The ISA establishes a connection between the tilting parameters and the parameters of an assumed missingness model (e.g., logistic model). Finally, by making informed guesses about the parameters of the model, we perform sensitivity analysis where perturbations are translated into meaningful real-world quantities. ISA utilizes a new modified identifiability assumption, which we call *edge-wise identifiability assumption*. In this paper, we refer to the assumption introduced by Chen[1] as the *original* assumption. The relations among pattern graphs, m-graphs and ISA are depicted in Figure 2.

   In this section, we start by introducing the edge-wise assumption and show how it is an effective assumption for pattern graphs. We then introduce ISA's modified imputation and sensitivity analysis algorithms and discuss how assessability and interpretability are achieved.

3.1 | Notations

Let $G = (\mathbf{R}, \mathbf{E})$ represent the pattern graph for a missing data problem. For a given missingness pattern ($r \in \mathbf{R}, r \neq \mathbf{1}_d$), the pattern graph imputation Algorithm 1 selects an *imputation path* $\pi \in \Pi_G$ by sampling iteratively on different depths from the ancestors of r. We denote the nodes lying on the imputation path by $r^{(k)}$, where the first node $r^{(0)}$ is the current pattern, $r^{(1)}$ is the selected parent of the pattern, and last node $r^{(K)}$ is always the fully-observed pattern, that is, $r^{(K)} = \mathbf{1}_d$. We denote the edge from $r^{(k)}$ to $r^{(k-1)}$ by $e^{(k)} \in \mathbf{E}$. We index the components that are imputed during moving across the edge $e^{(k)}$, namely the components in $L_{r^{(k)} - r^{(k-1)}}$, by the subscript Δ_k. Subsequently, all components in R_{Δ_k} are 0 for the child pattern, and 1 for the parent pattern, that is, $R_{\Delta_k}^{(k-1)} = \mathbf{0}_{|r^{(k)} - r^{(k-1)}|}$ and $R_{\Delta_k}^{(k)} = \mathbf{1}_{|r^{(k)} - r^{(k-1)}|}$. For ease of notation, we write $R_{\Delta_k} = \vec{0}$ and $R_{\Delta_k} = \vec{1}$ to refer to these two values. Finally, subscript $-\Delta_k$ selects the remaining components other than Δ_k. For every edge $e^{(k)}$, the missingness vector $R_{-\Delta_k}$ has a constant value of

$$R_{-\Delta_k} = (\underbrace{0, \ldots, 0}_{|\overline{r}^{(k)}|}, \underbrace{1, \ldots, 1}_{|r^{(k-1)}|}).$$

For simplicity, we refer to this value as $r_{-\Delta_k}$. Figure 3 illustrates our notations for an arbitrary imputation path (a) as well as an example of a path in a pattern graph (b). The values for the selected imputation path are:

$$\text{nodes (patterns):} \quad r^{(0)} = (1000), \quad r^{(1)} = (1110), \quad r^{(2)} = (1111), \tag{6a}$$

$$\text{edges:} \quad e^{(1)} = (1110) \rightarrow (1000), \quad e^{(2)} = (1111) \rightarrow (1110), \tag{6b}$$

$$\text{pattern-indexed components:} \quad L_{r^{(0)}} = (L_1), \quad L_{r^{(1)}} = (L_1, L_2, L_3), \quad L_{r^{(2)}} = (L_1, L_2, L_3, L_4), \tag{6c}$$

$$\Delta_k \text{ components:} \quad L_{\Delta_1} = (L_2, L_3), \quad L_{\Delta_2} = (L_4). \tag{6d}$$

(1111)

(1101) (1110) (0111)

$r^{(0)} \xleftarrow{e^{(1)}} r^{(1)} \xleftarrow{\cdots\cdots} r^{(K-1)} \xleftarrow{e^{(K)}} r^{(K)}$ (1000) (0000)
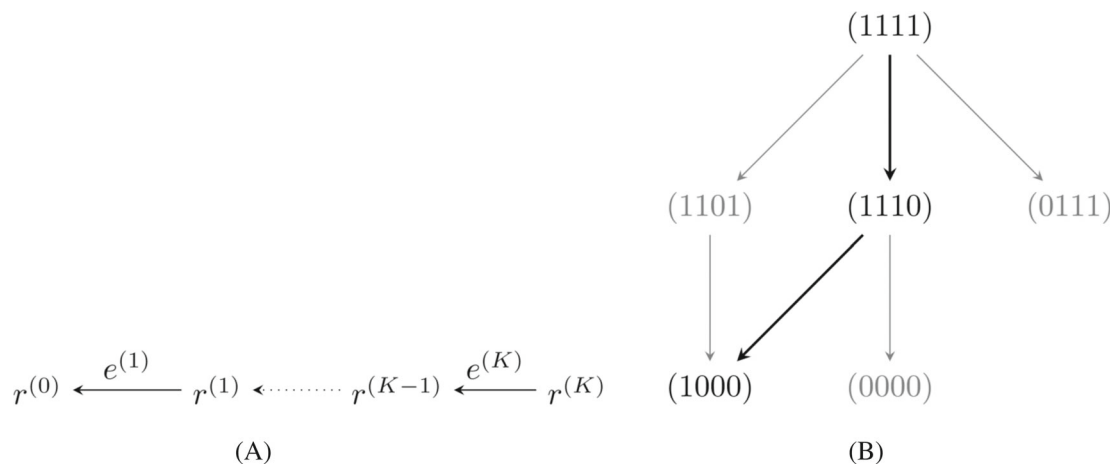
(A)   (B)

FIGURE 3   Illustrating examples for the introduced notation.

3.2 | Edge-wise identifiability assumption and sensitivity analysis

M-graph models describe the causal relationships among individual components. In addition, the imputation Algorithm 2 recursively imputes the components in $L_{\bar{r}}$ along an imputation path. Therefore, in order to leverage structure of the m-graph model in the imputation algorithm, we propose breaking the identifiability assumption into edge-wise statements about components that are imputed on each edge of the pattern graph. This way, it becomes more amenable to simplification on the basis of conditional independence relations holding in the m-graph model. Specifically, we consider the *edge-wise identifiability assumption*, which requires that for all edges $e^{(k)}$ lying on a selected imputation path $\pi \in \Pi_G$, and for every imputation path,

$$\prod_{i=1}^{k} \frac{p\left(l_{\Delta_k} | l_{r^{(k-1)}}, r^{(i-1)}\right)}{p\left(l_{\Delta_k} | l_{r^{(k-1)}}, r^{(i)}\right)} = 1. \tag{7}$$

We shall emphasize that Equation (7) is stated for an edge, conditioning on the imputation path being selected. This means that for an individual edge in a pattern graph, the associated edge-wise assumption differs when the edge belongs to different imputation paths. An example of such an edge is $1111 \rightarrow 1110$ in Figure 3 (right), which belongs to two paths $1111 \rightarrow 1110 \rightarrow 1000$ and $1111 \rightarrow 1110 \rightarrow 0000$. Given that the subsequent discussions in this paper primarily focus on conditioning on an imputation path being selected, we will omit the use of selected path index (π) in the equations for the sake of brevity. In case of analyzing multiple imputation paths simultaneously, we will employ appropriate clarifying path indices for precision.

While the original pattern graph assumption is made about the entire unobserved components of a sample, Equation (7) makes assumptions separately about components that are imputed across a single edge. In other words, the former assumes that the imputed data in one complete round of Algorithm 2 ($L_{\bar{r}}$) is sampled from the unbiased extrapolation density, while the latter breaks this assumption down to each turn of partial-imputation (L_{Δ_k}). Nevertheless, the edge-wise assumption is as strong for the pattern graph as the original assumption, as each ratio in Equation (7) appears in the original assumption, although corresponding to a different pattern. In other words, edge-wise assumption is a reordering and regrouping version of the original assumption (with this implicit consideration that it must structurally hold, as the equality must be true for two different ordering and grouping of a set of terms). However, as we will show later in Subsection 3.3, edge-wise assumption proves beneficial in constructing the ISA method. In Subsection 3.4 We will discuss the relation between two assumptions as well as the necessity of the edge-wise assumption in detail. The validity of edge-wise assumption is discussed in Theorem 1.

> **Theorem 1** (Identifiability under edge-wise assumption). *Let $G = (\mathbf{R}, \mathbf{E})$ be a pattern graph for a missing data problem modeled by (L, R), and the imputation Algorithm 1 imputes the partially observed data. Under the edge-wise assumption, sampling extrapolation densities in Algorithm 2 are correct.*

Proof of Theorem 1 is provided in Appendix A.1.

The ratio terms in Equation (7) are assumptions about the extrapolation densities. We can replace each ratio with an equivalent assumption about the missingness mechanism.

Corollary 1. _Let $G = (\mathbf{R}, \mathbf{E})$ be a pattern graph for a missing data problem modeled by (L, R). For every path and edge selection ($e \in \pi$, $\pi \in \Pi_G$), the edge-wise assumption holds iff_

$$\prod_{h \in H} \frac{p\left(l_{\Delta_k} | l_{r^{(k-1)}}, r^{(h-1)}\right)}{p\left(l_{\Delta_k} | l_{r^{(k-1)}}, r^{(h)}\right)} \prod_{j \in J} \frac{p\left(R_{\Delta_j} = \vec{0} | l_{r^{(k)}}, r_{-\Delta_j}\right)}{p\left(R_{\Delta_j} = \vec{1} | l_{r^{(k)}}, r_{-\Delta_j}\right)} = C, \tag{8}$$

for any H, J s.t. $H \cup J = \{1, \dots, k\}$ and $H \cap J = \emptyset$, where C is an arbitrary constant independent of L_{Δ_k}.

For simplicity, we call the ratio terms in first and second product terms in Equation (8), the _L-sorted_ and _R-sorted_ ratios respectively. Corollary 1 states that we can replace any chosen subset of the ratios by its _R_-sorted counterpart. Note that the validity of Equation (8) still depends on the imputed components across a single edge, that is, L_{Δ_k}, which is included in $L_r^{(k)}$. Since Equation (7) is a special case of Equation (8) with $J = \emptyset$, for the rest of the paper we refer to Equation (8) as the edge-wise assumption.

Associated to the edge-wise assumption, we now define the edge-wise sensitivity analysis by perturbing Equation (8) using the exponential tilting concept, as

$$\prod_{h \in H} \frac{p\left(l_{\Delta_k} | l_{r^{(k-1)}}, r^{(h-1)}\right)}{p\left(l_{\Delta_k} | l_{r^{(k-1)}}, r^{(h)}\right)} \prod_{j \in J} \frac{p\left(R_{\Delta_j} = \vec{0} | l_{r^{(k)}}, r_{-\Delta_j}\right)}{p\left(R_{\Delta_j} = \vec{1} | l_{r^{(k)}}, r_{-\Delta_j}\right)} = C \exp(\omega_k^\top l_{\Delta_k}), \tag{9}$$

where ω_k is the tilting parameter with the size of $|\Delta_k|$ for the edge $e^{(k)}$ on a selected imputation path. Three illustrating examples for Theorem 1, Corollary 1 and Equation (9) are presented in Appendix B.

As Equation (8) makes edge-wise assumptions, we perform the rejection sampling for each partial-imputation step. Algorithm 4 presents the modified edge-wise sensitivity analysis Algorithm. The associated rejection probability for Algorithm 4 is obtained as

$$\frac{\exp\left(\omega_k^\top l_{\Delta_k}^\dagger\right)}{\exp\left(\omega_k^\top \mathcal{U}_{\Delta_k}\right)}. \tag{10}$$

Next, we discuss how the edge-wise perspective of the new sensitivity analysis helps us formalize a way to incorporate substantive knowledge in the framework.

3.3 | Informing the edge-wise sensitivity analysis: ISA

As shown above, the edge-wise assumption in Equation (8) comprises (_R_-sorted) terms which relate to the missingness mechanism of study. In the absence of further information, Equation (8) must be assumed for inference. However, if mechanism information is available for a problem, some _R_-sorted terms can be evaluated. Consequently, the evaluated terms will give us an evidence for an informed choice of tilting parameter in Equation (9). As the extreme case, the terms will be canceled out, and hence do not appear in the sensitivity analysis step. Such a practice is carried out in our proposed method of informed sensitivity analysis (ISA). Upon such incorporation of missingness mechanism information, the global sensitivity model of the pattern graph (what we named the original sensitivity analysis) converts to a set of local sensitivity models, each of which is utilized for imputation of the single component.

The ISA method is built on the following setup: (i) Assume that besides the knowledge about missingness patterns, there exists prior knowledge about causes and the corresponding causal effects of the missingness indicators. This knowledge is encoded via m-graphs and the parameters of the missingness model (e.g., logistic model). (ii) In ISA, we simplify the left-hand side of Equation (9) using the conditional independence statements of m-graph, and then (iii) bring the remaining terms to an exponential form which consists of the parameters of the missingness model. This way, we translate

prior knowledge directly to the range of variation for ω in the sensitivity analysis step. In summary, we instruct the ISA method in the following steps:

1. *(prerequisite)* Construct the pattern graph and m-graph for a given problem;
2. **Incorporating m-graph information:** Extract the sufficient conditional independence statements from the m-graph to simplify the edge-wise exponential tilting terms;
3. **Decomposing global sensitivity model into local sensitivity models:** Obtain the local tilting parameters as a function of coefficients of the assumed missingness indicators' model for each imputation component;
4. **Choice of local tilting parameters:** Select an informed range of variation for local tilting parameters ω based on the prior knowledge about the aforementioned coefficients.

The structure of ISA is illustrated in Figure 2. We continue through the rest of the section by discussing details of these steps.

3.3.1 | Step 1: Incorporating m-graph information

As introduced in Subsection 2.2, m-graphs are causal DAGs for modeling the missing data problem. For a data distribution (L, R), an m-graph is defined as $G_{\text{m-graph}} = (\mathbf{V}, \mathbf{F})$ where $\mathbf{V} = L \cup R$ and $f \in \mathbf{F}$. An edge $f = (V_i, V_j)$ represents a causal effect from node V_i to V_j. The edges received by R nodes (missingness indicator) determine the missingness mechanism in a problem. As an example, an m-graph with Rs without parents (within $\mathbf{V}$) represents the ignorable *Missing-Completely-At-Random* (MCAR) mechanism. In addition to the general causal DAG assumptions, m-graphs often follow the *no-direct-effect* (NDE) assumption, which states that testing/measurement has no direct and unmediated effect on the study variables.[19] This implies "no $R \rightarrow L$ edge" in the m-graphs. Through the rest of the paper, we make the NDE assumption for all inferences.

