Analysis of Missingness Scenarios for Observational Health Data

Alireza Zamanian *, Henrik von Kleist, Octavia Andreea Ciora, Marta Piperno, Gino Lancho, Narges Ahmidi

Posted Date: 5 April 2024

doi: 10.20944/preprints202404.0429.v1

Keywords: Missing Data Analysis; Observational Health Data; Missingness Scenarios; Missing Data Assumptions; Missingness distribution shift

Article

Analysis of Missingness Scenarios for Observational Health Data

Alireza Zamanian [1,2,*], Henrik von Kleist [1,3], Octavia-Andreea Ciora [2], Marta Piperno [2], Gino Lancho [2] and Narges Ahmidi [2]

[1] TUM School of Computation, Information and Technology, Department of Computer Science, Technical University of Munich, Munich, Germany;
[2] Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany;
[3] Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany.
* Correspondence: alireza.zamanian@iks.fraunhofer.de

Simple Summary: This paper argues the importance of considering domain knowledge when dealing with missing data in healthcare. We identify fundamental missingness scenarios in healthcare facilities and show how they impact analysis methods.

Abstract: Despite the extensive literature on missing data theory and cautionary articles emphasizing the importance of realistic analysis for healthcare data, a critical gap persists in incorporating domain knowledge into missing data problem formulation, assumption specification, and method development. In this paper, we highlight the gap particularly for observational data from healthcare facilities. We address this gap by identifying ten fundamental missingness scenarios arising during data measurement, recording, and pre-processing in observational health data, influenced by physicians, patients, healthcare facilities, and data scientists. We analyze the effect of scenarios on estimand formulation, missing data identification, estimation, and sensitivity analysis. To emphasize how domain-informed analysis can improve method reliability, we conduct simulation studies under the influence of various missingness scenarios. We compare the results of three common methods in medical data analysis (complete-case analysis, Missforest imputation, and inverse probability weighting estimation) for two estimands (variable mean estimation and classification accuracy). We advocate for our analysis approach as a reference for the analysis of observational health data. Furthermore, we posit that the proposed analysis framework is applicable to other medical domains, including medical wearable data analysis.

Keywords: missing data analysis; observational health data; missingness scenarios; missing data assumptions; missingness distribution shift

1. Introduction

Healthcare data encompasses a wide range of variables related to various diseases and health conditions collected from different facilities under the supervision of distinct and even contrasting guidelines. It is, therefore, naive to expect a medical dataset in which a sufficient number of informative variables are available for all patients. This is why data-driven research in healthcare almost always faces the challenges of missing data.

As two out of many existing approaches, data scientists may choose to simply discard the incomplete data samples and use only the complete ones for analysis (complete case analysis) or utilize complex deep learning models to fill in the missing entries and create a complete semi-synthetic dataset (imputation). In any case, the reliability of the methods depends on the nature of the missingness problem, e.g., how a variable distribution changes when the variable is observed or missed or why physicians decided to measure a variable for a patient and not the other. Inevitably, we must make assumptions about these questions, which are often not testable using the data itself [1,2]. Hence, we can trust the analysis result only if the assumptions are made explicit, the method conforms to them, and the sensitivity of results to departures are investigated [1].

Even though the scholarly discourse on missing data consistently highlights this concern [3–7], surprisingly, the applied machine learning (ML) research in healthcare often suffers from ambiguous

 doi:10.20944/preprints202404.0429.v1

missingness problem statements [6] and lack of transparent reports of assumptions for methods which are detached from the reality of healthcare facilities [4]. Complete-randomness of missingness is assumed even though observations in healthcare facilities are clearly conducted by guidelines and protocols [6,7]. Missing data methods are often chosen inconsistently when training and evaluating prediction models [8]. The problem intensifies as the survey and review articles on missing data in healthcare [5,9–12] are often limited to merely imputing the medical datasets without scrutiny of assumptions.

We believe the remedy is a guide for mapping data collection procedures and missing data scenarios in healthcare facilities to the missing data theory. Such a mapping could enable medical data scientists to collect necessary information about the problem, formalize it, and perform a domain-informed and reliable analysis. In this paper, we follow this purpose by identifying and categorizing ten fundamental missingness scenarios in observational healthcare data and analyzing them in the scope of missing data theory. Section 1.1 provides a more in-depth discussion about the research gap. Section 1.2 discusses the paper's contribution.

1.1. Research Gap

To highlight the gap between missing data theory and applied data analysis, we illustrate the ideal analysis process under missing data in Figure 1, as previously formalized in the literature [2,13]. The process begins by (1) formulating the objective (e.g., risk factors for a disease) as a mathematical expression, namely an estimand, which should be learned using incomplete data. (2) Next, assumptions are collected about the missingness status of the variables in the estimand, e.g., regarding how other variables influence missingness in a variable. (3) In the light of the assumptions, whether and how the estimand can be learned from incomplete data is determined. This step is referred to as *identification* (a.k.a., *recovery*). The next step is (4) to learn the estimand using data, e.g., by employing ML methods (*estimation*). The right method and learning setting for this step is also decided by assumptions. (5) To increase the reliability and robustness of the results, the results' sensitivity against model perturbations and violation of assumptions is measured (*sensitivity analysis*). Similar to step 3 and 4, this step is also influenced by the assumptions when choosing meaningful perturbation ranges according to the problem at hand.

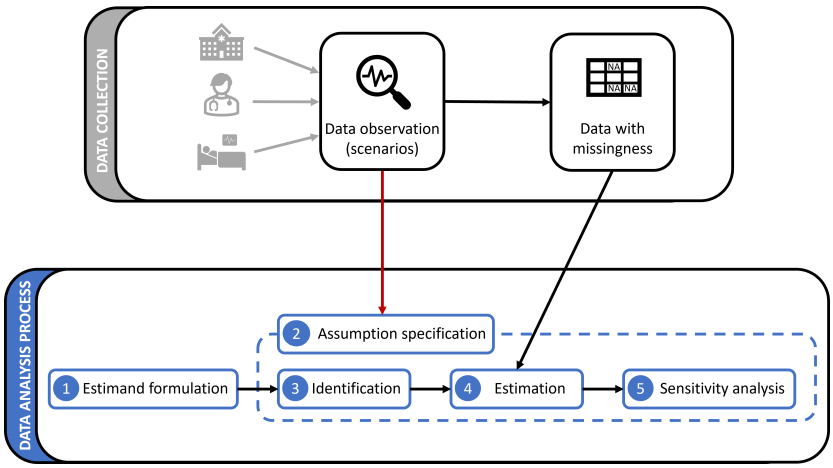


Figure 1. The ideal analysis process under missing data, consisting of (1) estimand formulation, (2) assumption specification, (3) identification, (4) estimation, and (5) sensitivity analysis. Table 1 presents supporting literature for each step. The red arrow, representing the influence of missingness scenarios on the assumption specification step, highlights the research gap.

Table 1 presents the examples for the related works supporting the analysis steps: (1) correct formulation of different estimands, (3) identification theory for missing data problems, (4) efficient estimation for missing data, and (5) sensitivity analysis. Regarding step 2 (assumption specification),

 doi:10.20944/preprints202404.0429.v1

cautionary articles exist for advocating domain-informed analysis [3,5,7], and providing high-level strategies for correct interpretation and reporting of missingness scenarios [4,14]. Several papers have performed such domain-informed analyses (for example, Mirkes et al. [15] and Millard et al. [16]), yet only for specific diagnoses (trauma and COVID-19), limited to unique missingness complications (selection bias) and estimands (outcome prediction). To our knowledge, no comprehensive taxonomy and analysis of missingness scenarios exists for observational data in healthcare facilities. Our paper is inspired by Moreno-Betancur et al. [17] and Marino et al. [18], developing general guidelines for treating missing data in epidemiology and clinical studies.

Table 1. Examples of literature supporting the steps of the missing data analysis process, presented in Figure 1.

Step	Subject matter	References
1	Estimand formulation under missing data	[2,19–22]
2	Domain-informed missing data formulation and assumptions*	[15,16]
3	Missing data identification theory	[2,23–28]
4	Estimation with missing data	[20,21,29–32]
5	Sensitivity analysis for unidentifiable missingness	[33–36]

* Research gap.

1.2. Contribution

Addressing the research gap in Table 1, this paper falls under the translational research category. We mainly focus on missingness scenarios in healthcare facilities such as clinics and hospitals. We exclude planned clinical studies, as they comprise considerably different scenarios (e.g., case drop-out and planned missingness). We achieve the goal of the paper through the following steps:

- We introduce ten fundamental and prevalent scenarios in healthcare facilities that lead to observation, recording, or missingness of data. Table 2 gives an overview of the scenarios.
- To make the application-theory connection, we introduce theoretical inquiries to be made about each missingness scenario. Table 3 gives an overview of the inquiries.
- For each scenario, we make the inquiries above and analyze the theoretical implications, along with various examples from the medical data analysis literature.
- To demonstrate the effect of domain-informed assumption on the method reliability, we perform a simulation study, showing how domain-agnostic analysis may lead to different levels of bias depending on the active scenarios and different estimands.

We propose this paper as a reference point for correct data analysis and reporting using observational health data. The methodology of this paper is not limited to the selected missingness scenarios and can be applied to less common yet equally essential scenarios that the readers may encounter. For the scope of the paper, we mainly focus on the assumptions and their interpretation without going into detailed discussions about the implementation of missing data algorithms and methods. Nevertheless, when required, we provide sufficient references to influential works in the missing data literature.

 doi:10.20944/preprints202404.0429.v1

Table 2. Overview of missingness scenarios in healthcare facilities, analyzed in this paper.