In m-graphs, conditional independence relations can be discovered via the rules of d-separation. Obviously, different m-graphs encompass different conditional independence properties. In ISA, we seek such properties related to R nodes to reduce their corresponding R-sorted ratios in Equation (9) up to a constant (independent of L_{Δ_k}) or into conditional odds of individual R variables, as these could be expressed by meaningful and interpretable parameters. According to Corollary 1, we freely split the k index set for $e^{(k)}$ into the H and J subsets, that is, rewrite necessary ratio terms in their R-sorted form, such that applying the independence statements becomes possible.

As m-graph and pattern graph frameworks model two distinct types of information, scenarios with distinct m-graphs may share a single pattern graph. This means that the extracted conditional independence statements, and hence the simplification of ratio terms in the pattern graph assumptions, may differ case by case. Below, we discuss a sufficient condition for m-graphs which leads to simplification of R-sorted ratios. In the absence of this condition, we invoke a second sufficient condition which leads to a useful factorization for the ratios, such that incorporation of available knowledge becomes possible.

Condition 1. The following conditional independences hold:

$$R_{\Delta_j} \perp\!\!\!\perp L_{\Delta_k} | L_{r^{(k-1)}}, R_{-\Delta_j}. \tag{11a}$$

In other words, Condition 1 requires that $\{L_{r^{(k-1)}}, R_{-\Delta_j}\}$ is a d-separation set for R_{Δ_j} and L_{Δ_k} in the associated m-graph. Given Condition 1, the R-sorted ratio of the Δ_j index, reduces to a constant factor in the edge-wise sensitivity analysis (9) (proof in Appendix A.3).

An m-graph which follows three restrictions, satisfies Condition 1[†]: *(i)* no path $L_{\Delta_k} \rightarrow R_{\Delta_j}$ which is not mediated by the d-separation set (e.g., direct edge) should exist; *(ii)* No confounder path of the form $L_{\Delta_k} \leftarrow L_{\mathbf{1}_d - r^{(k)}} \rightarrow R_{\Delta_j}$ which is not mediated by the d-separation set should exist. Recall, that the subscript $\mathbf{1}_d - r^{(k)}$ selects all variables that remain missing after the partial imputation along the edge $e^{(k)}$; *(iii)* No collider path of the form R_{Δ_j} and $L_{\Delta_k} \rightarrow R_{-\Delta_j} \leftarrow R_{\Delta_j}$ should exist. Figure 4 depicts an imputation path and three example m-graphs structures, demonstrating when the restrictions are met or violated. Details of derivation of the restrictions are presented in Appendix A.3.

Among possible missingness mechanisms which conform to this condition are the ignorable MCAR and MAR missingness mechanisms. For a detailed discussion about these mechanisms and their corresponding m-graphs, see Appendix C.
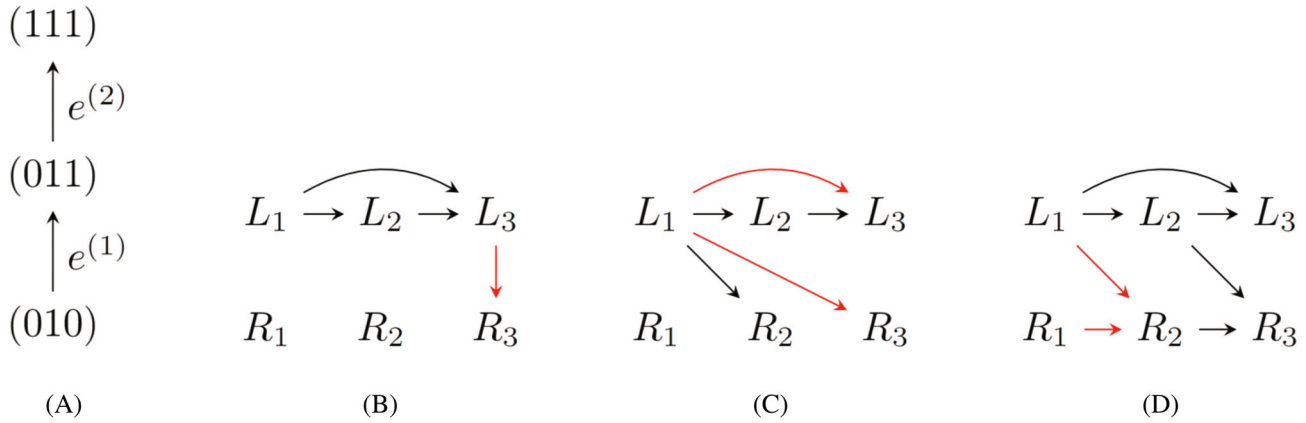
(111)

$\uparrow e^{(2)}$

(011)

$\uparrow e^{(1)}$

(010)

	$L_1 \to L_2 \to L_3$	$L_1 \to L_2 \to L_3$	$L_1 \to L_2 \to L_3$
	$\quad\quad\quad\downarrow$	$\quad\searrow\quad\searrow$	$\quad\searrow\quad\searrow$
$R_1 \quad R_2 \quad R_3$	$R_1 \quad R_2 \quad R_3$	$R_1 \quad R_2 \quad R_3$	$R_1 \to R_2 \to R_3$

(A) (B) (C) (D)

FIGURE 4 Three possible m-graphs (b, c, d) for a pattern graph and an imputation path in (a) to study the restrictions of Condition 1: Red edges highlight the violating paths: For partial imputation across $e^{(1)}$, in (b) the direct edge $L_3 \to R_3$ violates restriction (*i*), and in (c) the confounding $\{L_1\}$ belongs to $L_{1_{d-r^{(1)}}}$, hence violates restriction (*ii*). for partial imputation across $e^{(2)}$ and $j = 2$, in (d) the node R_2 is a collider for L_1 and R_1, hence violates restriction (*iii*), whereas for $j = 1$, the set $\{L_2, R_2\}$ *d*-separates L_1 and R_3. In all three examples, Condition 1 holds if the red 'spurious' paths are interrupted, i.e, one removes: $L_3 \to R_3$ from (b), $L_1 \to R_3$ or $L_1 \to L_3$ from (c), and $R_1 \to R_2$ or $L_1 \to R_2$ from (d).

In the absence of Condition 1, we can alternatively invoke the following

Condition 2. R_{Δ_j} consists of only one component ($|\Delta_j| = 1$), or we have the conditional independences

$$\forall \delta_m \in \Delta_j : \quad R_{\delta_m} \perp\!\!\!\perp R_{\Delta_j \setminus \delta_m} | L_{r^{(k)}}, R_{-\Delta_j}, \tag{11b}$$

where δ_m enumerates the individual components of R_{Δ_j} in an arbitrary order.

To explain Condition 2 further, we define the *conditional odds* as follows.

Definition 1 (Conditional odds). Given a set of variables Q and values q, the conditional odds for R given $Q = q$ is defined as

$$O(R|q) = \frac{p(R = \vec{0}|q)}{p(R = \vec{1}|q)}.$$

We will show in the next step of ISA that the conditional odds for *individual R* components (i.e., $|R| = 1$ in Definition 1) may be conveniently connected to explainable and meaningful parameters in the sensitivity analysis. Thus, it is desirable if the *R*-sorted ratios which have escaped Condition 1, can be factorized into conditional odds for individual *R* components. Condition 2 describes the sufficient conditions for such a scenario:

First part of Condition 2 ($|\Delta_j = 1|$) trivially gives the desired results. To explain the second part (when $|\Delta_j| > 1$), we factorize the *R*-sorted ratio following the chain rule as

$$\frac{p\left(R_{\Delta_j} = \vec{0}|l_{r^{(k)}}, r_{-\Delta_j}\right)}{p\left(R_{\Delta_j} = \vec{1}|l_{r^{(k)}}, r_{-\Delta_j}\right)} = \prod_{m=1}^{|\Delta_j|} \frac{p\left(R_{\delta_m} = 0|l_{r^{(k)}}, r_{-\Delta_j}, R_{\delta_1 : \delta_{m-1}} = 0\right)}{p\left(R_{\delta_m} = 1|l_{r^{(k)}}, r_{-\Delta_j}, R_{\delta_1 : \delta_{m-1}} = 1\right)}, \tag{11c}$$

with a slight abuse of notation for the case $m = 1$, as it doesn't have the $R_{\delta_1 : \delta_{m-1}}$ term in the conditioning set. Without further assumptions, the crude factorization in (11c) cannot be simplified further to individual conditional odds terms. An effective assumption for simplification is stated by the second part of Condition 2 and the conditional independences (11b), namely, if the missingness indicators in the R_{Δ_j} *block* are independent given $\{l_{r^{(k)}}, r_{-\Delta_j}\}$. This lets us drop $R_{\delta_1 : \delta_{m-1}}$ from the conditioning set of the right-hand side of (11c) and obtain

$$= \prod_{m=1}^{|\Delta_j|} O\left(R_{\delta_m}|l_{r^{(k)}}, r_{-\Delta_j}\right), \tag{11d}$$
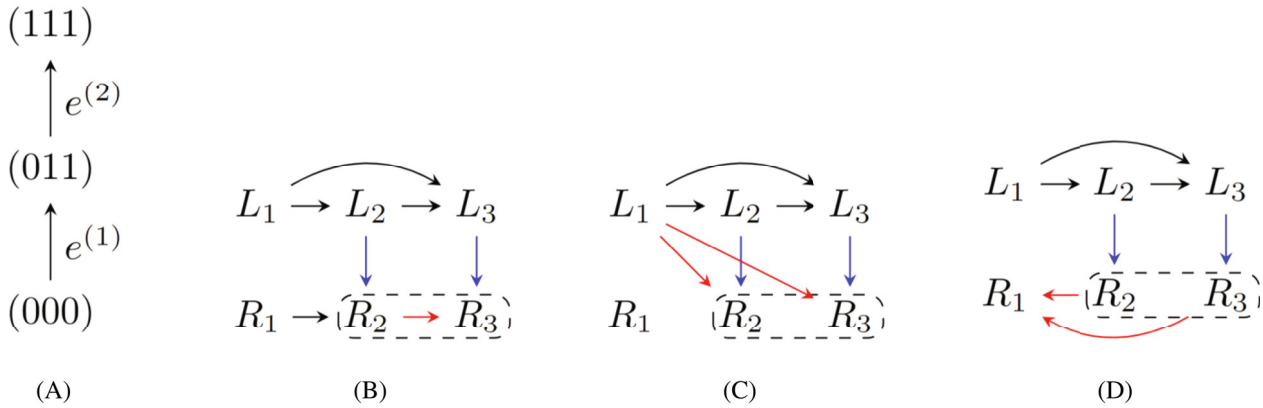
FIGURE 5 Three possible m-graphs (b, c, d) which violate Condition 2 for a pattern graph and an imputation path in (a). The blue edges highlight the edges which violate Condition 1 (thus leads us to check Condition 2), while the red edges highlight the violating paths: For partial imputation across $e^{(1)}$, we have $R_{\Delta_1} = \{R_2, R_3\}$ (dashed bounding box). In (b) the direct edge $R_2 \rightarrow R_3$ violates the restriction (i); in (c) the confounder $\{L_1\}$ belongs to $L_{1_{d-r^{(1)}}}$, hence violates restriction (ii); In (d) the node R_1 is a collider for R_2 and R_3, hence violates restriction (iii). In all three examples, Condition 2 holds if the red spurious paths are interrupted, that is, one removes: $R_2 \rightarrow R_3$ from (b), $L_1 \rightarrow R_2$ or $L_1 \rightarrow R_3$ from (c), and $R_2 \rightarrow R_1$ or $R_3 \rightarrow R_1$ from (d).

for which the discussed case of $|\Delta_j| = 1$ is a special case, where only one conditional odds term exists. Step 2 of ISA will demonstrate how conditional odds terms in (11d) are employed further in the ISA procedure.

An m-graph which follows three restrictions, satisfies the second part of Condition 2, namely the conditional independences in (11b): for all $\delta_m, \delta_{m'} \in \Delta_j$, (i) no path $R_{\delta_m} \rightarrow R_{\delta_{m'}}$ which is not mediated by the d-separation set should exist; (ii) No confounder path of the form $R_{\delta_m} \leftarrow L_{1_{d-r^{(k)}}} \rightarrow R_{\delta_{m'}}$ which is not mediated by the d-separation set should exist; (iii) No collider path of the form $R_{\delta_m} \rightarrow R_{-\Delta_j} \leftarrow R_{\delta_{m'}}$ should exist (the proofs follow an identical path to those of Condition 1, presented in Appendix A.3). Figure 5 depicts an imputation path and three example m-graphs structures, demonstrating when these restrictions are met or violated. It is worthy of note that these simplifying assumptions concern the edges of an imputation path. It is possible, and in fact highly likely to have a mixture of simplifying cases applied to different edges and imputation paths.

In conclusion of step 1 of ISA, we establish an important realization: in the absence of Conditions 1 and 2, ISA does not provide a viable alternative to the default sensitivity analysis within the scope of this paper. In Section 5, we express our anticipation of the discovery of additional conditions in future research.

3.3.2 | Step 2: Decomposing global sensitivity model into local sensitivity models

In this step, we consider a working model according to our prior knowledge of the missing data problem, to bring the remaining ratios as in (11d) to an exponential form. This way, the parameters of the working model have a direct connection to the tilting parameters ω_k in the right-hand side of (9). In other words, we use the substantive prior knowledge embedded in the working model, in the sensitivity analysis step.

Suppose the following logistic model for a missingness indicator R_i:

$$O(R_{\Delta_j}|l_{r^{(k)}}, r_{-\Delta_j}) = \exp\left(\beta_{jk}^\top l_{\Delta_k} + f(l_{r^{(k-1)}}, r_{-\Delta_j})\right), \quad (12)$$

where β_{jk} is the logistic coefficient for Δ_k (to-be-imputed) components in the R_i model, and $f(.)$ is an arbitrary model over the rest of remaining factors. In (12), the exponential terms are grouped into two groups of dependent and independent factors of L_{Δ_k}. We treat the second group of terms as a constant, and rewrite Equation (12) as

$$O(R_{\Delta_j}|l_{r^{(k)}}, r_{-\Delta_j}) = C_{jk}(\pi) \exp(\beta_{jk}^\top l_{\Delta_k}), \quad (13)$$

where $C_{jk}(\pi)$ are the obtained constants for the j-th odds in J for $e^{(k)}$, conditioned on the path π being selected. The resulting exponential form in (13) matches the assumed exponential tilt in Equation (9). This means that the coefficients β in the assumed logistic model per each R-sorted ratio, is connected to the tilting parameters ω.