Ref.	Title
	Scenarios related to patients
1	Patient complete non-visit
2	Missing follow-up visit due to health status
3	Missing measurements due to health-related events during hospitalization
4	Missing measurements due to patient's refusal
	Scenarios related to physicians
5	Missing measurements due to diagnostic irrelevance
	Scenarios related to healthcare facilities
6	Missing measurements outside protocols requirements
7	Unavailability or shortage of resources
8	Unrecorded observations
	Scenarios related to data pre-processing
9	Omission of data samples based on inclusion/exclusion criteria
10	Omission of invalid data entries

Table 3. Overview of theoretical inquiries for missingness scenarios analyzed in this paper.

Ref.	Description
	at identification step
1	What missingness mechanism is induced by a scenario
2	Whether a scenario is subjected to missingness parametric distribution shift
3	Whether a scenario permits no-direct-effect assumption
4	Whether a scenario permits no-interference assumption
5	Whether a scenario induces selection bias
	at estimation step
6	Whether a scenario induces monotone missingness patterns
	at sensitivity analysis step
7	Whether a scenario gives informed guesses about sensitivity parameters

1.3. Structure

The rest of the paper is structured as follows: Section 2 introduces the missingness scenarios in healthcare facilities. In Section 3 we review the missing data theory, highlight the inquiries about missingness scenarios, and analyze each for all the introduced scenarios in Table 2. An empirical investigation is presented in Section 4. Discussions and conclusion are presented in Section 5.

2. Missingness Scenarios in Healthcare Data

Observational data from healthcare facilities, such as clinics and hospitals, comprises information about outpatient and inpatient (hospitalization) visits. Variables in healthcare data include patient demographic information (e.g., age and gender), medical history, signs and symptoms (e.g., blood pressure value and pain symptom), lab tests (e.g., blood chemistry test), diagnoses, and prescribed medications. The variable list extends for inpatient visits, including higher resolution observations and prescription information (e.g., oxygen saturation from bedside monitoring, and the input/output chart). In addition, new variables are collected during different hospitalization modes, such as ICU

 doi:10.20944/preprints202404.0429.v1

hospitalization. For more information about collected observations in healthcare facilities, see the documentation of publicly available datasets such as MIMIC-IV electronic health record dataset [37].

Health variables are collected under various scenarios that are influenced by the healthcare facilities, physicians, patients, and several medically related and unrelated factors. It is crucial to investigate if, for example, a variable is intentionally not measured and unrecorded based on the physician's opinion about the medical case, or unintentionally due to unexpected software issues. In this section, we explore ten fundamental and prevalent scenarios that drive data measurements and recordings in healthcare facilities. For each scenario, we provide real-world examples within the text, as well as in Appendix A. The majority of examples are extracted from the clinical prediction model (CPM) literature, introduced by Tsvetanova et al. [8].

2.1. Scenarios Related to Patients

By default, data is only collected during a patient's visit to a healthcare facility, resulting in a gap in data for the time between visits. At the same time, the sub-population that has not visited the healthcare facility is not observed at all. These are situations where we inevitably encounter the missing data problem unless additional data sources are used to complement the primary dataset.

It is naive to assume that visits are decided randomly and unrelated to the patient's health condition. This motivates the following key missingness scenario:

> **Scenario 1** (*Patient complete non-visit*)**:** Sub-populations with no healthcare facility visits during data collection are not included in the dataset.

In scenario 1, a specific sub-population is completely missing, e.g., due to health status. For instance, healthy people with no serious health complications infrequently visit clinics with different intentions such as preventive check-ups. Likewise, the data of patients deceased before any visit (e.g., dead-on-arrival) is often absent from the facility database. Other factors, such as socioeconomic status, can also influence the non-visit. References to scenario 1 are presented in Appendix A.1.

Scenario 2 describes another type of non-visit, namely missing follow-ups for patients with at least one recorded visit.

> **Scenario 2** (*Missing follow-up visit due to health status*)**:** Patients may miss a follow-up visit due to death, facility transfer, or if they decide not to continue the treatment.

The difference between scenarios 1 and 2 lies mainly in the reasons for missingness; patients have potentially different reasons not to visit a healthcare facility for the first time, or to drop the follow-up visits, possibly despite the physician's recommendations. References to scenario 2 are presented in Appendix A.2.

Health status factors influence not only the visits but also measurements during the visits. As highlighted by scenario 3, location transfer within the facility due to health conditions influences the observed variables.

> **Scenario 3** (*Missing measurements due to health-related events during hospitalization*)**:** Observations may be interrupted or limited by extreme health conditions or the transfer to a different location.

In scenario 3, observations may be interrupted due to events such as the occurrence of *code blue* and the resulting disconnection of devices for resuscitation [11,38], or patient transfer, e.g., to the operation room or ICU ward [14]. These events may also lead to observing new health variables that have not been recorded prior to the event, e.g., monitoring during operation [37]. References to scenario 3 are presented in Appendix A.3.

Another reason for missingness in a variable is the patient's refusal to take a test or consent to data sharing, as stated by scenario 4.

 doi:10.20944/preprints202404.0429.v1

> **Scenario 4** (*Patient's refusal*)**:** Patients may actively refuse specific observations or decline consent to data sharing.

Overall, patients' personal decisions, whether for medical (e.g., pain intolerance) or non-medical (e.g., fear of examination) reasons, may induce missingness. References to scenario 4 are presented in Appendix A.4.

2.2. Scenarios Related to Physicians

During a visit, the attending physicians decide which variables to observe. Bickley and Szilagyi [39] describe the examination and diagnosis practice as a step-by-step process in which physicians use basic observations such as vital signs and symptoms to form a first diagnostic belief, referred to as *impression*. To prove or rule out the possible diagnoses, physicians then order more specific, expensive, and sometimes invasive tests. Therefore, as scenario 5 states, the primary reason for taking or skipping a measurement is the diagnostic information it provides, compared to the cost of observation (e.g., monetary, time, harm to the patient.).

> **Scenario 5** (*Missing measurements due to diagnostic irrelevance*)**:** Variables that are less relevant to the physician's impressions are less likely to be observed.

Scenario 5 concerns diagnostic flowcharts and score systems in a dataset (see Elovic and Pourmand [40]). These provide rules for selecting the following observation until the final diagnosis. Nevertheless, observation patterns may not entirely reflect one particular guideline, as many guidelines are used during the data collection phase within a cohort, and other scenarios also affect the data.

One should note the implications and subtle differences between these tools when conducting a missingness analysis. For example, in flowcharts, the value range for the parent variable(s) determines the next observation. In contrast, in a score system, the cumulative score of all related variables determines whether more observations are necessary for concluding the decision [41]. References to scenario 5 are presented in Appendix A.5.

2.3. Scenarios Related to Healthcare Facilities

Measurement decisions are not only determined by physicians but also by protocols and guidelines in healthcare facilities, as stated by scenario 6.

> **Scenario 6** (*Missing measurements outside protocol requirements*)**:** Data collection protocols decide the measurements in different conditions during hospitalization.

For instance, hospital protocols may mandate specific data (e.g., demographic information, basic blood tests) to be collected upon admission. Similarly, there are measurements only performed in particular conditions, e.g., pre- and post-surgical measurements. It is, therefore, crucial to consider the role of protocols within healthcare facilities when investigating the causes of missingness or observation of variables. References to scenario 6 are presented in Appendix A.6.

Scenarios 5 and 6 assume that measurements can always be conducted if required. While this may generally be true, especially for routine tests, a measurement may sometimes be hindered by the unavailability or shortage of necessary resources. Diagnostic tests may be dropped or delayed for a patient due to prioritization in long waiting queues or temporary unavailability of equipment or staff. Scenario 7 describes the situation where measurement orders were not realized despite physicians' decisions.

> **Scenario 7** (*Unavailability or shortage of resources*)**:** The physician's order for observation may not be realized due to unavailability or shortage of resources.

References to scenario 7 are presented in Appendix A.7.

 doi:10.20944/preprints202404.0429.v1

Another assumption for scenarios 5 and 6, which does not always hold is that the measurements and physicians' observations are all recorded in the database. As stated by scenario 8, variables might be observed and influence medical decisions, yet they are withheld from the dataset.

> **Scenario 8** (*Unrecorded observations*)**:** Some variables are not recorded in the database or used in the data analysis, even though they have been observed and relied upon in medical practice.

There might be aspects characterizing the overall patient's health, which are not explicitly recorded but implicitly considered in the decision-making process. In addition, certain modalities of data may not be efficiently recorded or integrated into the medical record. Further, some modalities, such as textual data, may be excluded from data analysis due to complexity. All these reasons are mainly determined by the quality of data collection software in healthcare facilities, the physician's style of practice, and the choice of data modalities for analysis. References to scenario 8 are presented in Appendix A.8.

2.4. Scenarios Related to Data Pre-Processing

For the analysis of collected and recorded data, the first step is dataset selection and pre-processing. Depending on the analysis objective, data scientists apply inclusion-exclusion criteria based on demographic information, patient cohort, or variable availability. As stated by scenario 9, this should be considered a missingness scenario induced in the data analysis step.

> **Scenario 9** (*Data sample omission based on inclusion/exclusion criteria*)**:** Samples are included or excluded depending on data and missingness characteristics, such as measurement availability, values within a specific range, or patient cohort.

References to scenario 9 are presented in Appendix A.9.

Another common reason for data omission during pre-processing is the presence of invalid, unextractable, or erroneous values, as stated by scenario 10.

> **Scenario 10** (*Missingness of invalid data entries*)**:** Data rows with invalid or erroneous entries are removed from the data during data pre-processing.

Examples are omission due to poor handwriting or corrupted medical chart pages [14], negative age values, or entries specified by `ERROR` code. References to scenario 10 are presented in Appendix A.10.

3. Analysis of Missingness Scenarios

This section presents the foundation of the missing data theory necessary for analyzing the introduced scenarios in Section 2. After briefly explaining the missing data problem formulation, we describe the steps required for solving the problem under missing data, according to Figure 1. Throughout the steps presentation, we identify theoretical questions to answer for each scenario to bridge the gap between the analysis and application domains. We call these questions *inquiries* about the scenarios. Extensive details for the inquiries are presented in Appendix C.

3.1. Setting and Notation

Let the random vector $X \in \mathbb{R}^d$ comprise d study variables $X_i \in X$, $i \in \{1, \dots, d\}$. For ease of reference, we denote a specific variable of interest beside X (e.g., the class labels in the ML classification problem) as Y. Furthermore, we denote independence between X_i, X_j as $X_i \perp\!\!\!\perp X_j$. Independence by conditioning on a variable X_h is denoted as $X_i \perp\!\!\!\perp X_j | X_h$.

In reality, the variable X_i is realized for all patients, though it may or may not always be available(i.e., observed, recorded, and present in the dataset). We therefore refer to X_i as a *counterfactual* variable since this is what the data would have been if it had always been available, possibly contrary to reality. Corresponding to each X_i, we define a binary variable $R_i \in \{0, 1\}$, called the *missingness indicator*, to express X_i's availability: we set $R_i = 1$ when X_i is available, and $R_i = 0$ otherwise. The

 doi:10.20944/preprints202404.0429.v1

version of X_i which is masked by missingness is called *proxy variable*, denoted as $X_i^* \in \mathbb{R} \cup \{\texttt{NaN}\}$ where NaN represents the missing entries. By this definition, a proxy variable is modeled as

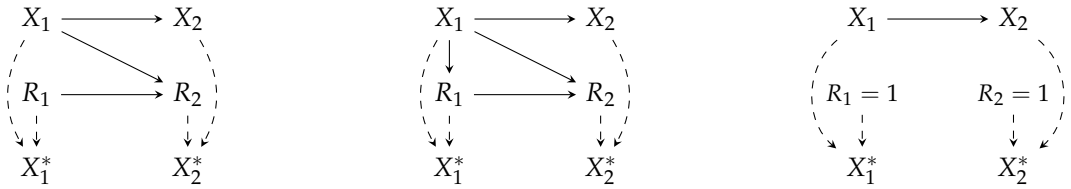
$$X_i^* = \begin{cases} X_i, & R_i = 1, \\ \texttt{NaN}, & R_i = 0. \end{cases} \tag{1}$$

The distribution of R is determined by the subset of scenarios from Section 2, which describe the data observation and recording in a healthcare facility and dataset selection for analysis. A data availability policy π represents the union of scenarios, such that the missingness distribution follows the policy, i.e., $R \sim \pi$. Subsequently, the resulting distribution given a policy π is denoted as p_π. We denote three special policies to reference in the paper: (1) the initial policy during data collection as π_{init}, (2) the full-availability policy as π_{full}, under which X is always available, and (3) any other policy as π_{new}, which is neither π_{init} nor π_{full}. This notation yields $p(X) = p_{\pi_{\text{full}}}(X^*|R = 1)$.

Example 1 (missingness under availability policies)**.** *Suppose a variable X_1 is realized for four patients, giving $X_1 = (2, 3, 4, 7)^\top$. If the fourth patient has missing values under a policy π_{init}, we have $X_1^* = (2, 3, 4, \texttt{NaN})^\top$. In this case, the mean estimations for X and X^* are defined as $\mathbb{E}_{\pi_{init}}(X_1^*|R = 1) = 3$ and $\mathbb{E}(X) = \mathbb{E}_{\pi_{full}}(X_1^*|R = 1) = 4$. Under a new policy π_{new} where only the $X \geq 4$ observations are available, we have $\mathbb{E}_{\pi_{new}}(X_1^*|R = 1) = 5.5$*

An availability policy for X_i is, in general, parameterized by all variables (including X_i itself) as well as other missingness indicators, i.e., $\{X, R_{\setminus i}\}$. To encode these dependencies, we model the joint (X, R) distribution using m-graphs Mohan et al. [2]. An m-graph under the availability policy π, denoted as $\mathcal{G}_\pi(V)$, is a causal directed acyclic graph (DAG) with the node set $V = \{X \cup X^* \cup R\}$. The edges in the structure $X_i \to X_i^* \leftarrow R_i$ are deterministic, representing equation 1. While non-graphical approaches for missing data exist, we focus on m-graphs for their effectiveness and popularity in this paper. Section 3.3 will provide more details about m-graphs.

Example illustrations of m-graphs are depicted in Figure 2, where three m-graphs model different policies for a similar (X_1, X_2) distribution.



(a) π_{init} **(b)** π_{new} **(c)** π_{full}

Figure 2. Three example m-graphs, modeling the joint distribution (X, X^*, R) for a bivariate dataset: (a) $(R_1, R_2) \sim (\pi_{\text{init},1}, \pi_{\text{init},2}(R_1, X_1))$ represents the missingness distribution induced by missingness scenarios at data collection; (b) $(R_1, R_2) \sim (\pi_{\text{new},1}(X_1), \pi_{\text{new},2}(R_1, X_1))$ represents a new missingness distribution due to a change in the scenarios for R_1; $(R_1, R_2) = (1, 1)$ represents the full-availability case (no missingness scenario). Dashed edges are deterministic, encoding the definition in equation 1.

3.2. Defining the Estimand

In the first step of data analysis, an objective must be set by the domain expert and the data scientist and translated into an estimand, which will be fitted to the data. Examples include finding the weights of a prediction model for patient morbidity or the mean value of a biomarker for a population. Based on the form of the estimand, and whether and how it depends on the unavailable data distribution under missingness, we may face diverse challenges.

A basic question of interest is the mean of an outcome variable Y (e.g., the mean value of a test or the chance of recovery). If Y is partially available under the policy π_{init}, one may formulate the question directly as $\mathbb{E}_{\pi_{\text{init}}}(Y^*|R_Y = 1)$, which reads as "mean of Y when it is available". However,

 doi:10.20944/preprints202404.0429.v1

we are often interested in estimating the entire population regardless of the missingness status "had Y for all samples been available for analysis." This objective, referred to as the counterfactual mean estimation, is presented as

$$\mathbb{E}(Y) = \mathbb{E}_{\pi_{\text{full}}}(Y^*|R_Y = 1). \tag{2}$$

Example 2 (counterfactual mean LDL cholesterol level)**.** *As part of public health research, we aim to estimate the nationwide average LDL cholesterol level, denoted as Y. Available datasets are collected from a hospital where LDL levels are not available for all the patients.*

- $\mathbb{E}_{\pi_{init}}(Y^*|R_y = 1)$ *gives the average observed value in the hospital.*
- *As a possible new policy,* $\mathbb{E}_{\pi_{new}}(Y^*|R_y = 1)$ *gives the average value if the LDL level had been observed for all patients in the hospital.*
- $\mathbb{E}(Y) = \mathbb{E}_{\pi_{full}}(Y^*|R_y = 1)$ *gives the target estimand, i.e., the nationwide average LDL level.*

As a more advanced objective, we may be interested in developing a prediction model for the outcome variable Y using the covariate vector X. i.e., $\mathbb{E}(Y|X = x)$, which reads as "conditional mean of Y given X". We often choose an ML model for estimation, such as a multi-layer perceptron neural network $f(x; w)$, parameterized by w. The weights of the network are learned by minimizing a loss function, e.g., the mean squared error (MSE): $\mathbb{E}[(y - f(x; w))^2]$. Model performance at deployment can also be evaluated using the same formula.

Given a fully observed outcome and missing covariates, the estimand

$$\mathbb{E}_{\pi_{\text{init}}}\left[(y - f(x^*, r_X; w))^2\right] \tag{3}$$

formulates the MSE loss for the available X. Estimand 3 suits the situation where the prediction model is to be deployed in an environment with the same observation policy, meaning that all missingness scenarios are the same during deployment as during the data collection stage. In estimand 3, we may use the information in R_X, e.g., we train (maximum) 2^d separate sub-models $g(x_j), R_j = 1$ for each unique value (pattern) of R [20].

Example 3 (health status estimation at hospital discharge)**.** *We aim to develop a prediction model for the 6-month outcome based on the observed variables during hospitalization, queried at discharge. The model deployment will not influence the physicians' decisions. The fact that the hospitalization data is being analyzed retrospectively can justify the assumption that the observation and recording policy will not change at deployment. The MSE loss for this case is given by estimand 3.*

Alternatively, we may be interested in learning a prediction model that is deployed in healthcare facilities with different missingness scenarios, e.g., with varying guidelines of observation and protocols (scenario 5 and 6), for a different patient cohort (scenario 1 and 9), or in the same healthcare facility, but with a change of observation policy because the physicians would measure different variables to "feed" it to the prediction model. In particular, suppose a training dataset generated given the m-graph in Figure 2a will be deployed in an environment modeled by the m-graph in Figure 2b. The estimand for such a case is

$$\mathbb{E}_{\pi_{\text{new}}}\left[(y - f(x^*, r_X; w))^2\right], \tag{4}$$

which reads as "MSE loss under new missingness scenarios at deployment," where π_{new} represents the new policy.

Example 4 (change of hospital discharge protocols)**.** *Suppose the hospital in example 3 adopts a new discharge protocol mandating performing a medical test for all patients before discharge. The MSE loss under the newly adopted policy is given by estimand 4*

 doi:10.20944/preprints202404.0429.v1

A special case of estimand 4 is when the prediction model is expected to make predictions always using full covariates (Figure 2c). The estimand for this case is $\mathbb{E}_{\pi_{\text{full}}}[(y - f(x;w))^2]$, with only one missingness pattern, the full-availability $R = \vec{1}$. This objective is employed for most clinical prediction models (see Tsvetanova et al. [8]). For more examples, Appendix B presents the estimands for prediction using decision trees and feature importance.