After the first step of ISA, for all the irreducible R-sorted ratios which are expressed in the form of (11d), and given the exponential model in (13), we rewrite the terms as

$$\prod_i O\left(R_i | l_{\Delta_k}, \ldots \right) = C \exp\left(\sum_i \beta_{ik}^\top \, l_{\Delta_k}\right). \tag{14}$$

where $C = \Pi_i C_{ik}(\pi)$, while i enumerates over all the conditional odds terms. By accepting a slight abuse of notation, we drop the k subscript and π argument from C to highlight the connection of Equations (14) to (9); comparing right-hand sides of Equations (14) and (9) gives

$$\omega_k = \sum_i \beta_{ik}. \tag{15}$$

Equation (15) connects the parameters of an assumed working model for missingness indicators to the tilting parameters in the pattern graphs' sensitivity models. By associating the logistic coefficients to substantive and interpretable prior knowledge, we have the possibility of leveraging prior knowledge in the sensitivity analysis step. This will be discussed in the next ISA step.

3.3.3 | Step 3: Informed choice of local tilting parameters

Let R_u be the missingness indicator of a single component L_u, and $R_u \in R_{\Delta_j}$. Consider the component $L_v \in L_{\Delta_k}$ is one of the components to impute across the edge $e^{(k)}$. Assuming a logistic model for R_u, we have a logistic coefficient β_{uv} that appears in the exponential tilting sensitivity analysis for imputing L_{Δ_k}, that is, in (14). Let l_v and l'_v be two arbitrary values for L_v. We evaluate the conditional odds of R_u at these two values (13):

$$O(R_u|.) = A \, \exp\left(\beta_{uv} l_v\right),$$
$$O(R_u|.) = A \, \exp\left(\beta_{uv} l'_v\right),$$

where A is a term including the covariates excluding L_j. Solving for β_{uv}, we have

$$\beta_{uv} = \frac{1}{(l_v - l'_v)} \log \mathrm{OR}(R_u|l_v \, : \, l'_v), \tag{16}$$

where $\mathrm{OR}(R_u|l_v : l'_v) := O(R_u|l_v)/O(R_u|l'_v)$ is the odds ratio of R_u between two levels of L_v. We can translate a piece of prior knowledge to an informed guess for β_{uv} if it resembles this odds ratio. Two possible cases from clinical studies are presented below:

- If L_v is a binary variable representing the patients' health status (1=healthy, 0=sick), and we know that "*the odds of a missing observation for healthy patients is approximately λ times greater than the respective odds for sick patients*", a choice of $\beta \approx \log \lambda$ would reflect the prior knowledge/assumption.

- Concerning the test protocols and diagnostic flowcharts, likewise, we might know from the clinical protocols that *the odds of ordering test u after the occurrence of a symptom v is approximately λ times greater than if symptom v is absent*, which yields $\beta \approx -\log \lambda$.

The λ values associated to the odds ratio in Equation (16) give initial guesses for β and subsequently ω parameters. We denote this initial guess for both β and ω by a superscript asterisk, that is, β^*, ω^*. Since λ is often a guess and cannot be inferred nor can be validated via the observations, we select a range by adding and subtracting an offset value to and from the initial guess. We define the offset value ϵ such that $[\beta^* - \epsilon, \beta^* + \epsilon]$ is considered as the range for β in the process of sensitivity analysis.

In case of the *default* sensitivity analysis, the range of $[-1, +1]$ (one unit of tilt toward both directions) is chosen in the global sensitivity model of the pattern graph, thus the ω parameters of all edges are set to a common value.[1] However, using the ISA method, we obtain different ranges for local sensitivity model. Therefore, we need a procedure to perform sensitivity analysis on different parameter ranges. For that purpose, let $\beta_1, \ldots, \beta_i$ be the coefficients used in ISA, with the corresponding ranges $\mathcal{B}_i := [\beta_i^* - \epsilon_i, \beta_i^* + \epsilon_i]$, $i = 1, \ldots, e$. One option is then to perform calculations on a grid defined

over the space $\mathcal{B}_1 \times \cdots \times \mathcal{B}_e$, and report the corresponding varying range of the parameter of interest. Another option is to vary the tilting parameters simultaneously from the minimum to the maximum values of their range and in equal relative steps. This is possible using a reference tilting parameter $\phi \in [-1, +1]$ which simultaneously controls the variation on all β_i parameters, via

$$\beta_i = \epsilon_i \phi + \beta_i^*. \tag{17}$$

Using the definition in (17), we have a way to compare the estimates from the pattern graph approach under default sensitivity analysis with our ISA method, as we set the tilting parameter of default method ω to be equal to the reference parameter ϕ. This idea will be used in the comparisons in Section 4.

We established step 3 of ISA on the assumption that the missingness indicators follow a logistic model. However, the validity of ISA method is not limited to a specific working model. For any assumed missingness mechanism model that comprises interpretable parameters, and fits the form of rejection sampling step of the imputation algorithm, step 3 of ISA can be re-defined and repeated. In this paper, we continue with only the logistic models as it is also considered an effective working model for most missing data scenarios.[18] Also, note that the parameters of the working model may still be uninterpretable, not because of a lack of prior knowledge, but because of the structure of the assumed m-graph. As an example, the R-sorted ratios for $O(R_3|L_3, \ldots)$ and $O(R_1|L_1, \ldots)$ in Figure 4c,d and Figure 5b,c,d respectively, cannot be canceled out and thus appear in the local sensitivity models, while the components are not causally connected (unlike in Figure 4b). Forming an informative choice of tilting parameter for such cases is not straightforward. With this regard, further extensions of ISA methodology are needed. This can extend the list of sufficient conditions to apply ISA to a missing data problem.

3.4 | Relation between the original and edge-wise assumptions

In the opening of the current section, we claimed that developing the edge-wise assumption is necessary for the introduction of ISA method. We break the discussion about this claim in three questions: (i) whether edge-wise assumption makes a stronger or weaker assumption about pattern graphs than the original assumption; (ii) how it facilitates assessing the identifiability assumption, and (iii) how it facilitates the interpretability of the sensitivity analysis results. In this subsection, we discuss these three questions.

Regarding the first question, Theorem 2 states that edge-wise assumption is as strong as the original identifiability assumption for pattern graphs.

Theorem 2 (Equivalency of original and edge-wise identifiability assumptions). *Let $G = (\mathbf{R}, \mathbf{E})$ be a pattern graph for a missing data problem modeled by (L, R). The identifiability assumption in (1) holds iff for every path, and for every edge on the path the edge-wise assumption in (7) holds.*

In order to discuss the second question, we recall the assessability process in ISA. In step 1 of ISA, we arrive at L-sorted or R-sorted ratios which we simplify and cancel out by incorporating the m-graph conditional independence statements. This implies the need for having such conditional odds terms in order to be able to assess the assumptions. In the original pattern graph assumption, we have the $R \in \text{pa}_{\text{PG}}(r)$ term in the denominator of the ratio. This means that at least one missingness indicator in the denominator is marginalized over and canceled out of the condition set, while it is still present in the conditioning set of the numerator. As an example, take a case where $r = (100)$ and $\text{pa}_{\text{PG}}(r) = \{(110), (101)\}$. Then, in the left-hand side of Equation (1) we have

$$\frac{p(l_2, l_3|l_1, R_1 = 1, R_2 = 0, R_3 = 0)}{p(l_2, l_3|l_1, R_1 = 1)}.$$

This ratio cannot be factorized into similar ratio terms in the edge-wise case since $R_2 = 0, R_3 = 0$ do not have the counterparts in the denominator to form a conditional odds term, thus inapplicable for ISA method. This leads to a similar issue to the absence of Conditions 1 and 2.

Subsequently, this issue propagates to step 2 of ISA, where we establish a connection between ω and β parameters via the conditional odds terms for individual R components. This brings us to the third question. If we do not make any connection between these two sets of parameters, explainability is lost as well. Furthermore, we bring the readers' attention to the fact that the tilting parameter in the default sensitivity analysis is a vector of the size $|l_{\bar{r}}|$, acting upon the entire imputed components in $L_{\bar{r}}$, and with equal components ranging from -1 to $+1$. This is due to the fact that in the
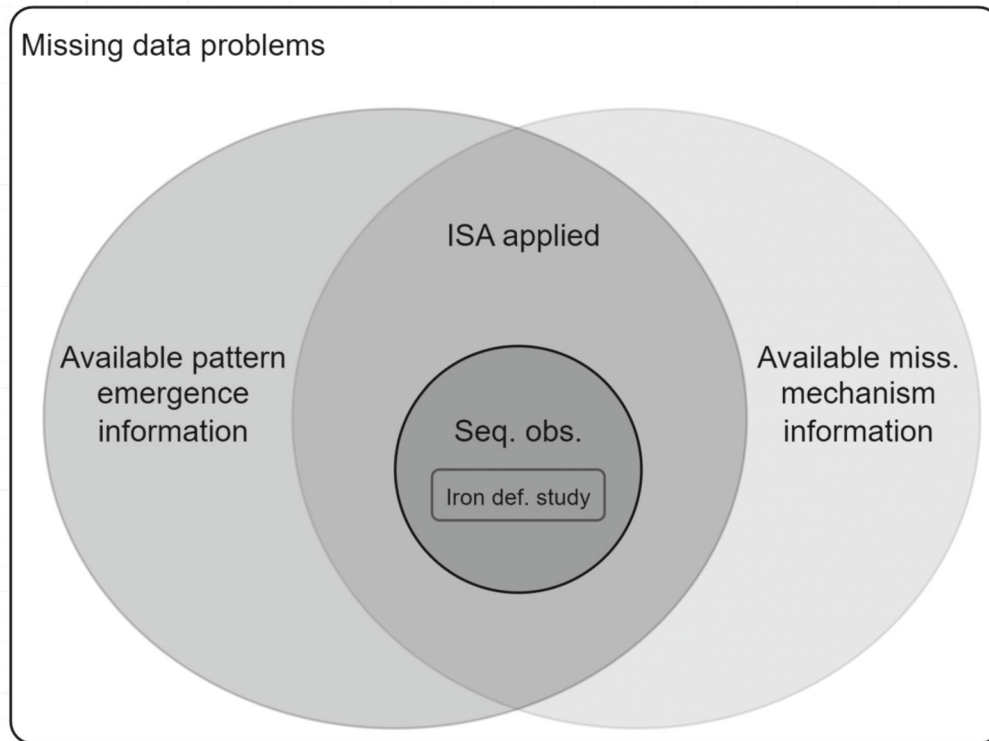
FIGURE 6 A schematic of the domain of applicability of ISA in missing data problems. If only pattern emergence information is available (Left circle), then we employ the pattern graph framework with its original assumption and default sensitivity analysis. If missingness mechanism information is available as well (intersection region), we can employ ISA. One of the scenarios for ISA is the introduced clinical sequential observations process. iron def, iron deficiency; Miss, missingness; Seq. obs, sequential observation.

default sensitivity analysis, we assume no further knowledge or assumption helps us with choosing the tilting parameters. In ISA, we assume that such knowledge exists, and it is encoded in the assumed m-graph, and the missingness model. This allows us to make informed guesses about the nature of the missing data problem, and choose accordingly each tilting component individually.

As a concluding remark, Theorem 2 states that in the absence of any extra information other than what we utilized to derive the pattern graph, ISA has no superiority over the default sensitivity analysis. ISA adopts the same pattern graph framework while incorporating applicable substantive knowledge from m-graphs.

4 | VALIDATION AND EXPERIMENTS

In Section 1.1, we introduced the clinical sequential observations missing data problem, in which both the pattern- and the mechanism-related information are often available; hence, this problem can benefit from the ISA method. This may not be the only missing data problem that ISA is able to address. In general, ISA can be used to approach all problems modeled via both m-graph and pattern graph, for which a working model is assumed. Figure 6 illustrates the domain of applicability of ISA method in relation to the pattern graph and m-graph frameworks, as well as the experiments for this section.

In this section, first, we present the ISA results for the bivariate sequential observations problem. Next, we set up a simulation study and generate a synthesized dataset according to the scenario to compare our method with the following four commonly used missing data methods: Unweighted CCA, KNN imputer, MICE, MissForest (See Appendix D for implementation details). We further compare the sensitivity analysis results with the default sensitivity analysis. Finally, we use real-world clinical data from the MIMIC-III public dataset[20] to evaluate our method. There, we estimate the cohort's mean iron level in the presence of missing data. We compare it with the estimated mean from the aforementioned four missing data methods. Furthermore, we demonstrate that in the presence of extra information about the problem, how one is able to validate the method and interpret the sensitivity analysis results.
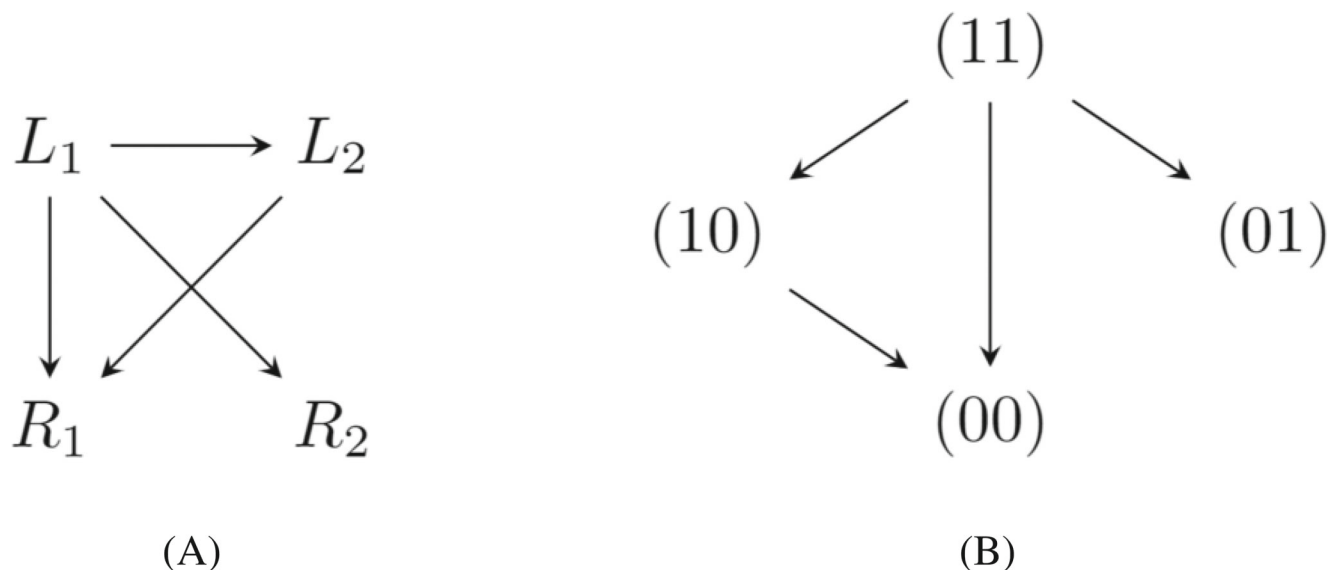
FIGURE 7 The assumed m-graph (a) and pattern graph (b) for the sequential observations problem.