Example 5 (clinical prediction model)**.** *Suppose a clinical prediction model is developed using an incomplete dataset. As a result of successful development, physicians will use the model while they actively collect all study variables every time to feed to the model. The MSE loss at deployment is given by* $\mathbb{E}_{\pi_{full}}[(y - f(x;w))^2]$.

3.3. Identification

As shown in the previous step, estimands may query different missingness distributions, while the only available distribution is given by the data collection policy π_{init}. If an estimand queries π_{init}, such as estimand 3, it can be computed directly using the training dataset $\mathcal{D}$. On the other hand, estimands such as 2 and 4 query different distributions and hence are subjected to the distribution shift problem. In the identification step, we find a procedure that computes a consistent estimate of an estimand under a target distribution using the available π_{init} [2].

To elaborate further, we consider an estimation approach under distribution shift, namely *importance sampling*: for a functional θ of the distribution at deployment $q(X,R)$, θ is estimated using the data collection distribution $p(X,R)$ as

$$\int \theta(x,r) \cdot q(x,r)dxdr = \int \theta(x,r) \cdot \lambda(x,r)p(x,r)dxdr, \tag{5}$$

where the fraction $\lambda(X,R) = p(X,R)/q(X,R)$ is called the importance ratio. By equation 5, samples are drawn from $p(.)$ but re-weighted by their "importance" in reflecting $q(.)$. Equation 5 states that a θ estimation is possible given the $p(X,R)$ samples when λ is known for all (x,r) over the support of p.

The importance ratio can be re-written using the selection model factorization [42], as

$$\lambda(X,R) = \frac{q(X)q(R|X)}{p(X)p(R|X)}. \tag{6}$$

The conditional terms $p(R|X)$ and $q(R|X)$ in the fraction are the data collection and deployment availability policies, respectively. Assuming no additional counterfactual data distribution shift, i.e., $p(X) = q(X)$, equation 6 is simplified as $\lambda(X,R) = q(R|X)/p(R|X)$, i.e., the ratio of missingness models at the data collection and deployment stages. When the availability policy does not change at deployment, the ratio is re-written as

$$\lambda(X,R) = \frac{p_{\pi_{\text{init}}}(R|X)}{p_{\pi_{\text{init}}}(R|X)} = 1, \tag{7a}$$

and when a new policy is adopted at deployment, it is re-written as

$$\lambda(X,R) = \frac{p_{\pi_{\text{new}}}(R|X)}{p_{\pi_{\text{init}}}(R|X)}. \tag{7b}$$

While the following arguments are valid for equation 5 in general, we consider a special case where the full-availability policy π_{full} is running at deployment (e.g., estimand 2). In this case, we trivially have $\lambda(X,R) = 0$ for all incomplete data, since the numerator $p_{\pi_{\text{full}}}(X,R)$ is zero when $R \neq \vec{1}$. This

 doi:10.20944/preprints202404.0429.v1

means that only the complete cases (R=1) are used for computation, for which $\lambda = 1/p_{\pi_{\text{init}}}(R = 1|x)$. The resulting estimator according to equation 5, is expressed (for estimation using $\mathcal{D}$) as

$$\hat{\theta}_{\text{IPW}} = \frac{1}{N} \sum_{X,R \sim \mathcal{D}} \theta(x, r) \cdot \frac{\mathbf{1}(r = 1)}{p_{\pi_{\text{init}}}(R = 1|x)}, \tag{8}$$

for N samples, where $\mathbf{1}(r = 1)$ selects only the complete cases. Equation 8 is known as the *inverse-probability weighting* (IPW) estimator. The denominator in equation 8 is referred to as the *propensity score*, often denoted as $\text{PS}(x)$.

The challenge of identification lies in the conditioning set of importance ratio terms, as they generally depend on the counterfactual distribution (X), which is only partially available. As a solution, we assume an m-graph for the problem and seek independence properties among (X, R) variables that allow us to express the importance ratio in terms of factors that can be estimated using the available distribution (X^*, R)[1] [2,43].

Example 6 (identification w.r.t. an m-graph)**.** *Suppose a functional $\theta(X_1, X_2)$ is to be estimated, given the data collection and deployment policies π_{init} and π_{full}, respectively. The propensity score for IPW estimator is $p_{\pi_{init}}(R_1 = 1, R_2 = 1|X_1, X_2)$ which cannot be estimated using $\mathcal{D}$. Assuming the m-graph in Figure 2a, we proceed as follows (we drop the distribution index for brevity):*

- *factorize: $PS(X_1, X_2) = p(R_1 = 1|X_1, X_2)p(R_2 = 1|R_1 = 1, X_1, X_2)$*
- *the assumed m-graph gives $R_1 \perp\!\!\!\perp X_1, X_2$ and $R_2 \perp\!\!\!\perp X_2|R_1, X_1$. The propensity score is thus rewritten as $p(R_1 = 1)p(R_2 = 1|R_1 = 1, X_1)$*
- *By the missingness definition in equation 1, we express the second term using the proxy variable and rewrite the propensity score as $p(R_1 = 1)p(R_2 = 1|R_1 = 1, X_1^*)$.*

Both factors in the propensity score can be estimated using (X^*, R)

In conclusion, identification in this manner requires an m-graph model, and within it, the causal relations of the missingness indicators are specifically important. It is therefore necessary to discover what kind of causal structures the missingness scenarios induce for R. In particular, we specify the parents and ancestors (direct and indirect causes) for the R nodes, as stated by inquiry 1. The causes of R nodes are commonly referred to as the missingness mechanism.

Inquiry 1 (*missingness mechanism*)**:** Which causes for R nodes a scenario implies.

To facilitate identifying the causes, we categorize all potential causes to search for, in the following three categories:

1. *(X and R components)* First candidates for causes of R are the study variables and their corresponding missingness indicators within the dataset. Examples can be found in Figures 2a and 2b, where X is a cause of indicator R.
2. *(latent/hidden confounders)* Variables that have not been collected and available in the dataset may also causally influence R. More importantly, they may confound two or more study variables within the estimand, and may therefore hinder the identification process.
3. *(exogenous causes)* Other variables that may lie outside the dataset and do not confound the study variables of interest, are considered exogenous causes, having, in general, no identification implications.

Missingness in health-related variables such as lab test items are mainly caused by physicians under scenario 5 (missing due to diagnostic irrelevance), where they make measurement decisions based

[1] For the scope of this paper, we mainly focus on identification with respect to m-graphs. See section 3 of Mohan et al. [2] for other identification approaches for missing data.

 doi:10.20944/preprints202404.0429.v1

on the observed history. Therefore, in this case, R indicators for health variables have incoming edges from the previous measurements (recorded or unrecorded). Other potential causes include the health status under scenarios 1 (patient complete non-visit) and 2 (missing follow-up visit due to health status). Examples of the latent/hidden confounders include socioeconomic variables as well as variables in secondary datasets with information about the non-visit population under scenario 1. As for the exogenous causes, many causes may be recognized, such as simply forgetting to enter the data for a patient, under scenario 8 (unrecorded observations). However, one should be cautious about treating all medically unrelated variables as exogenous causes, as they may still confound the study variables and missingness indicators. A detailed analysis of missingness scenarios with respect to inquiry 1 is presented in Table A1.

Inquiry 1 explores the structural distribution shift caused by a change of m-graph between the data collection and deployment stages. Another possibility is that the m-graph stays invariant, but the causal relations are subjected to a parametric shift. For example, assume the m-graph in Figure 2a holds for both data collection and deployment, but the missingness probability in X_2 changes from $\sigma(R_1 + 2X_1)$ to $\sigma(0.2R_1 + 5X_1)$. As stated by inquiry 2, it is crucial to explore the potential parametric shift at deployment due to a change in observation and recording policies.

> **Inquiry 2** (*Missingness distribution shift*)**:** Whether a scenario is subjected to missingness parametric distribution shift at deployment.

Parametric shift may occur in scenario 5 (missing due to diagnostic irrelevance), if the definition of normal/abnormal ranges for a health marker changes. In this case, the results of primary tests still influence the performing decision of later tests, however, via different rules. As another example, a parametric shift may occur in scenario 7 (missing due to resource unavailability), if the monetary cost of a medical test decreases as a result of equipment upgrade or insurance plans, leading physicians to order the test more often. A detailed analysis of missingness scenarios with respect to inquiry 2 is presented in Table A2.

Example 7 (parametric shift due to decreased test costs)**.** *Consider a primary test X_1 and a secondary and more expensive test X_2. Patients with abnormal primary test values ($X_1 > 5$) are more likely to give the X_2 test. After a cost reduction for the X_2 test, the overall frequency of the test ($R_2 = 1$) increases, such that now the relative number of tests for patients with abnormal X_1 values is ρ times larger than before; yet the association between X_1 and R_2 is retained. This statistics gives*

$$\frac{p_{\pi_{new}}(R_2 = 1|X_1 > 5)}{p_{\pi_{init}}(R_2 = 1|X_1 > 5)} = \rho,$$

which is the importance ratio for $(X_1 = 1, R_2 = 1)$ samples in equation 5. We leave it to the readers to calculate other importance ratios based on assumed statistics about this hypothetical problem.

So far, the described identification methodology has been based on the selection model factorization in equation 6 and the no-distribution-shift assumption for the counterfactual variables. However, there might exist missingness scenarios under which this assumption is violated. A case of violation is when the observation and measurement decisions directly affect the counterfactual variables. In terms of m-graphs, this translates to an $R \to X$ edge. The assumption that such a causal relation does not exist is referred to as *no-direct-effect* (NDE) [22], discussed in m-graph identifiability literature [25]. Since violation of NDE influences the identification procedure, it is crucial to know whether the problem setting permits it, as stated by inquiry 3.

> **Inquiry 3** (*no-direct-effect assumption*)**:** Whether a scenario implies outgoing edges from missingness indicators to counterfactual variables.

 doi:10.20944/preprints202404.0429.v1

A crucial case of NDE violation occurs when invasive tests such as biopsy affect the health status of patients. The effect of observation may be exerted on the corresponding counterfactual variable itself or other variables. This effect may also be exerted indirectly, e.g., through temporarily stopping a certain medication before a medical test. For example, due to the contraindication of radiology contrast agents and metformin, it is recommended for diabetic patients that the medication is stopped before performing angiography [44]. Note that under violation of the NDE assumption, the problem definition stated in Section 3.2 becomes ill-posed and requires further elaborations. An example of a problem definition under NDE violation is discussed in example 8. A detailed analysis of missingness scenarios with respect to inquiry 3 is presented in Table A3.

Example 8 (problem definition under NDE violation)**.** *Assume the following m-graph $X \rightarrow Y \leftarrow R_X$, describing an (X, Y) dataset with fully observed Y, where measurement of X negatively influences Y. This problem cannot be analyzed similarly to example 1, as the counterfactual realizations cannot be described ignoring the missingness status. Assuming a hypothetical data generation mechanism, the $X - Y$ relation follows $Y = wX + \epsilon$, $\epsilon \sim \mathcal{N}(0,1)$ in the absence of any measurement ($R_X = 0$). When X is measured ($R_X = 1$), Y distribution changes to $Y = wX + w_0 + \epsilon$. Therefore $\mathbb{E}_{\pi_{init}}(Y) \neq \mathbb{E}_{\pi_{full}}(Y)$. Possible questions to pose with regard to a target quantity θ are:*

- *if the observation policies remain unchanged,*
- *if we begin to always observe X,*
- *if we knew the value of X but without negative influences on Y, e.g., using a new testing technology.*

Another common assumption for missing data problem is the *no-interference* assumption, stating that the measurement decisions for one individual do not affect other individuals [22]. This is similar to the i.i.d. assumption in general ML problems: having interfered measurements, the i.i.d. assumption cannot be made for the R distribution. It is therefore important to check whether the no-interference assumption is permitted for observation scenarios, as stated by inquiry 4.

> **Inquiry 4** (*no-interference assumption*)**:** Whether a scenario causes interference among the availability status of data samples.

Similar to the NDE assumption, one may find realistic scenarios where the no-interference assumption is violated. In general, competing for limited resources under scenario 7 (unavailability or shortage of resources) or for available hospitalization services under scenario 1 and 2 (complete non-visit and missing follow-up) imply interference. A detailed analysis of missingness scenarios with respect to inquiry 4 is presented in Table A4.

Finally, we discuss a unique case of missingness, where data samples are completely omitted from the dataset prior to any analysis. This case can be modeled in m-graphs via an $R^\dagger$ node that influences all R_i, such that if $R^\dagger = 0$, then $R_i = 0, \forall i$ (Figure 3). The risk in this situation lies in the fact that we cannot infer the occurrence of such omissions from a dataset without additional information, which may thus lead to the wrong conclusion that the dataset is complete and free of missingness. This case is commonly referred to as *selection bias* in causal inference literature. Selection bias is argued in inquiry 5.

> **Inquiry 5** (*selection bias*)**:** Whether a scenario causes the omission of an entire data sample in the form of selection bias.

Clearly, sample omission can be a result of non-visit under scenario 1 and inclusion/exclusion criteria under scenario 9. Whether or not this should be conceived as a bias depends on whether the target parameter (e.g. Y in Estimand 2) is believed to vary between the observed and the unobserved sub-populations. A detailed analysis of missingness scenarios with respect to inquiry 5 is presented in Table A5.

 doi:10.20944/preprints202404.0429.v1

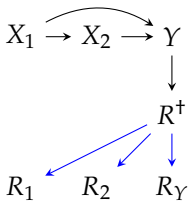


Figure 3. An example m-graph to model selection bias, determined by the Y values. Blue edges represent the deterministic masking relation between $R^\dagger$ and R_is.

3.4. Estimation

There are several methods for estimation with missing data, including likelihood-based methods such as Expectation-Maximization (EM) algorithm, multiple imputation (MI), IPW estimator, and outcome regression (OR) [27,42]. In the scope of this paper, we continue with the importance sampling approach in equation 5, in particular, the IPW estimator in equation 8 and the estimation of the propensity score.

Even though a successful identification step guarantees that the propensity score can be estimated using the available data, we still face some challenges, e.g., when the missingness pattern is *non-monotone*. A missingness pattern is called monotone if there is at least one ordering of the variables, such that observing the j-th variable ensures that all variables $k > j$ in the ordering are all observed for all samples (Figure 4a). Estimation of the propensity score has a straightforward solution for monotone patterns. Example 9 showcases propensity score estimation for identifiable monotone missingness.

Example 9 (Propensity score estimation for identifiable monotone missingness)**.** *For the missingness in Figure 4a, we have* $\sum_{j=1}^{4} p(S_j|X) = 1$*, while* $PS(X) \equiv p(S_1|X)$*. Assuming identifiability, each* $p(S_j|X)$ *can be estimated using only the variables available in* S_j*. As a result, the propensity score is estimated as*

$$PS(X) = 1 - p(S_2|X_1, X_2, X_3) - p(S_3|X_1, X_2) - p(S_4|X_1). \tag{9}$$

While methods have been developed for an effective estimation under non-monotone missingness [27,31], it is beneficial to adopt monotone solutions if applicable. In that regard, inquiry 6 argues whether a missingness scenario individually induces monotone missingness patterns.

> **Inquiry 6** (*monotonicity*)**:** Whether a scenario induces missingness with monotone patterns.

If an individual missingness scenario is active, monotonicity can be directly inferred from the emerged patterns, revealed by a simple sorting of the variables with respect to their missingness ratio (Figure 4a). However, in practice, several scenarios influence a dataset. In such cases, the monotone pattern attributed to one scenario is broken by other scenarios. If we can attribute the emerged non-monotone pattern to a dominant monotone-inducing scenario along with less effective non-monotone scenarios (hypothetically in Figure 4b), then methods exist based on resolving the missing entries up to recovery of the monotone pattern, e.g., via imputation, and proceeding with IPW estimation for monotone missingness [45]. A noteworthy scenario likely inducing monotonicity is the sequential observations of physicians under scenario 5 (missing due to diagnostic irrelevance). Given a specific diagnostic flowchart, it is reasonable to assume that more specific secondary tests shall not be made unless primary tests are done. As said, many reasons may break this pattern, including more than one diagnostic flowchart being used, other scenarios such as 4 (patient's refusal) or 7 (resource unavailability). A detailed analysis of missingness scenarios concerning inquiry 6 is presented in Table A6.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 April 2024 doi:10.20944/preprints202404.0429.v1

(a) Monotone pattern **(b)** Non-monotone pattern

Figure 4. Two monotone (a) and non-monotone (b) missingness patterns. The monotone pattern is described by four $S_1 - S_4$ patterns. One can infer, using prior knowledge, that the non-monotone pattern in (b) is a result of some non-monotone missingness scenarios, interrupting the monotone pattern in (a).

3.5. Sensitivity Analysis

The assumptions made for handling missing data may not hold under all circumstances. They might be too strong for practical implementation, or we may expect the environment to undergo some perturbations that violate them. To ensure the robustness of the analysis, it is crucial to measure the sensitivity of results to departures from the assumptions and report the variation. Sensitivity analysis is usually done by perturbing the m-graph model.

In addition, it is possible that due to the nature of the problem, assumptions do not lead to a successful identification. In this case, we may impose stronger assumptions that lead to identifiability, model the departures from the actual assumptions, and finally measure the sensitivity to different degrees of magnitude of those departures.

Example 10 (sensitivity analysis for the unidentifiable self-masking missingness)**.** *Consider an outcome variable Y that is subjected to missingness under the following mechanism $Y \to R_Y$. The estimand $\mathbb{E}(Y)$ is unidentifiable under this mechanism, referred to as self-censoring [25] or self-masking [20].*

We can assume the mean of the unobserved population is δ units away from the observed population, additively $\mathbb{E}_{\pi_{init}}(Y|R=0) = \mathbb{E}_{\pi_{init}}(Y^*|R=1) + \delta$, or multiplicatively $\mathbb{E}_{\pi_{init}}(Y|R=0) = \delta\mathbb{E}_{\pi_{init}}(Y^*|R=1)$ [6,33]. We then measure the variation of $\mathbb{E}_{\pi_{full}}(Y^*|R=1)$ assuming a range of values for the sensitivity parameter δ.

For reliable meaningful sensitivity analysis results, it is crucial to interpret the sensitivity parameters based on meaningful real-world quantities. Inquiry 7 states that scenarios may carry valuable information for choosing meaningful parameters.

> **Inquiry 7** (*meaningful sensitivity parameters*)**:** Given a scenario, what are the meaningful units and ranges of parameters for sensitivity analysis.