4.1 | Case study: Clinical sequential observations

Consider a problem with two study variables L_1 and L_2 and the corresponding missingness indicators R_1 and R_2. Let the variable L_1 represent the severity of a disease (0=healthy, 1=sick) and L_2 represent an associated symptom for the disease (0=normal, 1=abnormal). To derive the m-graph and pattern graph, we explore the reasonable causal relations as well as the most frequent missingness scenarios in the clinical settings:

Causal structure of L: We define L_2 as the manifested symptom of the underlying disease L_1.

Missingness scenario 1: Occurs when a patient does not show up in the hospital (*missing-visit*); as an example, extremely healthy patients are less likely to visit doctors and therefore their associated variables are missing.

Missingness scenario 2: Occurs for patients with abnormal (eg, high) primary test values i.e. L_2. Here, the physician becomes suspicious about a particular disease and orders the secondary tests, that is, L_1.

Missingness scenario 3: Occurs when the variable L_2 is observed, but not recorded in the database, for example, when L_2 represents a symptom. Symptoms often are recorded in a non-standard textual format and extracted via text mining and natural language processing algorithms. An error in these algorithms leads to invalid and hence missing values in the collected dataset. On the other hand, the secondary test results are gathered by labs in a structured fashion, hence less prone to missingness and mistakes.

Figure 7 presents the derived m-graph and pattern graph for this case study (See Appendices E.1 and E.2 for more details about the derivation processes). In summary, the resulting pattern graph includes the following nodes, edges and imputation paths (denoted by π):

$$R_{1:4} = \{(1,1),(1,0),(0,1),(0,0)\},$$
$$E_{1:4} = \{r_1 \to r_2, r_1 \to r_3, r_1 \to r_4, r_2 \to r_4\},$$
$$\pi = \begin{cases} (e_1) & \text{to impute } r_2 \\ (e_2) & \text{to impute } r_3 \\ (e_3) & \text{to impute } r_4 \\ (e_4, e_1) & \text{to impute } r_4 \end{cases}. \tag{18}$$

By assuming the logistic models R_1 and R_2 as

$$\text{logit } p(R_1 = 0 | R_{-1}, L) = \beta_{11}L_1 + \beta_{12}L_2 + \beta_{10},$$
$$\text{logit } p(R_2 = 0 | R_{-2}, L) = \beta_{21}L_1 + \beta_{20},$$

we obtain the informed tilting parameters via the ISA method as

$$
\begin{aligned}
\pi_1 : \quad & \omega_1^{\pi_1} = 0 \\
\pi_2 : \quad & \omega_1^{\pi_2} = \beta_{11} \\
\pi_3 : \quad & \omega_1^{\pi_3} = (\beta_{11} + \beta_{21}, \beta_{12}) \\
\pi_4 : \quad & \omega_1^{\pi_4} = \beta_{11}, \quad \omega_2^{\pi_4} = 0.
\end{aligned} \tag{19}
$$

The details of ISA derivation are presented in Appendix E.3.

The obtained tilting parameters in (19) imply a need for informed guesses for the three logistic coefficients, for example:

- The odds of missing primary and secondary test results for healthy against sick patients are approximately λ_{21} and λ_{11} times larger. This implies $\beta_{11}^* \approx -\log \lambda_{11}$ and $\beta_{21}^* \approx -\log \lambda_{21}$.
- The odds of ordering the secondary test for patients with abnormal primary test results is approximately λ_{12} times larger for patients with normal primary test results. This implies $\beta_{12}^* \approx -\log \lambda_{12}$.

Next, we use this case study in the following two validation experiments.

4.2 | Validation: Simulation experiment

Our goal in this simulation step was to demonstrate that the ISA method yields unbiased results if prior knowledge is available, and missing data assumptions are correct. We further attempt to demonstrate the successful incorporation of prior knowledge in the ISA's sensitivity analysis by comparing the results with the default sensitivity analysis. Throughout the experiments, we thus utilized true assumptions, models, and parameters (ones which are used to synthesize data). For a detailed explanation of the experiments' setup as well as the results, see Appendix F.

Figure 8 presents the estimation bias of expected values for the designed bivariate clinical sequential observations, as introduced in Section 4.1. The ISA method obtained better results compared to unweighted CCA, the single imputation KNN method, and multiple imputation methods: MICE and Missforest.

For default versus edge-wise sensitivity analysis comparison, we made informed choices of tilting parameters according to ISA steps. Figure 9 presents the results of two sensitivity analysis approaches for the simulated dataset. As depicted, the varying range under ISA complies better with the ground truth. Improvement in estimating L_1 is more significant, as
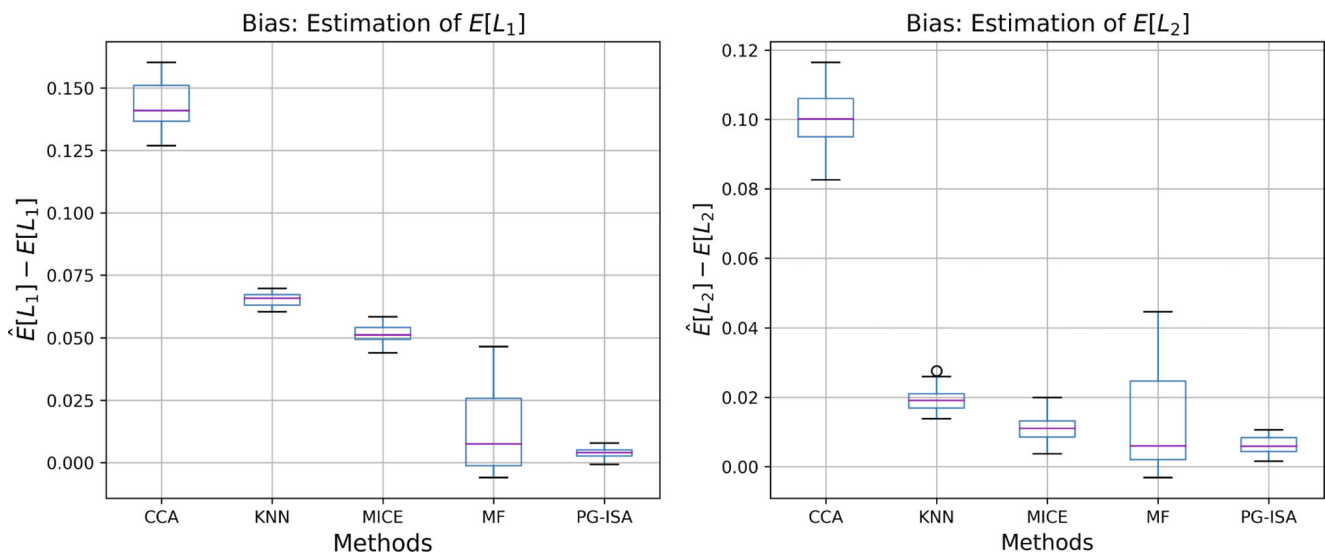

FIGURE 8 Estimation results for the bivariate clinical sequential observations case study for a simulation round.
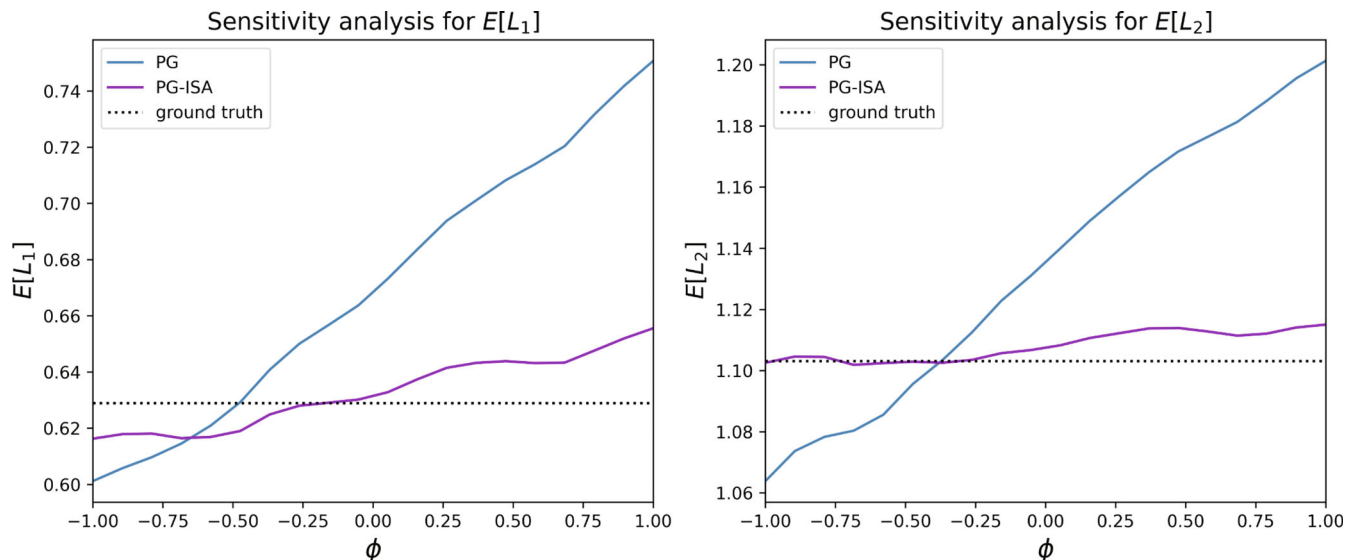
FIGURE 9 Sensitivity Analysis for estimation of $E[L_1], E[L_2]$ with agnostic and informed sensitivity parameters.

it is subjected to self-censoring, implying a higher departure from the pattern graph assumption. Figure 9 demonstrates how ISA improves the estimations if prior knowledge is valid.

4.3 | Demonstration: Iron deficiency experiment

Section 1.1 introduced the diagnostics flowchart for thalassemia carrier identification. We further discussed how it can be considered as a clinical sequential observations scenario. This section aims to estimate the true average iron level (IL) in a hospital cohort in the presence of missing data. As a brief recall, we know, according to the flowchart (Figure 1), that physicians diagnose iron deficiency based on the values of the following tests (if available): MCV, MCH, Hb pattern, iron level and a number of genetic tests.[4] We took the set of MCV, MCH, and Hb pattern tests as primary since all are part of the *chemistry* category blood test, and IL test as secondary tests. Our dataset was the MIMIC-III public dataset (see Appendix G for details about the dataset).

We employ the ISA's solution for sequential observations, obtained in Section 4.1. The remaining step to take is step 3 of ISA i.e. making informed choices of the tilting parameters. According to (19), we make informed guesses for β_{11}, β_{12}, and β_{21}.

1. The odds of missing primary and secondary test results for healthy patients are approximately λ_{21} and λ_{11} times larger compared to sick patients. This implies $\beta_{11}^* \approx -\log \lambda_{11}$ and $\beta_{21}^* \approx -\log \lambda_{21}$.
2. The odds of ordering the secondary test for patients with abnormal primary test results is approximately λ_{12} times larger for patients with normal primary test results. This implies $\beta_{12}^* \approx -\log \lambda_{12}$.

The first statement points to the epidemiological statistics of iron deficiency in the population of interest, which can be extracted from related literature. In order to proceed, we assumed that the missing values of the study variables positively correlate with the patients being extremely healthy (as the sick population visits the hospital to take the tests more often). Therefore, we associated the missing follow-ups with high values of IL, MCV and MCH. To formulate our prior knowledge, we assumed that: *the odds of missing follow-ups (both variables) for extremely healthy patients is $\lambda_{11} = \lambda_{21} = 2.5$ times higher compared to extremely sick patients*. Using Equation (16) this assumption translates to $\beta_{11}^* = \beta_{21}^* \approx -0.9$. We utilized these guessed values, knowing that the domain experts are able to validate and adjust the numbers if required. Regarding the large size of missingness for HbA2, the thalassemia flowchart[4] supports this assumption that the Hematology lab tests are performed regardless of their direct observed values, since any range and value of three items (MCV, MCH, HbA2) are used to order further tests. This means that we can consider the missingness in HbA2 as not self-censoring. Based on Figure 10 we modeled the missingness mechanism of HbA2 as a logistic regression model given the values of MCV and
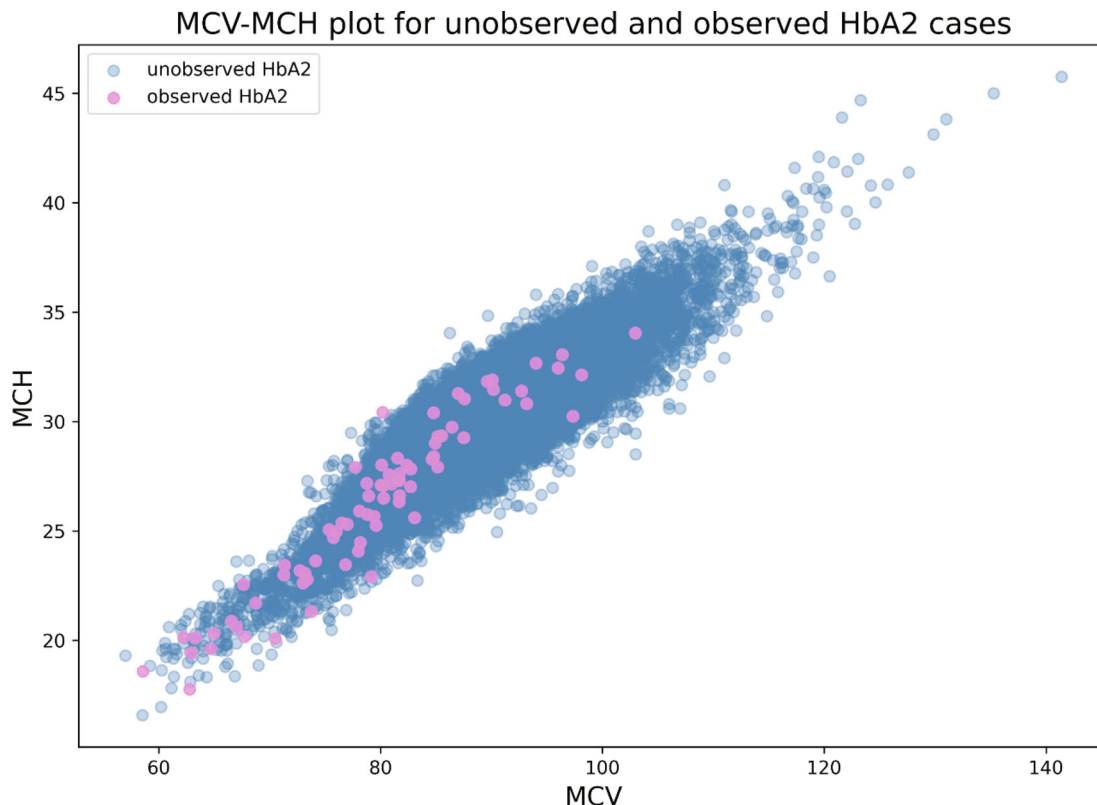
FIGURE 10 MCV-MCH plot for observed and unobserved HbA2 cases.

MCH. In other words, we imputed the HbA2 variable (only for the rows with observed MCV and MCH) as a function of MCV and MCH variables.

Second statement points to physicians' opinions and the diagnostic flowchart associated to the iron deficiency, formulated as

$$\lambda_{12} = OR(R_1|L_2 = \text{low} : \text{high}). \tag{20}$$

Similar to the previous item, one can make informed guesses for this item by analyzing physicians' style of practice in a target hospital. In the absence of such information, we estimated $E[R_1|L_2]$ by assuming

$$p(R_1 = 1|L_2, R_2 = 0) = p(R_1 = 1|L_2, R_2 = 1) \quad \text{(MAR)},$$

in order to make a reasonable initial guess for β_{12}^*. We admitted that this is potentially a biased estimation (as we must estimate $E[\lambda_{12}|L_2]$), but we accepted it, to take into account the potential bias later by selecting a range of variation and doing sensitivity analysis. This gave $\lambda_{12} \approx 1.35$, and thus $\beta_{12}^* = -log(\lambda_{12}) \approx -0.3$ is the preliminary approximated value obtained from (20). Apart from the MAR assumption, we made another simplifying assumption: even though L_2 is continuous, we guessed a constant value for β_{12}^* based on the odds ratio of lowest and highest values of L_2. In order to account for the approximations and simplifying assumptions, in the sensitivity analysis step, we vary the parameter β_{12} from its initial guess in order to account for the potential effect of these simplifying assumptions.

As the final step, we select an offset value for β_{11}, β_{21}, and one for β_{21}. With $\epsilon_{11} = \epsilon_{21} = 0.3$ and $\epsilon_{21} = 0.15$. According to (17), we have

$$\phi \in [-1, +1],$$
$$\beta_{11} = \beta_{21} = 0.3\phi + 0.9,$$
$$\beta_{12} = 0.15\phi + 0.3. \tag{21}$$

Subsequently, we calculated the informed tilting parameters according to Equation (19) and set the varying range for the tilting parameter of the default sensitivity analysis method to $\omega_{\text{default}} = \phi$ for future comparisons.
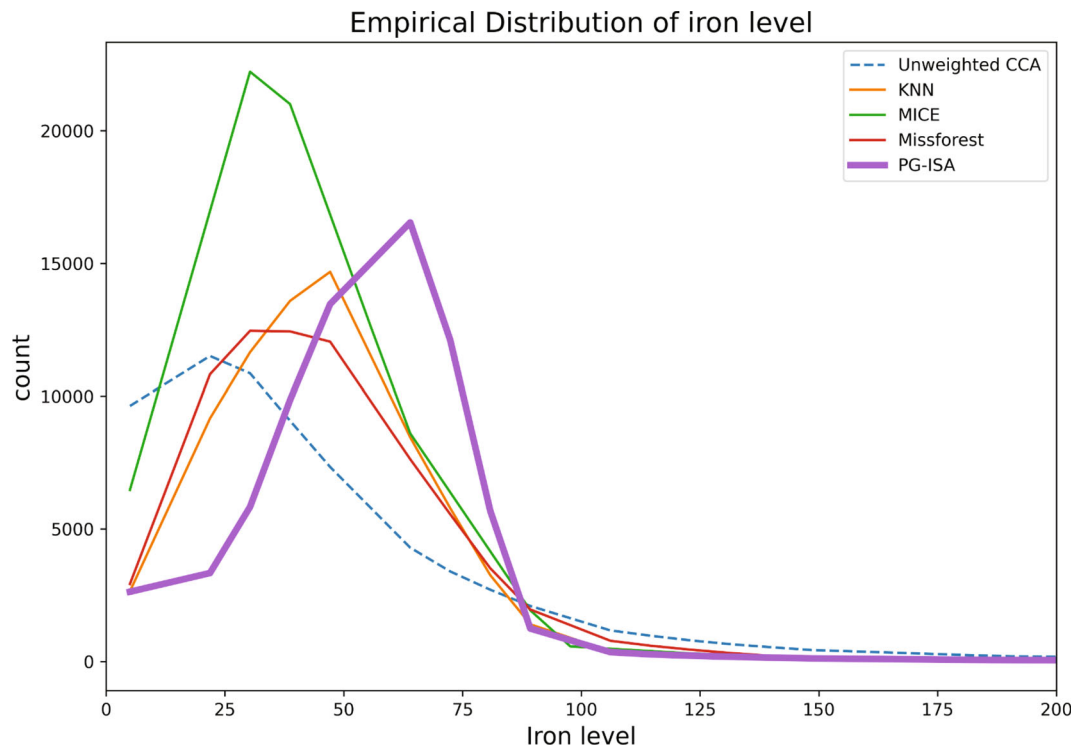
FIGURE 11 The estimated empirical distribution of IL for the entire cohort using different missing data approaches.

TABLE 1 The estimated mean and standard deviation of the IL in the entire cohort using different missing data approaches.

Method	Mean	STD
Unweighted CCA	45.32	32.30
KNN	46.49	27.15
MICE	41.70	22.74
MissForest	46.62	34.72
ISA	54.89	25.65

As explained throughout the paper, the choices in (21) are not solid inferred values but rather only guesses based on assumed available prior knowledge. What we achieved in the ISA method, however, was assessability (as clinical experts are able to verify and adjust above statements) and also interpretability of the results (as the sensitivity analysis results has a clear mapping to the selected values).

Figure 11 and Table 1 presents the estimated distribution of IL for the entire cohort using different methods. We imputed the partially-observed IL variable using unweighted CCA, KNN (K=33) imputation, MICE, MissForest, and ISA methods. We performed ISA using the initial guesses for tilting parameters, presented in (21).

Sensitivity analysis results for ISA and default approach are $\hat{E}_{\mathrm{ISA}}[\mathrm{iron}] \in [42.51, 65.20]$ and $\hat{E}_{\mathrm{default}}[\mathrm{iron}] \in [29.22, 55.90]$ for the default approach. While the difference in estimated ranges are significant, we cannot claim that the measured difference implies any superiority of a particular approach. An advantage of our method, however, is that one has the possibility of discussing and validating the edge-wise assumptions, encoded scenario, and the tilting parameters, by the medical experts. Furthermore, ISA results are interpretable, as the obtained range corresponds to meaningful quantities.

5 | DISCUSSION AND FUTURE WORK

In this paper, we introduced a new identifiability assumption for pattern graphs and, subsequently the ISA method for incorporating available prior knowledge about the missingness mechanism. ISA method allowed us to validate the

framework's assumption in a given problem and correct the bias accordingly (assessability). It also allowed us to make an interpretable connection between domain-expert knowledge and the sensitivity analysis results (interpretability). We concluded that ISA method enables the use of pattern graphs in a broader range of nonignorable missingness problems which otherwise would violate the framework's assumption.

We demonstrated how the causal independence statements from m-graphs are used to simplify the pattern graph assumptions. We also introduced examples of prior knowledge statements and how they relate to the parameters of the missingness model.

We implemented the ISA method to solve the bivariate *sequential observations* clinical case study, which involved nonignorable missingness, violating the pattern graph assumption. We assessed the edge-wise assumptions and calculated the corresponding tilting parameters. We identified two pieces of prior knowledge to make informed guesses for the logistic coefficients: information about rate of hospital visit for health and sick patients, and protocols of ordering a test based on primary findings. We compared the performance of our method with widely-used missing data methods (including original pattern graph) in a simulation study. We studied a routine real-world clinical data and compared our results with other methods.

One may adopt the methods presented in Section 3 for other missing data scenarios. The adoptation includes the calculation of tilting parameters and identifying interpretable prior knowledge for those scenarios. We considered using the logistic model for missingness indicators. There are other working models suggested in the literature, such as the Gaussian model.[21] One may investigate the use of such models in the ISA framework. Furthermore, we introduced sufficient conditions for incorporation of m-graph information in the pattern graph framework. This list can be extended considering more real-world scenarios. In particular, one may investigate the sensitivity models with irreducible conditional odds with no immediate interpretable missingness model coefficients despite available causal prior knowledge.

AUTHOR CONTRIBUTIONS

Alireza Zamanian: Conceptualization (lead); Formal Analysis (lead); Methodology (lead); Software (lead); Validation (equal); Visualization (lead); Writing – Original Draft Preparation (lead). **Narges Ahmidi:** Funding Acquisition (lead); Supervision (equal); Validation (equal); Writing – Review & Editing (supporting). **Mathias Drton:** Conceptualization (supporting); Methodology (supporting); Supervision (equal); Validation (equal); Writing – Review & Editing (lead).

ACKNOWLEDGMENTS

FUNDING INFORMATION

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The experiments in this paper are conducted using simulated and real-world data. Simulations were done according to guidelines presented in Subsection 4.2. Real-world data is extracted from the MIMIC-III public dataset,[20] requested and accessed from the Physionet repository.[22]

ENDNOTES

*In Machine learning community, the general sequential observations process is also known as Active Feature-value Acquisition (AFA).[3] This field is closely related to Reinforcement Learning, whose focus is training an effective *observer agent*. We chose not to use the AFA name, to avoid any potential confusion regarding the scope of the paper.

†Throughout the items, note that all four components in (11a) may include more than one variables.

‡We use the notations a and b to represent α and β parameters of the Beta distribution to avoid confusion with the beta coefficients of the logistic models.

ORCID

Alireza Zamanian 🔟 https://orcid.org/0000-0001-7412-6755

REFERENCES

1. Chen YC. Pattern graphs: a graphical approach to nonmonotone missing data. *Ann Stat*. 2022;50(1):129-146.
2. Bickley L, Szilagyi PG. *Bates' Guide to Physical Examination and History-Taking*. Philadelphia, PA: Lippincott Williams & Wilkins; 2012.
3. Saar-Tsechansky M, Melville P, Provost F. Active feature-value acquisition. *Manag Sci*. 2009;55(4):664-684.
4. Cao A, Galanello R. Beta-thalassemia. *Genet Med*. 2010;12(2):61-76.
5. Shimozawa N. Molecular and clinical findings and diagnostic flowchart of peroxisomal diseases. *Brain Develop*. 2011;33(9):770-776.
6. Deconinck B, Verschakelen J, Coolen J, Verbeken E, Verleden G, Wuyts W. Diagnostic workup for diffuse parenchymal lung disease: schematic flowchart, literature review, and pitfalls. *Lung*. 2013;191(1):19-25.
7. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Vol 793. Hoboken, New Jersey: John Wiley & Sons; 2019.
8. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866.
9. Robins JM. Causal inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality*. New York, NY: Springer; 1997:69-117.
10. Zhou Y, Little RJ, Kalbfleisch JD, et al. Block-conditional missing at random models for missing data. *Stat Sci*. 2010;25(4):517-532.
11. Mohan K, Pearl J, Tian J. Graphical models for inference with missing data. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K, eds. *Advances in Neural Information Processing Systems*. Vol 26. Red Hook, NY: Curran Associates, Inc; 2013.
12. Mohan K, Pearl J. Graphical models for recovering probabilistic and causal queries from missing data. *Probabilistic and Causal Inference: the Works of Judea Pearl*. New York, NY:ACM; 2022:413-432.
13. Shpitser I, Mohan K, Pearl J. Missing data as a causal and probabilistic problem. *Uncertainty in Artificial Intelligence-Proceedings of the 31st Conference*. Amsterdam, The Netherlands: AUAI Press; 2015:802-811.
14. Nabi R, Bhattacharya R, Shpitser I. Full law identification in graphical models of missing data: completeness results. *International Conference on Machine Learning*. Brookline, Massachusetts: PMLR; 2020:7153-7163.
15. Kim JK, Yu CL. A semiparametric estimation of mean functionals with nonignorable missing data. *J Am Stat Assoc*. 2011;106(493):157-165.
16. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Stat Assoc*. 1999;94(448):1096-1120.
17. Franks AM, Airoldi EM, Rubin DB. Nonstandard conditionally specified models for nonignorable missing data. *Proc Natl Acad Sci*. 2020;117(32):19045-19053.
18. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res*. 2007;16(3):259-275.
19. Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Stat Med*. 2008;27(23):4678-4721.
20. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):1-9.
21. Le Morvan M, Josse J, Moreau T, Scornet E, Varoquaux G. NeuMiss networks: differentiable programming for supervised learning with missing values. *Adv Neural Inf Process Syst*. 2020;33:5980-5990.
22. Johnson A, Pollard T, Mark R III. MIMIC-III clinical database. *Phys Net*. 2016;10:C2XW26.
23. Guénais T, Vamvourellis D, Yacoby Y, Doshi-Velez F, Pan W. BaCOUn: Bayesian Classifers with out-of-distribution uncertainty. arXiv preprint arXiv:2007.06096. 2020.
24. Blitzstein JK, Hwang J. *Introduction to Probability*. New York, NY: Chapman and Hall/CRC; 2019.
25. Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP. Paper presented at: International Conference on Learning Representations, OpenReview. 2017.
26. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell*. 1984;6:721-741.
27. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.