Specific to the importance sampling approach and equation 8, the unidentifiable terms appear in the importance ratio. The importance ratio captures the difference in the levels of availability for different covariate strata. To make an informed guess about this quantity, we may refer to other research works or collaborations with health domain experts. For instance, Zamanian et al. [36] suggest that the sensitivity parameters for physicians' observations (scenario 5) is related to the odds of making an observation for relatively healthy or sick patients, which can be inferred based on the guidelines, protocols, and referring to the attending physicians. The sensitivity parameter for this case

 doi:10.20944/preprints202404.0429.v1

is formulated for the model in [36], assuming a logistic model for missingness, by the following *Odds ratio term*,

$$\delta = \log \frac{O(R|\text{healthy})}{O(R|\text{sick})}, \tag{10}$$

where $O(R|X) = p(R = 0|X)/p(R = 1|X)$. Equation 10 follows the so-called exponential tilting model, where the multiplicative departure for R is modeled as an exponential term [34].

Likewise, the parameters for hospital visits (scenario 1) are related to the odds of visiting a healthcare facility for healthy and sick populations, formulated similarly as equation 10, for which some information can be extracted from epidemiologic studies and public health reports. Overall, the form of the sensitivity model depends on the estimand and the estimator. Yet, a similar ratio as in equation 10 often appears in the analysis, which must be specified. A detailed analysis of missingness scenarios concerning inquiry 7 is presented in Table A7.

4. Experiments

In this section, we perform two experiments demonstrating the impact of different scenarios and their corresponding assumptions on the results of the missing data analysis. By conducting simulation studies, we can induce different missingness scenarios in a controlled manner and evaluate the reliability of analysis compared to the ground truth. Details regarding the study design and implementation of simulations, medical use case, missingness scenarios, estimands, algorithm derivations, and missing data methods are presented in Appendix D.

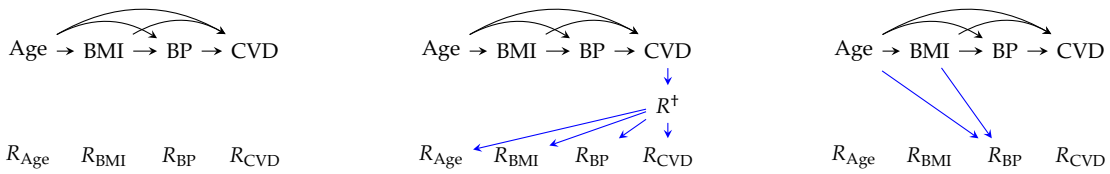
Figure 5 presents three m-graphs with a shared counterfactual causal DAG for cardiovascular disease (CVD) influenced by baseline variables and measurements, namely age, body mass index (BMI) at admission, (systolic) blood pressure (BP), and CVD outcome. The causal DAG was inspired by the work of Bakhtiyari et al. [46].

For the CVD causal model, three unique cases were induced using different combinations of missingness scenarios:

Case 1. (Figure 5a) We induced missingness under scenario 10, where invalid entries were dropped from the dataset.

Case 2. (Figure 5b) In addition to scenario 10, we induced a selection-bias missingness under scenario 1, where the healthy sub-population, in terms of CVD, visited the healthcare facility less frequently.

Case 3. (Figure 5c) In addition to scenario 10, we assumed that age and BMI values were always measured and influenced the measurement of BP under scenario 5, while they may have not been recorded always (scenario 8).



(a) Scenario 10 **(b)** Scenario 1+10 **(c)** Scenario 5+8+10

Figure 5. Example m-graphs for cardiovascular disease under the influence of different constellations of missingness scenarios: Scenario 1 (patient complete non-visit), 5 (missing measurements due to diagnostic irrelevance), 8 (unrecorded observations), and 10 (omission of invalid data entries). Top row nodes and edges are the shared counterfactual sub-graph. Blue edges encode the missingness mechanisms

For each case, we studied two estimands:

Estimand 1. (counterfactual mean BP) We estimated the counterfactual mean blood pressure in the form of estimand 2.

 doi:10.20944/preprints202404.0429.v1

Estimand 2. (classification accuracy for CVD) We evaluated the MSE loss for a trained prediction model for CVD under the full-availability policy. The estimand was $\mathbb{E}_{\pi_{\text{full}}}[(y - f(x))^2]$. The model $f(.)$ was a logistic regression classifier trained on a mean-imputed dataset. The focus of this study was only to estimate the performance of the existing model at deployment. Therefore, perfect model training or fine-tuning was not necessary.

Finally, the analysis was done using three different methods for both estimands:

Method 1. We performed complete-case analysis (CCA). For the mean BP estimand, we took the average BP value for the observed cases. For the model performance estimand, we tested $f(.)$ using complete cases only.

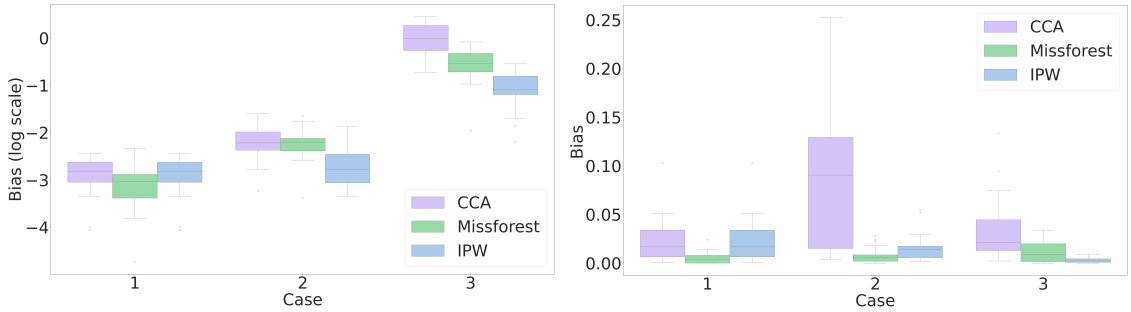
Method 2. We used the Missforest imputation method [29] for the incomplete datasets. The analyses were then carried out using the imputed dataset.

Method 3. We performed IPW estimation using equation 8. After the identification step, propensity scores were estimated via re-weighted complete cases.

We modeled the structure of the m-graphs for each case and simulated its distribution parameters in 20 iterations by sampling uniform-randomly from a search space. In this way, various simulation cases were generated, following the imposed structure [47]. For each iteration, we reported the estimation bias regarding absolute error using the ground truths, namely the counterfactual mean BP (on a log scale) and factual classification accuracy.

Figure 6a presents the estimands 1 and 2 results. For method 1 (CCA), only case one aligns with its assumption, giving the most negligible bias among the three cases. For method 2 (Missforest), cases 1 and 3 align with their assumption, yet interestingly, case 3 has the highest bias. The IPW estimator is designed domain-informed. Higher biases in case 3 for method 3 (IPW) can be attributed to possible model misspecifications for the propensity score estimation. Overall, the domain-informed IPW estimation gives the lowest bias.

Figure 6b presents the results for the classification accuracy estimation at deployment. All three methods give approximately 0.1 bias (average classification accuracy over all cases was 0.77), except for the CCA estimation in case 2. The considerable difference in the estimation bias between the two estimand shows how the reliability of the methods may depend on the estimand, in addition to the dataset characteristics. In both experiments, CCA and IPW are identical in case 1, as the IP-weights are 1 for all samples (completely random missingness).



(a) Counterfactual mean BP **(b)** CVD classification accuracy

Figure 6. Results of the simulation studies: Estimation bias for (a) the counterfactual mean BP, (b) classification accuracy of the logistic regression model for CVD.

5. Discussion

In order to overcome the challenges of a missing data problem, it is recommended to report the problem transparently [3,4], collect complementary datasets from various sources [7,14], and choose a suitable method accordingly. However, the question remains: what kind of data should we collect and incorporate into the analysis, and how should the missing data be reported?

 doi:10.20944/preprints202404.0429.v1

This paper propounds a framework for analyzing missingness in observational health data. Via recognition and categorization of fundamental scenarios, we provided a basis for understanding the data collection processes in healthcare facilities. According to this framework, focusing on human agents (physicians, patients) and environments (hospitals, clinics) helps discover and report the scenarios. As shown in this paper, the implications of scenarios must be formulated for identification, estimation, and sensitivity analysis steps. Once the corresponding assumptions are specified, a suitable method can be selected. During the analysis, complementary datasets can be utilized when they enable and facilitate the analysis steps, e.g., when new variables make the missingness mechanism identifiable (inquiry 1) or when a dataset offers meaningful interpretations of the sensitivity parameters (inquiry 7).

This paper is intended for not only the medical data scientists but also the developers of missing data methods in the ML community as a response to the recurring theme of devising sophisticated (deep learning) imputation models [48–50]. While imputation (or IPW) methods with high learning capabilities may considerably enhance the estimation step, most assumptions and inquiries concern the identification and sensitivity analysis steps (inquires 1-5, 7). Therefore, to ensure the method's effectiveness, it is necessary to actively scrutinize the assumptions, especially if the application domain is safety-critical.

In fact, methodology papers often report the conditions using Rubin's at-randomness categorization [23], according to which a missingness mechanism can be completely-at-random (MCAR), at-random (MAR), and not-at-random (MNAR). However, this categorization has been a blessing and a curse; even though it effectively formalizes the first step of identification[2] (addressed by inquiry 1), it was introduced in the original work specifically for estimand 2 and was later extended for the joint-distribution estimand $p(X, Y)$ [25]. The works we criticize often make at-randomness assumptions disconnected from the reality of the use case and denuded of their context, ignoring the requirements of the estimand. If these works were more sensitive to the nuances of the problem and adhered to the standard missing data analysis process, they would positively influence the medical data analysis works that adopt them. We suggest that future works discuss the assumptions in a form similar to the inquiries in this paper and present some real-world examples for validity and violation cases.

The current paper studied ten fundamental and prevalent scenarios considering healthcare facilities. Nevertheless, new and different scenarios may arise under new data observation, recording, and collection circumstances. Likewise, the inquiries in this paper were made regarding the general frame of the missing data theory, which is relevant for most analysis methods. It is conceivable that specific methods have unique assumptions and, therefore, further inquiries to explore. Nevertheless, the analysis process in this paper can be applied to new scenarios and inquiries.

The scope of our paper is limited to observational data from healthcare facilities. Other similar health domains, such as medical wearable sensors, can be similarly analyzed; e.g., one can study the scenarios induced by patients (e.g., missing measurement due to taking off the smartwatch because of irritation) or technical issues (operating system issues or low battery), and connect it to the same inquiries introduced in this paper.

In addition, exploring all details for the introduced scenarios scenarios would have indeed exceeded the publication's scope. Future research endeavors would benefit from deeper investigations into each individual missingness scenario with greater granularity.

Finally, it is worthy of note that the missing data problem lies in the broader category of *data coarsening processes*, where the data veracity, representativeness, or completeness is impaired. Other data coarsening processes in healthcare include the patient recall bias [51], variations in medical outcome definition over time or depending on the defining consortium [52], or coarsened representation of pertinent negative/positive values [14,53]. Extending our analysis framework to other data

[2] The work also bears extreme historical significance for the missing data analysis methodology

 doi:10.20944/preprints202404.0429.v1

coarsening problems in future work could bridge further translational gaps in medical and healthcare data analysis.

Author Contributions: The authors' contribution to this paper is as follows: Conceptualization, A.Z., H.V., and N.A.; methodology, A.Z., and H.V.; software, G.L., and A.Z.; formal analysis, A.Z., and H.V.; investigation, A.Z., O.A.C., and M.P.; writing—original draft preparation, A.Z., O.A.C.; writing—review and editing, O.A.C., H.V, M.P. and N.A.; visualization, A.Z., O.A.C., G.L.; supervision, N.A.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Bavarian Ministry for Economic Affairs, Regional Development, and Energy as part of a project to support the thematic development of the Fraunhofer Institute for Cognitive Systems.

Data Availability Statement: The data used in the Experiment section were synthesized using the PyPARCS Python library for causal simulation. The simulation configuration is presented in Appendix D.

Acknowledgments: We thank Ruijie Chen for his support in reviewing the related clinical papers. We thank Leopold Mareis, Elisabeth Pachl, and Patrick Rockenschaub for providing instructive feedback.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. References to Missingness Scenarios

Appendix A.1. Scenario 1

Cheng et al. [21] assumed higher HbA1c values in the data than the overall population due to the non-visit of healthier patients. Wells et al. [14] indicated that data is missing, partly because patients expired before seeking treatment.

Appendix A.2. Scenario 2

Wells et al. [14] assumed a bias in the systolic blood pressure value with respect to the hospital visits, as healthier patients are less likely to utilize the healthcare system. Schafer and Graham [1] suggested higher blood pressure measurements for hypertension patients who returned for a follow-up visit. Ayilara et al. [9], Phung et al. [10] and Mirkes et al. [15] highlighted the fragmentation of datasets due to the exclusion of deceased patients and those who missed clinic visits. Lip et al. [54] reported a connection between unknown outcome value (Thromboembolism) due to missing follow-ups and the death of the patients. In the study on Apgar score for evaluating newborn infants, Apgar [55] reported that 1.5%

Appendix A.3. Scenario 3

Lip et al. [54] reported that some patients were not started on VKA treatment (an inclusion criterion), partly because of comorbidities and intolerance of anticoagulation.

Appendix A.4. Scenario 4

Lip et al. [54] reported that some patients were not started on VKA treatment (an inclusion criterion), partly because of poor compliance. Ayilara et al. [9] introduces a data category, the Patient-Report Outcomes (PROs), whose missingness entirely depends on the patient's willingness to report. Penny and Atkinson [3] point out the case where offering monetary incentives substantially increases response rates, though it is unknown whether this would predispose a particular type of person to provide data.

Appendix A.5. Scenario 5

Zachariasse et al. [56] reported missing values in the documentation of vital signs, which were measured at the nurse's discretion during triage. As a result, some patients had incomplete sets of

 doi:10.20944/preprints202404.0429.v1

vital signs recorded. Missing vital signs were assumed to be within normal range and were more often encountered in less severe patients.

Appendix A.6. Scenario 6

Zachariasse et al. [56] reported that some hospitals' information (high care admission, emergency surgery) was unavailable.

Appendix A.7. Scenario 7

Zachariasse et al. [56] reported high missingness in the triage urgency of one of the emergency departments, resulting from "the absence of triage nurses during night shifts at the start of the study." Limb [57] describes the patient discharge situation in NHS hospitals after the emergence of the COVID-19 omicron variant in order to "release the maximum number of beds."

Appendix A.8. Scenario 8

Wells et al. [14] stated that an important variable related to diabetes research is the length of time that patients fast before having blood drawn for the metabolic panel; however, this variable is only recorded on the paper and hence unavailable in the dataset. Gray et al. [58] reported that, even though more variables were recorded in their dataset, they only analyzed ones that did not require a laboratory or third party to be measured. Falcoz et al. [59] considered a dataset that was populated using a fixed pull-down menu and mentioned that additional unrecorded features could influence the response variable.

Appendix A.9. Scenario 9

Lip et al. [54] included patients in the analysis only if they were without mitral stenosis or previous heart valve surgery and did not use vitamin K antagonists (VKA) or heparin upon discharge. Aguirre et al. [60] and Wishart et al. [61] excluded patients who did not have surgery, with incomplete local therapy, and patients with less than four nodes removed with a diagnosis of node-negative disease. Aguirre et al. [60] also omitted patients who had metastasis or ductal carcinoma in situ, underwent conservative surgery without having radiotherapy afterward, did not undergo sentinel lymph node biopsy or axillary lymph node dissection, or whose cause of death was unknown. Gray et al. [58] only considered patients older than 40. In contrast, Falcoz et al. [59] restricted their study to patients older than 16. In their study of warning scores among hospitalized patients assessed by a rapid response team, Fernando et al. [62] excluded cardiac arrest cases, for which a different response team is responsible.

Appendix A.10. Scenario 10

Aguirre et al. [60] removed observations with any missing data in HER2 status or Ki67. Falcoz et al. [59] included solely patients with more than 95% complete information and discarded patients with any missing values in the variables considered in the study. Additionally, they excluded features with too many inconsistent or missing values. Blatchford et al. [63] excluded patients with incomplete records and patients whose outcomes could not be determined. Fernando et al. [62] reported excluding patients with incomplete demographics, missing outcomes, and those for whom the warning scores could not be computed.

 doi:10.20944/preprints202404.0429.v1

Appendix B. Other Estimands

Appendix B.1. Full-Availability Estimand for Decision Trees Learning

One way to train a decision tree is via computation of the Gini impurity index and selection of a split that leads to the largest reduction in the index. The Gini index for a (binary classification) dataset is given by

$$g(\mathcal{D}) = 2\eta(1-\eta), \tag{A1}$$

where $\eta = p(Y=1)$ is the positive class probability. Upon a split concerning a (binary) covariate X_i, as $\mathcal{D} = \mathcal{D}_1 + \mathcal{D}_2$ with N_1 and N_2 sample sizes respectively, the new index is calculated as

$$g_{X_i}(\mathcal{D}) = \frac{N_1}{N} g(\mathcal{D}_1) + \frac{N_2}{N} g(\mathcal{D}_2), \tag{A2}$$

The best split is found via $\arg\max_{X_i} g(\mathcal{D}) - g_{X_i}(\mathcal{D})$.

Consider the objective of classification given the full availability at deployment. The estimand for this objective is written, according to equation A2, as $\mathbb{E}_{\pi_{\text{full}}}(g_{X_i})$ at each split. This can be broken down into two estimations: (1) the class probability η for $\mathcal{D}_i, i \in \{1,2\}$, (2) the split ratios N_1/N and N_2/N, which requires estimation of one of N_1 or N_2.

Assuming $\mathcal{D}$ at a split is already conditioned by splits concerning the covariate set X_j; the first estimation is then written as

$$\mathbb{E}_{\pi_{\text{full}}}(Y|X_i, X_j). \tag{A3}$$

For the second objective, suppose the realizations for X_i are $X_i = (1,1,0,\texttt{NaN})^\top$. One way to estimate N_1 (or N_2) is by directly estimating the missing entry for X_i. Hence, the second estimand is written as

$$\mathbb{E}_{\pi_{\text{full}}}(X_i|X_j). \tag{A4}$$

Appendix B.2. Feature Importance

A method for calculating the importance of a covariate X_i in a supervised ML problem is to compare the estimation accuracy with or without the covariate. Feature importance is often posed under the full-availability policy. Therefore, the estimand for feature importance can be given by

$$\mathbb{E}_{\pi_{\text{full}}}[(y - f(x;w))^2 - (y - f(x_{\setminus i};w))^2] \tag{A5}$$

 doi:10.20944/preprints202404.0429.v1

 doi:10.20944/preprints202404.0429.v1

Appendix C. Tables of Inquiry Results

Table A1. Analysis of scenarios with respect to inquiry 1 - missingness mechanism.