APPENDIX A. PROOFS

A.1 Proof of Theorem 1

Assume an imputation path, as illustrated in Figure 3, which starts from $r^{(0)}$, ends at $r^{(K)} = \mathbf{1}_d$, and goes over edges $e^{(1)}$ to $e^{(K)}$. The goal is to sample imputations for the components in $L_{\bar{r}^{(0)}}$ from the following density

$$p(l_{\bar{r}^{(0)}}|l_{r^{(0)}}, R = r^{(0)}).$$

Algorithm 1 breaks this sampling into K steps of partial-imputations across K edges. Therefore, at each edge the goal is to sample from the density

$$p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(0)}).$$

However, the algorithm samples the imputation from the parent of the edge, namely $r^{(k)}$. Thus, the imputation is correct if these two densities are equal, that is, the following ratio holds:

$$\frac{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(0)})}{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(k)})} = 1.$$

By the chain rule, we re-write the assumption as

$$\frac{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(0)})}{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(k)})} = \frac{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(0)})}{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(1)})} \frac{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(1)})}{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(2)})} \cdots \frac{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(K-1)})}{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(K)})}$$

$$= \prod_{i=1}^{K} \frac{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(i-1)})}{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(i)})}.$$

□

A.2 Proof of Corollary 1

We reorder an L-sorted ratio as the following:

$$\frac{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(i-1)})}{p(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(i)})} = \frac{p(l_{r^{(k)}}, r^{(i-1)})}{p(l_{r^{(k)}}, r^{(i)})} \frac{p(l_{r^{(k-1)}}, r^{(i)})}{p(l_{r^{(k-1)}}, r^{(i-1)})}.$$

The second ratio is independent of L_{Δ_k}, which we replace by a constant called A. We factorize the numerator and denominator of the first ratio with respect to the components which are different between $r^{(i-1)}$ and $r^{(i)}$; namely R_{Δ_i}:

$$= \frac{p\left(r^{(i-1)}_{\Delta_i}|l_{r^{(k)}}, r^{(i-1)}_{-\Delta_i}\right) p\left(l_{r^{(k)}}, r^{(i-1)}_{-\Delta_i}\right)}{p\left(r^{(i)}_{\Delta_i}|l_{r^{(k)}}, r^{(i)}_{-\Delta_i}\right) p\left(l_{r^{(k)}}, r^{(i)}_{-\Delta_i}\right)} A$$

By definition, the components other than R_{Δ_i} are the same between the parent and child patterns; namely $R^{(i-1)}_{-\Delta_i} = R^{(i)}_{-\Delta_i}$. Therefore the second term from numerator and denominator is cancelled out. Moreover, we know that $R^{(i-1)}_{\Delta_i} = \vec{0}$ and $R^{(i)}_{\Delta_i} = \vec{1}$. Thus, we rewrite the ratio as

$$= \frac{p\left(R_{\Delta_i} = \vec{0}|l_{r^{(k)}}, r^{(i-1)}_{-\Delta_i}\right)}{p\left(R_{\Delta_i} = \vec{1}|l_{r^{(k)}}, r^{(i)}_{-\Delta_i}\right)} A.$$

By choosing the index set J from the L-sorted ratios to do such reordering and leaving the rest of the ratios, that is, the ratios in the index set H, we rewrite Equation (7) as

$$\prod_{h\in H} \frac{p\left(l_{\Delta_k}|l_{r^{(k-1)}}, r^{(h-1)}\right)}{p\left(l_{\Delta_k}|l_{r^{(k-1)}}, r^{(h)}\right)} \prod_{j\in J} \frac{p\left(R_{\Delta_j} = \vec{0}|l_{r^{(k)}}, r_{-\Delta_j}\right)}{p\left(R_{\Delta_j} = \vec{1}|l_{r^{(k)}}, r_{-\Delta_j}\right)} = C,$$

where $C = \prod_{j\in J} A_j^{-1}$ the constant terms of each R-sorted ratio.

□

As this ratio is going to be utilized either in the rejection probability of the rejection sampling method, or the importance ratio of the importance sampling method, any term C independent of the sampled imputation will be cancelled out (by being cancelled out from the numerator and the denominator of the rejection probability or via normalization of the weights of each sample in the weighted-averaging step, respectively). Thus it is valid to have the R-sorted ratios up to a normalizing constant.

A.3 Proof of conditional independence (11a) and derivation of the three m-graph restrictions

Starting with the independence statement (11a), we recall an R-sorted ratio is written as

$$\frac{p\left(R_{\Delta_j} = \vec{0}|l_{r^{(k)}}, r_{-\Delta_j}\right)}{p\left(R_{\Delta_j} = \vec{1}|l_{r^{(k)}}, r_{-\Delta_j}\right)}.$$

By definition, we have $L_{r^{(k)}} = \{L_{\Delta_k}, L_{r^{(k-1)}}\}$. Hence, we rewrite the conditioning set as $\{L_{\Delta_k}, L_{r^{(k-1)}}, R_{-\Delta_j}\}$. Consequently, if the second and third term of the conditioning set, that is, $\{L_{r^{(k-1)}}, R_{-\Delta_j}\}$ d-separates R_{Δ_j} and L_{Δ_k}, the latter is removed from the conditioning set, thus the R-sorted ratio reduces to a constant (hence, (11a)).

By the rules of d-separation, a set C is a d-separation set for A, B if A and B do not have a direct causal edge, C blocks all the confounder paths, and no variable in C opens a collider path:

1. Restriction i represents the 'no direct edge' rule.
2. Restriction ii represents the 'no confounder' rule. The only component which is left out of the R-sorted conditioning set is $L_{\mathbf{1}_d - r^{(k)}}$, that is, the *upstream* components on the imputation path of the pattern graph. As this component does not appear in the conditioning set, any unmediated confounder path cannot be blocked by $\{L_{r^{(k-1)}}, R_{-\Delta_j}\}$.
3. Restriction iii represents the 'no collider' rule. Since we assumed NDE (see Section 3.3), no L variable can be a collider for Rs, hence only the $R_{-\Delta_j}$ component needs to be mentioned in the restriction.

A.4 Proof of Theorem 2

After the imputation Algorithm 1 selects a parent $r^{(k)}$ to impute the pattern $r^{(k-1)}$ (not necessarily at the beginning of the imputation; $r^{(k-1)}$ can be any intermediate node for imputation of a lower pattern), the right-hand side of the original pattern graph assumption is stated as

$$\frac{p(l_{\bar{r}^{(k-1)}}|l_{r^{(k-1)}}, R = r^{(k-1)})}{p(l_{\bar{r}^{(k-1)}}|l_{r^{(k-1)}}, R = r^{(k)})}.$$

We factorize the components in $L_{\bar{r}^{(k-1)}}$ with respect to the ancestors of the current child, as we expect it to be imputed:

$$= \frac{p\left(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(k-1)}\right)}{p\left(l_{\Delta_k}|l_{r^{(k-1)}}, R = r^{(k)}\right)} \frac{p\left(l_{\Delta_{k+1}}|l_{r^{(k-1)}}, R = r^{(k-1)}\right)}{p\left(l_{\Delta_{k+1}}|l_{r^{(k-1)}}, R = r^{(k)}\right)} \cdots \frac{p\left(l_{\Delta_K}|l_{r^{(k-1)}}, R = r^{(k-1)}\right)}{p\left(l_{\Delta_K}|l_{r^{(k-1)}}, R = r^{(k)}\right)}$$

$$= \prod_{i=0}^{K-k} \frac{p\left(l_{\Delta_{k+i}}|l_{r^{(k-1)}}, R = r^{(k-1)}\right)}{p\left(l_{\Delta_{k+i}}|l_{r^{(k-1)}}, R = r^{(k)}\right)}$$

Each ratio term written for $L_{\Delta_{k+i}}$ corresponds to a ratio term in the edge-wise assumption in (7) where $\Delta_{k+i} := \Delta_k$ and $r^{(k-1)} :=$; in other words, if we factorize the original assumption for all patterns, and write the edge-wise assumption for all edges, every term in one, is found in the other. This means that the set of all terms, the product of which must be equal to one is shared between two assumptions, but with a different order. In theory we consider a case where each term in an assumption statement is not one, but the product of all terms are. However, this is an arbitrary case that might or might not occur in a problem. In order for the assumptions to structurally hold, we should have all the terms being equal to one. □

As an example, assume a simple pattern graph with three missingness patterns of

$$\mathbf{R} = \{(111), (110), (100)\}$$

and two imputation paths

$$\pi = \{(111) \to (110), (111) \to (110) \to (100)\}$$

We write two pattern graph assumptions for two incomplete patterns:

$$\frac{p(l_2, l_3|l_1, R = (100))}{p(l_2|l_1, R = (110))} = \frac{p(l_2|l_1, R = (100))}{p(l_2|l_1, R = (110))} \frac{p(l_3|l_1, l_2, R = (100))}{p(l_3|l_1, l_2, R = (110))} = 1,$$

$$\frac{p(l_3|l_1, l_2, R = (110))}{p(l_3|l_1, l_2, R = (111))} = 1.$$

We write the edge-wise assumption for three edges (one for imputation of (110), two for imputation of (100)):

$$\frac{p(l_2|l_1, R = (100))}{p(l_2|l_1, R = (110))} = 1, \quad \frac{p(l_3|l_1, l_2, R = (100))}{p(l_3|l_1, l_2, R = (110))} \frac{p(l_3|l_1, l_2, R = (110))}{p(l_3|l_1, l_2, R = (111))} = 1,$$

$$\frac{p(l_3|l_1, l_2, R = (110))}{p(l_3|l_1, l_2, R = (111))} = 1.$$

It has been shown that three L-sorted ratios are shared between the two assumptions.

APPENDIX B. EXAMPLES OF EDGE-WISE ASSUMPTION

We provide three examples related to Figure 3 which better describe Theorem 1, Corollary 1 and the sensitivity model in Equation (9).

Example 1 (Edge-wise assumption). The pattern (1000) in Figure 3 is imputed according to the selected path by first imputing (L_2, L_3), and then L_4 across two edges in (6b). Edge-wise assumptions are as follows:

$$e^{(1)} : \quad \frac{p(l_2, l_3|l_1, R = (1000))}{p(l_2, l_3|l_1, R = (1110))} = 1,$$

$$e^{(2)} : \quad \frac{p(l_4|l_1, l_2, l_3, R = (1000))}{p(l_4|l_1, l_2, l_3, R = (1110))} \frac{p(l_4|l_1, l_2, l_3, R = (1110))}{p(l_4|l_1, l_2, l_3, R = (1111))} = 1.$$

Example 2 (Edge-wise assumption with L- and R-sorted ratios). We rewrite the edge-wise assumption for $e^{(2)}$ in Example 1 using L-sorted and R-sorted ratios. In Equation (8), let $H = \{1\}$ and $J = \{2\}$, that is, second ratio must be written in its R-sorted form. Then the edge-wise assumption for $e^{(2)}$ is expressed as

$$e^{(2)} : \quad \frac{p(l_4|l_1, l_2, l_3, R = (1000))}{p(l_4|l_1, l_2, l_3, R = (1110))} \frac{p(R_4 = 0|L, R_{1:3} = (111))}{p(R_4 = 1|L, R_{1:3} = (111))} = C.$$

Example 3 (Edge-wise local sensitivity model). Continuing Example 2, we write the local sensitivity model for $e^{(2)}$ based on Equation (9) as

$$e^{(2)} : \quad \frac{p(l_4|l_1, l_2, l_3, R = (1000))}{p(l_4|l_1, l_2, l_3, R = (1110))} \frac{p(R_4 = 0|L, R_{1:3} = (111))}{p(R_4 = 1|L, R_{1:3} = (111))} = C \exp(\omega_2 l_4).$$

where ω_2 is the tilting parameter for $e^{(2)}$ on the selected imputation path.

APPENDIX C. MCAR AND MAR MISSINGNESS IN ISA

MCAR missingness mechanism is defined by the independence assumption $R \perp\!\!\!\perp L$. In the corresponding m-graph, Rs accept no incoming edges from the L nodes. With a slight abuse of terminology, we assign the name "MCAR" to an individual R node of the same condition in an m-graph, that is, $R_i \perp\!\!\!\perp L, R_{-i}$ (receiving no edge from other R nodes as well). In this case, all three restrictions of Condition 1 holds, and hence the corresponding R-sorted ratio is cancelled out. In an extreme case, if all R nodes are of MCAR type, all partial imputations on all imputation paths are free of the need for sensitivity analysis.

Likewise, MAR missingness mechanism is defined by the independence assumption $R \perp\!\!\!\perp L_{\text{miss}} | L_{\text{obs}}$ where the index set *obs* selects the fully-observed variables, and *miss*, the remaining ones. In the corresponding m-graph, Rs are restricted to have incoming edges only from the fully-observed nodes. Similar to the MCAR case, we assign the name "MAR" to an individual R node of the same condition in an m-graph. Condition 1 holds for MAR R nodes as well, as the component $L_{r^{(k-1)}}$ (including observed variables) is a sufficient d-separation set. Condition 1 holds also for the weaker "non-monotone MAR" mechanism, where the subscript *obs* selects the observed *entries* for a sample. Obviously, this is a weaker assumption since L_{obs} may include the partially-observed variables which are observed for that particular sample. Nevertheless, the observed entries are still included in $L_{r^{(k-1)}}$, hence a similar discussion as for MAR is valid.

APPENDIX D. SELECTED MISSING DATA METHODS FOR COMPARISON

Table D1 introduces the missing data methods for comparison in our experiments.

TABLE D1 Missing data methods for comparison with informed sensitivity analysis pattern graph.

Method	Details
Unweighted CCA	By calculating the expected value of the variables using the complete cases.
KNN imputer	Scikit-learn[27] implementation with `K=13`.
MICE	Scikit-learn implementation (estimator: Bayesian Ridge Regression).
MissForest	Scikit-learn implementation of random forest (`n_trees=15, min_sample_leaf=2, min_samples_split=2, max_iter=10`).

APPENDIX E. DETAILS OF ISA METHOD FOR SEQUENTIAL OBSERVATIONS PROBLEM

E.1 Pattern graph

Chen[1] claims that generation of the correct pattern graph for a scenario is intrinsically an open problem. Nevertheless, a pattern graph structure selection procedure has been introduced by the author which is summarized as follows:

The general idea of the procedure is to have in mind a scenario of step-wise data collection, be it the order of questions in a questionnaire or the chronological order of lab tests through time which matches the reality of the scenario. We start from when it is not yet determined whether the study variables are missing, for example, at the beginning of the questionnaire or first admission day. From that point, we specify, step-by-step, the possibilities of observing/missing a variable in the process. For that purpose, in addition to 0 and 1 values for r, we use also a *placeholder* value "$-$". The placeholder implies that the observation of a variable has not yet been decided upon.

As an example, imagine a case of 3 lab tests, the first two of which are taken simultaneously while the third one is performed the next day. We start with the pattern $s_0 = (-, -, -)$. Patients may either miss or take the first two tests on the first day, therefore we arrive at two new patterns $s_1 = (1, 1, -)$ and $s_2 = (0, 0, -)$. For s_1 the third test might or might not be performed, therefore two new patterns emerge from s_1 which are $s_3 = (1, 1, 1)$ and $s_4 = (1, 1, 0)$. Moreover, suppose we know by the clinic protocols, that the third test must be performed if the first two are missing. Therefore only one pattern $s_5 = (0, 0, 1)$ is resulted from s_2.

To finish the process, we replace all the placeholders with the value 1. Finally, the emerging patterns and the in-between edges are used to form the pattern graph, with this consideration that a self-looping edge are dropped, and duplicated edges are presented once. For the example at hand, we arrive at the following graph:

$$V = \{\mathbf{1}_d = (1, 1, 1), r_1 = (1, 1, 0), r_2 = (0, 0, 1)\},$$
$$E = \{(\mathbf{1}_d \rightarrow r_2), (\mathbf{1}_d \rightarrow r_1)\}.$$

For our case study, considering the procedure explained above, we construct the pattern graph as follows:

- We start from the initial step $s_0 = (-, -)$. A patient may be absent from the site, therefore no observations will be made for (z_1, z_2), which represents the pattern $s_1 = (0, 0)$. For a present patient, the symptom z_2 might or might not be recorded which represents $s_2 = (-, 1)$ and $s_3 = (-, 0)$.

- Based on the value of z_2, the test z_1 might or might not be taken disregarding whether or not z_2 is recorded. These represent $s_4 = (1, 1)$, $s_5 = (0, 1)$ for the next step of s_2, and $s_6 = (1, 0)$ and $s_7 = (0, 0)$ for the next step of s_3.
- Replacing the placeholders with the value 1, we arrive at four different patterns $r_0 = \mathbf{1}_d = (1, 1)$, $r_1 = (1, 0)$, $r_2 = (0, 1)$ and $r_3 = (0, 0)$ and the edges are $\mathbf{1}_d \rightarrow r_1, r_2, r_3$ and $r_1 \rightarrow r_3$

Figure E1 depicts the process of choosing the pattern graph and the final result.
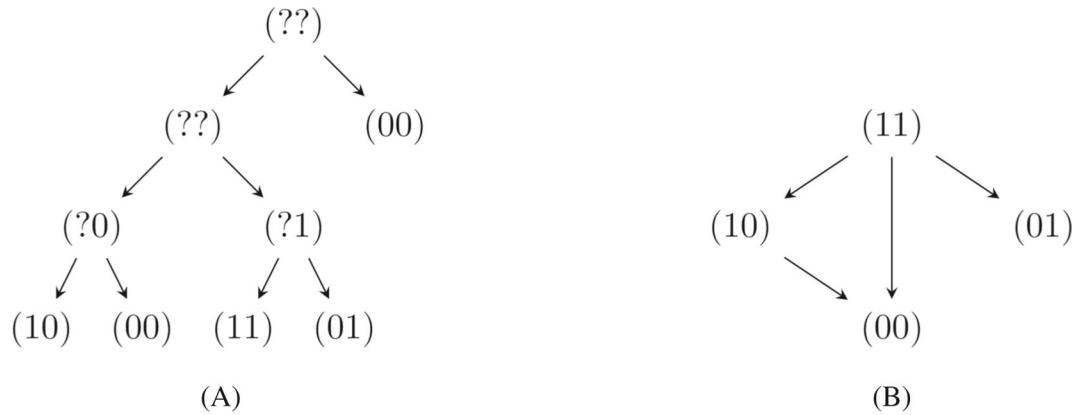

FIGURE E1 The corresponding pattern graph for the case study. (a): Encoding the data gathering process. (b): The constructed pattern graph

E.2 M-graph

To generate the m-graph, we introduced a causal relation as well as three missingness scenarios in Section 4.1. Here we discuss what those statements imply about the m-graph:

- **Causal structure of** L: It implies the causal edge $L_1 \rightarrow L_2$ in the m-graph.
- **Missingness scenario 1** This scenario implies $L_1 \rightarrow R_1$ and $L_1 \rightarrow R_2$ edges in the m-graph.
- **Missingness scenario 2** This scenario implies the edge $L_2 \rightarrow R_1$.
- **Missingness scenario 3** This scenario points to an external cause of R_2 missingness, that is, other than L_1 and L_2 and therefore does not imply any explicit edge in the m-graph.

In the constructed m-graph (Figure 7), the variable L_1 is subjected to self-censoring ($L_1 \rightarrow R_1$). By subtracting the self-censoring edge, the m-graph becomes a subgraph of the block-parallel model.[11] Also, note that the missingness in both variables depends on the unobserved data.

E.3 Deriving informed tilting parameters via ISA

We describe the first two steps of ISA to derive the informed tilting parameters for the sequential observations problem.

Step 1: Simplifying the edge-wise assumption ratios

Given the information in Equation (18), we have 1 assumption for imputing r_2, 1 assumption for imputing r_3, and 3 assumptions on 2 imputation paths for imputating r_4:

1. π_1: e_1 **edge for imputing** r_2: the L-sorted ratio for the edge-wise assumption is $p(l_2|l_1, R = (10))/p(l_2|l_1, R = (11))$. This is an identifiable case, as presented by (11a), since $L_1 \in \text{pa}_{\text{m-graph}}(L_2)$; therefore, no sensitivity analysis is needed:

$$\omega_1^{\pi_1} = 0 \tag{E1a}$$

2. π_2: e_2 **edge for imputing** r_3: the R-sorted exponential tilt for this edge, according to Equation (11b), is

$$O(R_1|l_1 l_2) = \exp\left(\omega_2^{\pi_2} l_1\right). \tag{E1b}$$

Note that the condition term $R_2 = 1$ is removed from the terms since $R_1 \perp\!\!\!\perp R_2|L$.

3. π_3: e_3 **edge for imputing** r_4: In this case the number of components for imputation is more than 1 (L_1, L_2), and according to the m-graph, missingness indicators of the components have no direct causal edges. Therefore Condition 2 is met. We use (11d) to write the R-sorted exponential tilting terms

$$O(R_1|L)O(R_2|L) = \exp\left(\omega_3^{\pi_3}L\right) \tag{E1c}$$

4. π_4: e_4 **edge for imputing** r_4: This is the first edge of the π_4 imputation path for R_4, where L_1 is being imputed. similar to π_2, we have

$$O(R_1|l_1) = \exp\left(\omega_4^{\pi_4}l_1\right) \tag{E1d}$$

5. π_4: e_1 **edge for imputing** r_4: This is the second edge of π_4 to complete the partially-imputed r_4 by sampling L_2. Similar to π_1, the variable L_2 is in the condition set of the L-sorted ratio for this edge, and therefore this case is identifiable:

$$\omega_1^{\pi_4} = 0 \tag{E1e}$$

Step 2: Obtaining the tilting parameters

Assuming the logistic model, we obtain the tilting parameters via Equation (16):

- Equations (E1a) and (E1e) directly give $\omega_1^{\pi_1} = 0$ and $\omega_1^{\pi_4} = 0$.
- In both Equations (E1b) and (E1d), we have the conditional odds of R_1 in the process of imputing L_1. It gives $\omega_2^{\pi_2} = \omega_4^{\pi_4} = \beta_{11}$.
- In Equation (E1b), we have two conditional odds of R_1 and R_2 in the process of imputing L_1 and L_2. The odds of R_1 gives β_{11} for L_1 and β_{12} for L_2, while odds of R_2 gives β_{21} for L_2 only. The result is the tilting vector $\omega_3^{\pi_3} = (\beta_{11} + \beta_{21}, \beta_{12})$.

E.4 MI Sampling

To estimate the probability densities $\hat{p}(L_{s-r}|L_r, R = s)$ in Algorithm 4, we categorize the estimation task in two classes:

1. when $L_r \neq \emptyset$ that is, the target is a conditional density. We can utilize probabilistic models such as Gaussian process or Bayesian neural network. For larger datasets, more scalable models such as neural linear model (NLM) are used.[23] Assuming linearity, one may also implement Bayesian linear regression for binary/categorical or the generalized linear model for continuous variables.

2. when $L_r = \emptyset$. For example, in a dataset with three variables $L = (L_1, L_2, L_3)$, to impute a row $L_i = (?, ?, ?)$ with the missingness indicator $R_i = r_0 = (0, 0, 0)$ where $s \rightarrow r_0$ and $s = (1, 1, 0)$, the target is $\hat{p}(l_1, l_2|R_1 = 1, R_2 = 1, R_3 = 0)$. This is the problem of *density estimation*, an unsupervised machine learning task where the goal is to establish an approach to sample from an unknown distribution from which the observations (in case of the above example (L_1, L_2)) have been drawn. In case of a univariate distribution with a large enough sample size, the most straightforward solution is based on the empirical cumulative distribution function (eCDF) and the universality of the uniform theorem.[24] The eCDF approach is equivalent to sampling with replacement from a realization vector. However, this solution is not feasible for high-dimensional data or for sparse and continuous variables. Dinh et al[25] summarize the following more advanced approaches:

   - Maximum likelihood models such as Restricted Boltzmann Machines or Deep Boltzmann Machines described by probabilistic undirected graphs, and Variational Autoencoders described by directed graphical models.
   - Adverserial models such as GAN, and the novel *real-valued non-volume preserving transformations* (real NVP) method proposed by Dinh et al.[25]

Finally, If the conditional distribution models are available for a set of variables, Gibbs sampling[26] can also be used for both aforementioned cases of conditional and marginal distributions.

In this study we chose GLM and eCDF approaches.

APPENDIX F. SIMULATION SETUP

The simulation steps are as follows:

1. For L_1, sample n independent realizations $l_1^{(1)}, \ldots, l_1^{(n)} \sim \text{Beta}(a, b)$.[‡]
2. Calculate L_2 as $l_2^{(i)} = \theta_2 l_1^{(i)2} + \theta_1 l_1^{(i)} + \theta_0 + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon)$ are independent Gaussian additive noise terms.
3. Calculate success probabilities P_1 and P_2 via the sigmoid functions $p_1^{(i)} = \sigma\left(\beta_{11} l_1^{(i)} + \beta_{12} l_2^{(i)} + \beta_{10}\right)$ and $p_2^{(i)} = \sigma\left(\beta_{21} l_1^{(i)} + \beta_{20}\right)$.
4. sample missingness indicators independently as $R_1^{(i)} \sim \text{Bern}\left(p_1^{(i)}\right)$ and $R_2^{(i)} \sim \text{Bern}\left(p_2^{(i)}\right)$.

The simulation parameters are a, b parameters of the L_1 distribution, $\theta = [\theta_0, \theta_1, \theta_2]$, σ_ϵ parameters of the structural equation of L_2, and β parameters of the missingness mechanisms. Figure F1 depicts an example of generated distribution in this simulation scenario.

First, we ran the simulations for the sample size $N = 1000$ with the set of parameters presented in Table F1. Table F2 and Figure 8 present the comparison results between different methods. The unweighted CCA gave the highest bias in the estimations. Also, MICE and MissForest (the MI methods) outperformed the single imputation KNN method. The ISA method obtained the best results. To showcase the process of making informed guesses based on prior knowledge, we considered two assumptions on the missingness model according to step 3 of ISA, corresponding to values $\beta_{11}, \beta_{21} \approx -0.7$, and $\beta_{12} \approx -0.4$. We used the reference parameter ϕ, as explained in (17), to compare ISA with the default sensitivity analysis.

To validate the results in a broader range of possibilities (According to Table F3 with a restriction that the missingness indicators satisfy $E[R_1] \geq 0.5$ and $E[R_2] \geq 0.5$), we ran the same simulation scenario 50 times with different simulation parameters. Similar to the previous scenario, we assumed that the missingness mechanism is known for ISA method. Results in Figure F2 show smaller differences among models in this round of simulation compared to the first run. This may be due to the fact that we intentionally devised the first scenario to showcase an extreme case where conventional methods such as MICE and MissForest would fail, while in the second run parameters were chosen randomly.
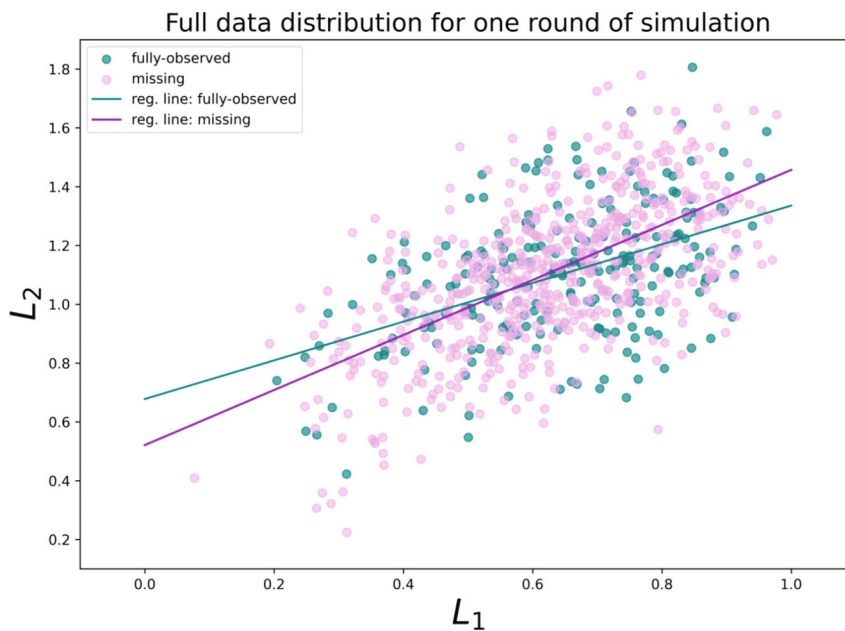

FIGURE F1 The full data distribution for the first simulation run with the parameters presented in Table F1. Fully-observed and missing (either L_1, L_2 or both) samples are presented with green and purple dots respectively. The plot exhibits the deviation of the regression line slope of the extrapolation density from that of the observed density.
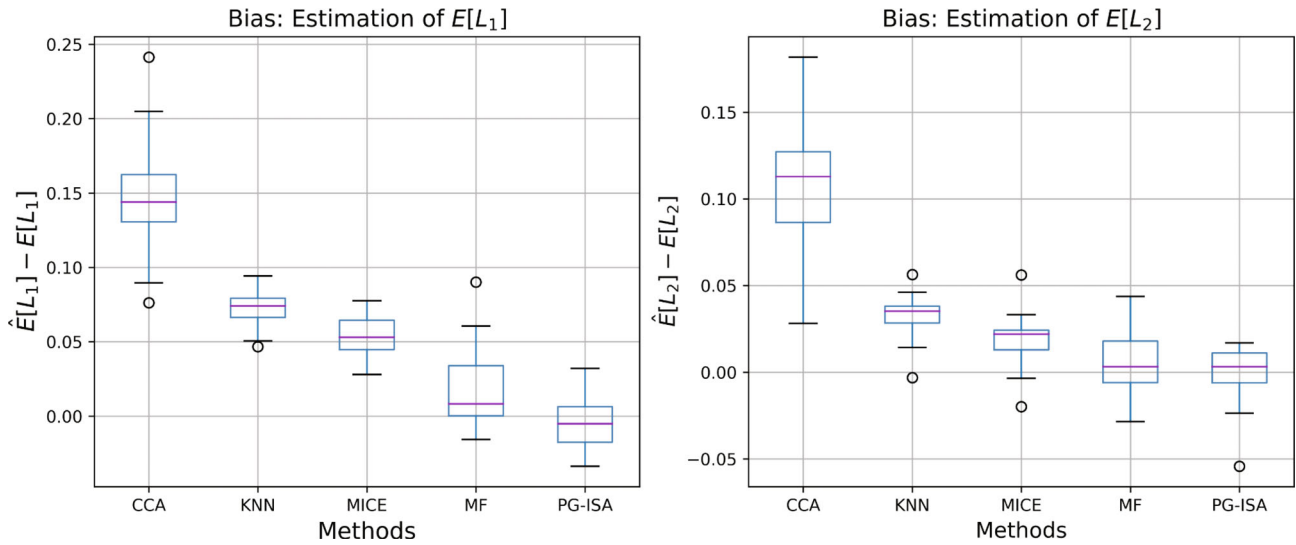
FIGURE F2 Box plot visualization for the results of different simulation runs using Table F3.

TABLE F1 The simulation parameters of the first round of simulation.

Parameter	Value
α, β	$5, 3$
$(\theta_0, \theta_1, \theta_2)$	$(0.3, 1.8, 0.8)$
σ_ϵ	0.2
$(\beta_{11}, \beta_{21}, \beta_{12}, \beta_{10}, \beta_{20})$	$(-0.6, -0.6, -0.3, 0.6, 0.3)$

TABLE F2 The mean estimation bias, and STD of bias in the first simulation for the study variables L_1, L_2. for $n = 1000$, number of iterations $= 50$, and simulation parameters according to Table F1.

Method	$E[L_1]$ (bias $\pm$ std)	$E[L_2]$ (bias $\pm$ std)
CCA	0.143 ± 0.010	0.100 ± 0.008
KNN	0.065 ± 0.003	0.019 ± 0.003
MICE	0.052 ± 0.004	0.011 ± 0.004
MissForest	0.013 ± 0.016	0.014 ± 0.015
ISA	0.004 ± 0.002	0.006 ± 0.003

TABLE F3 The simulation parameters for the first round of simulation.

Parameter	Value ranges
α	$[1, 10]$
β	$[1, 10]$
$(\theta_0, \theta_1, \theta_2)$	$[0, 2]^3$
σ_ϵ	$[0.1, 0.8]$
$(\beta_{11}, \beta_{21}, \beta_{12})$	$[-1, -0.3]^3$
(β_{10}, β_{20})	$[0.3, 1]^2$

APPENDIX G. MIMIC-III DATASET

MIMIC-III public database[20] is a comprehensive longitudinal database collected from over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Table G1 and Figure G1 provide the summary statistics and distributions of the variables MCH, MCV, HbA2 and IL for 38,542 patients (whose ages are higher than 10 years). For each patient, we averaged over multiple readings of each variable to obtain one value per patient. Apart from the missing values, we also considered the error codes, for example, ERROR as missingness. The result for the

T A B L E G1 Summary statistics for three study variables in MIMIC-III database.[20]

Variable	#-Miss.	Range	Mean (STD)
MCV	3858	57–141.4	89.65 (5.97)
MCH	3858	16.58–45.76	30.24 (2.27)
HbA2	38,465	0–5.4	2.64 (1.00)

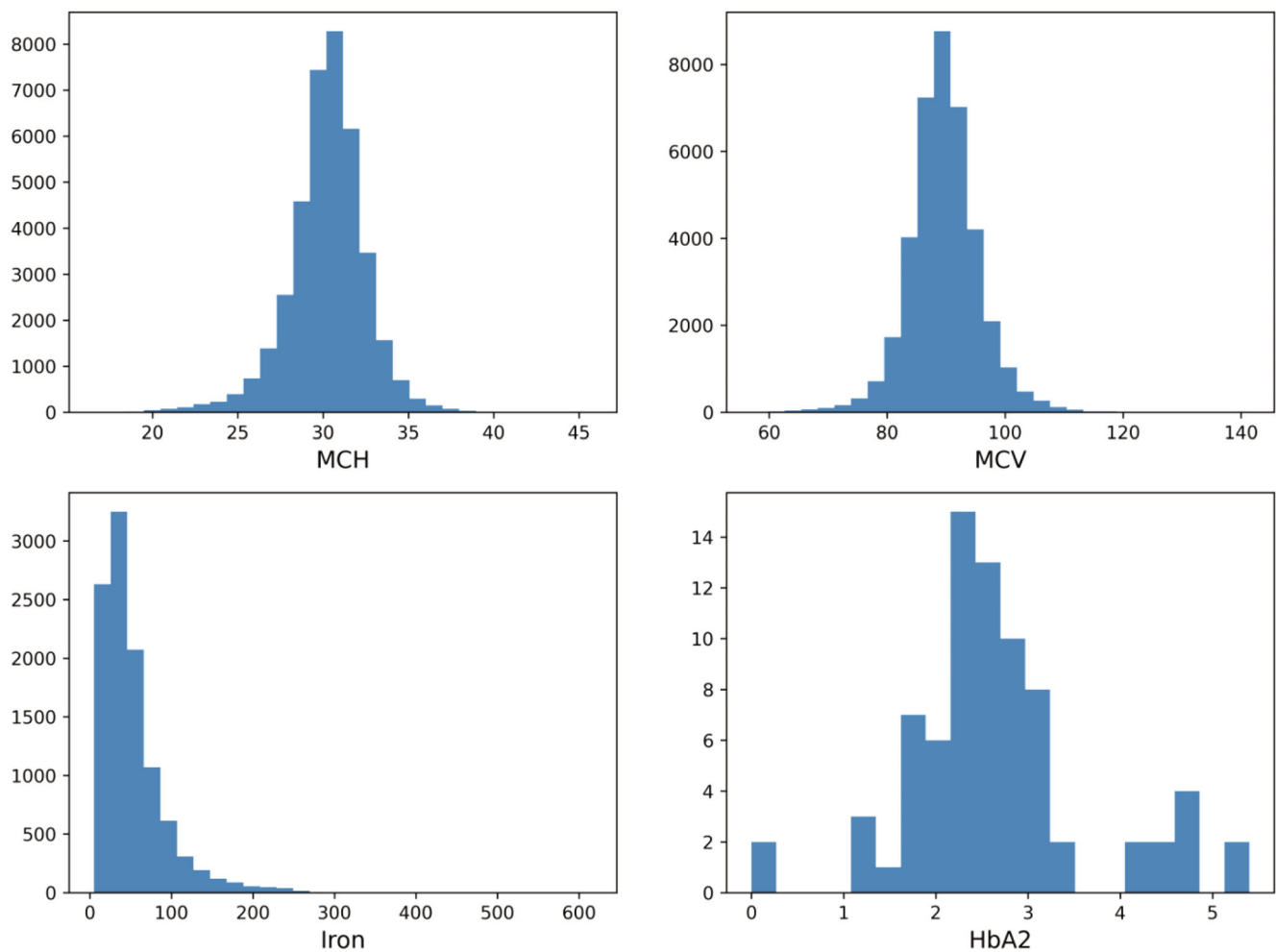*Note*: Range, mean and standard deviation values are calculated for observed cases.

F I G U R E G1 Distribution of MCV, MCH, IL and HbA2 in MIMIC-III database[20] for the cohort with age > 10.

percentages of missingness was 10.01% jointly for MCV and MCH, 89.83% for HbA2, and 76.74% for IL. Finally, we used the binary interpretation of HbA2, as suggested by the flowchart in Figure 1, based on the threshold of 3.6%.

APPENDIX H. IMPUTATION ALGORITHMS

The original pattern graph imputation is presented in Algorithm 2, with the partial imputation Algorithm 1. Modified imputation algorithm in the default sensitivity analysis is presented in Algorithm 3. The partial imputation algorithm is modified under edge-wise sensitivity analysis (in ISA) as in Algorithm 4. This modified partial imputation algorithm shall be utilized with the original imputation Algorithm 2.

Algorithm 1. Partial-imputation algorithm for pattern graphs1

Inputs: variables l_r; the pattern r is determined by the input l
Sample a random pattern S from the pattern set $\text{pa}_{\text{PG}}(r)$ with probability

$$p(S = s) = \frac{\hat{p}(l_r|R = s)n_s}{\sum_{\tau \in \text{pa}_{\text{PG}}(r)} \hat{p}(l_r|R = \tau)n_\tau},$$

where $n_s = \sum_{i=1}^{n} I(R_i = s)$.
Impute the components in $s - r$ by sampling from the conditional density:

$$L_{s-r}^{\dagger} \sim \hat{p}(l|L_r = l_r, R = s).$$

Return $L_s = (L_{s-r}, L_r)$

Algorithm 2. Pattern graph algorithm for imputing the entire data 1

1: **Requires** estimators $\hat{p}(l_r|R = r)$; a regular pattern graph $\mathcal{G}$
2: **for** $i = 1, \ldots, n$ **do**
3:     Set $L_{\text{now}} = L_{i,R_i}$ and $R_{\text{now}} = R_i$.
4:     Execute Algorithm 1 with inputs L_{now} and R_{now}.
5:     Update $L_{\text{now}}, R_{\text{now}}$ to be the return of the previous step.
6:     **if** $R_{\text{now}} \neq \mathbf{1}_d$ **then**
7:         return to 2.
8:     **else**
9:         update $L_i = L_{\text{now}}$.
10:     **end if**
11: **end for**
12: **Return** imputed $\hat{L}$.

Algorithm 3. Pattern graph imputation algorithm with rejection sampling step 1

1: **Requires** estimators $\hat{p}(l_r|R = r)$; a regular pattern graph $\mathcal{G}$
2: **for** $i = 1, \ldots, n$ **do**
3:      Set $L_{\text{now}} = L_{i,R_i}$ and $R_{\text{now}} = R_i$.
4:      Execute Algorithm 1 with inputs L_{now} and R_{now}.
5:      Update $L_{\text{now}}, R_{\text{now}}$ to be the return of the previous step.
6:      **if** $R_{\text{now}} \neq \mathbf{1}_d$ **then**
7:          return to 2.
8:      **else**
9:          draw $V \sim \text{Unif}(0, 1)$
10:          **if** $V \leq \frac{\exp(\omega_r^\top L_{\text{now},r})}{\exp(\omega_r^\top \mathcal{U}_r)}$ **then**
11:              update $L_i = L_{\text{now}}$;
12:          **else**
13:              return to 3
14:          **end if**
15:      **end if**
16: **end for**
17: **Return** imputed $\hat{L}$.

Algorithm 4. Partial-imputation algorithm with edge-wise sensitivity analysis

---

1: **Inputs:** variables l_r; the pattern r is determined by the input l; the tilting parameters ω for edges $r \leftarrow \text{pa}_{\text{PG}}(r)$
2: Sample a random pattern S from the pattern set $\text{pa}_{\text{PG}}(r)$ with probability

$$p(S = s) = \frac{\hat{p}(l_r|R = s)n_s}{\sum_{\tau \in \text{pa}_{\text{PG}}(r)} \hat{p}(l_r|R = \tau)n_\tau},$$

    where $n_s = \sum_{i=1}^{n} I(R_i = s)$.
3: Impute the components in $s - r$ by sampling from the conditional density:

$$L_{s-r}^\dagger \sim \hat{p}(l_{s-r}|L_r = l_r, R = s).$$

4: draw $V \sim \text{Unif}(0, 1)$
5: **if** $V \leq \frac{\exp(\omega_s^\top L_{s-r}^\dagger)}{\exp(\omega_s^\top \mathcal{U}_{s-r})}$ **then**
6:      update $L_{s-r} = L_{s-r}^\dagger$;
7: **else**
8:      return to 3
9: **end if**
10: **Return** L_{s-r}