Scenario No.	Interpretation
1	Since all study variables are missing in case of complete non-visit, a shared missingness indicator R_0 influences all the other R_i's, such that $\forall i, R_i = 0$ when $R_0 = 0$. Based on the potential reasons for complete non-visit, R_0 may have incoming edges from the main health outcome such as in fatal diseases (e.g., reflected in death status, possibly available in external data sources such as the Social Security Death Index database [14]), risk factors such as BMI, or non-medical factors such as socioeconomic status (e.g., reflected in the occupation and level of education variables in patient information [39]). Figure 3 presents a schematic m-graph structure for this scenario.
2	Similar to scenario 1, missingness of follow-up visits can be modeled using a shared indicator $R_j^\dagger$ for the j-th visit. The simplicity of this scenario compared to scenario 2 lies in the likelihood that $R_j^\dagger$ is influenced by observations at the $j-1$-th visit. For instance, missing follow-up can be influenced by the health status upon discharge or length of stay for inpatient admissions.
3	Observed and partially observed causes of missingness include health markers directly related to the reasons leading to the interruption of data measurement. For instance, code-blue missingness may happen right after anomalies in vital sign readings. It is unlikely for missingness under scenario 3 to have exogenous causes since measurement standards are in place for patients during their entire inpatient stay. Causes that might confound missingness with other study variables include event location and timestamp metadata, as they highlight the decision-making conditions.
4	Reasons for patients' refusal are usually predicated on personal characteristics that are not associated with other health-related variables. Thus, it is a fairly safe assumption to consider them exogenous unless disproved explicitly. A possible health-related cause for missingness under this scenario is refusal due to pain intolerance or distress, which may indicate a negative health status at the moment. In this case, indicators of the health status can be considered the causes of missingness.
5	The missingness indicators under this scenario are influenced by observed health variables X, mediated by the attending physician unless X itself is subjected to missingness under scenario 8. In this case, the edge is received from the counterfactual counterpart X^* (since it influences missingness regardless of its observation status). As an extreme case, a counterfactual variable that is completely missing under scenario 8 is a latent cause for missingness (see the corresponding entry for scenario 8). Causes that might confound missingness with other study variables include the attending physician's identifiers, which are proxies of medical practice styles.
6, 7	By definition, missingness under these scenarios can be predicted using the patient's location, transfer information, or type of resource required for making the observation. This information is considered crucial, being part of the management and billing data; therefore, causing variables for this scenario are likely to be fully observed and available, especially for datasets from large healthcare facilities.
8	recording of the counterfactual study variables X^* depends on the nature of the variable, as well as the style of the medical practice of the attending physician and the recording capabilities of the software tool. For instance, expensive and decisive tests such as medical imaging or lab tests are generally recorded, while qualitative examination results may escape from recording depending on the physician or if the software tool does not provide an entry for it. Overall, there is a possibility that the reasons for missingness under this scenario confound other study variables if they are also affected by the medical practice style, e.g., when a physician with a tendency to record the most variables also diagnoses and prescribes treatments more effectively.

(continued on the next page)

 doi:10.20944/preprints202404.0429.v1

Table A1. Analysis of scenarios with respect to inquiry 1 - missingness mechanism (continue).

Scenario No.	Interpretation
9	Inclusion/exclusion criteria directly indicate the reason for missingness under this scenario. For instance, age is the direct cause of missingness if, by design, data is selected according to the age criteria.
10	The potential reasons for invalid entries mentioned in this paper are likely unrelated to the analysis of interest, as they are mostly related to human and software tool errors. However, one should be cautious about treating all medically unrelated variables as exogenous causes. Instead, whether these variables can realistically be confounders for other variables should be investigated. For instance, socioeconomic or occupational status may affect overall health and healthcare facility visits.

Table A2. Analysis of scenarios with respect to inquiry 2 - missingness distribution shift.

Scenario No.	Interpretation
1, 2	Parametric shift mat occur if the population distributions change. Examples include: (i) conducting analysis using hospital data but deploying it for patients in local clinics or the general healthy population, such as in a preventive healthcare plan, (ii) conducting analysis using data from a specific cohort but deploying it for another cohort, (iii) when the target population visits healthcare facilities more or less frequently than during the data acquisition stage.
3	Shift occurs if the transfer protocols or observation protocols during hospitalization change. Examples include (i) observing more, fewer, or different variables in different hospital wards and (ii) encountering data availability or unavailability after the transfer despite being available during the data acquisition stage.
4	A no-shift assumption regarding patient refusal behavior for taking a test or answering questions appears reasonable. However, it is essential to consider the possibility of a shift due to data-sharing consent. In such cases, the data available for analysis could differ from the data available to physicians at deployment.
5, 6, 7, 8	Shift occurs when there are changes in the observation policy of physicians, healthcare facility protocols, available equipment, or data collection software. Examples include (i) alterations in the measurement decisions resulting from the deployment of a prediction model, (ii) modifications in the utilized diagnostic flowcharts and scores, (iii) fluctuations in the level of physicians' expertise, and (iv) enhancing data collection protocols following significant events such as an epidemic.
9	Inclusion/exclusion criteria typically imply a shift in missingness unless the same criteria are applied for the admission of patients, which is highly unlikely in most cases. An example of *no*-shift occurs when the data scientist restricts the general population to the cohort of interest for deployment, using inclusion/exclusion criteria.
10	Shift occurs only when the reasons behind errors and invalid entries in the data change.

 doi:10.20944/preprints202404.0429.v1

Table A3. Analysis of scenarios with respect to inquiry 3 - no-direct-effect assumption.

Scenario No.	Interpretation
1 & 2	non-visit possibly influences the health variables via a direct causal effect on the treatments: patients usually do not receive treatment until being admitted (except the self-medication case). This means that the NDE assumption is mostly violated. Take the example of missing follow-up after the first visit, compared with unrecorded observations in a realized follow-up visit. In the latter, missingness does not influence the health status at the end of the follow-up visit. In contrast, in the former, the health status, for example, may degrade due to discontinuation of diagnosis/treatment.
3, 4, 5, 6, 7	Assuming most of the variables measured under these scenarios are related to the patient's health status, the validity of the NDE assumption under these scenarios depends on the nature of the measurement. If the measurement directly influences the patient's health (e.g., invasive tests) or indirectly (temporary pause of a medication), the NDE assumption is violated. Following the discussion on scenarios 1 and 2, the NDE assumption is likely violated for the treatment and medication variables since treatment decisions usually depend on the observations.
8	Unless disproved explicitly, missingness under this scenario admits the NDE assumption since recording status of the variables cannot influence the variables by any conceivable means.
9, 10	Since missingness under these scenarios are related to the data analysis and occurs after the data collection step, the NDE assumption can be made.

Table A4. Analysis of scenarios with respect to inquiry 4 - no-interference assumption.

Scenario No.	Interpretation
1, 2, 3, 4, 5, 8	Observations and measurements under these scenarios permit the no-interference assumption, as the decisions are being generally made per individuals.
6	The healthcare facility protocols typically apply uniformly to individuals and remain consistent over a short period. Hence, it is reasonable to make the no-interference assumption in this scenario.
7	This scenario is the most critical and obvious example of violating the no-interference assumption. In this scenario, a prioritization scheme is usually adopted to allocate observation and measurement resources. Examples are (i) early discharge, no admission due to limited hospital capacity during the epidemic, and (ii) delayed or canceled measurements for healthier patients during staff overload.
9, 10	Unless for particular reasons, the data scientists do not induce interference by the inclusion/exclusion criteria, and the no-interference assumption holds. An example of a violation of the no-interference assumption (though not to be conceived as a meaningful scenario) is when performing sample selection based on the so-far selected samples from different cohorts, e.g., when we only choose up to 20 patients from an age stratum.

 doi:10.20944/preprints202404.0429.v1

Table A5. Analysis of scenarios with respect to inquiry 5 - selection bias.

Scenario No.	Interpretation
1, 9	By their definitions, these scenarios induce selection bias, discarding the entire sample (of a specific sub-population) from the dataset.
2	To miss a follow-up under this scenario implies that the patient still has recorded data in the database. However, if analysis is limited to a specific follow-up (e.g., analysis of health status in the second hospital visit), then patients with limited data are subjected to selection bias.
3, 5, 6, 7, 8, 10	These scenarios by default concern data entries and do not cause missingness of an entire data sample; hence, no selection bias occurs.
4	A situation where patients' refusal can lead to selection bias missingness is when they refuse to give data-sharing consent.

Table A6. Analysis of scenarios with respect to inquiry 6 - monotonicity.

Scenario No.	Interpretation
1, 9	Complete non-visit and sample exclusion induce only two complete case and all-missing patterns.
2	Missing follow-ups in clinical studies induce monotone missingness, since by the study design rules, the patients who are absent, for any reason, from a visit are excluded from the remaining visits (case drop-out). However, such a rule does not apply in healthcare facilities; therefore, this scenario, in general, leads to non-monotone missingness.
3, 4, 6, 7, 8, 10	The scenarios are not determined to induce a monotone missingness pattern, unless for a specific reason related to the problem at hand.
5	Observation according to the diagnostic flowcharts and score tables induce a monotone missingness pattern, where extensive secondary measurements are not made unless primary ones are. However, many diagnostic flowcharts are utilized across all patients in a healthcare facility dataset. The set of primary tests usually overlaps among different flowcharts; therefore, a monotone pattern may still emerge. The pattern graph framework [27] provides a powerful methodology for dealing with missing data in this situation.

 doi:10.20944/preprints202404.0429.v1

Table A7. Analysis of scenarios with respect to inquiry 7 - sensitivity parameters.

Scenario No.	Interpretation
1	Potential odds ratio information for sensitivity analysis include (i) difference in total hospital visits between the healthy and sick sub-populations, (ii) difference in medical care advantages received among different socioeconomic strata, (iii) difference in the death rates reported in the healthcare facility, and in total, specific to a disease.
2	Potential odds ratio information for sensitivity analysis include differences in the number of visits for healthy vs. sick patients, between groups with different socioeconomic status due to insurance plans, or morbidity rate for a specific diagnosis, possibly obtained from epidemiological research.
3	Potential odds ratio information for sensitivity analysis include the difference in the interrupted measurement level due to specific events, such as the code-blue.
4	For those types of missingness due to patients' health-related refusal (such as intolerance to pain), meaningful differences may be found for available and unavailable samples, e.g., conceivable level of infection which may cause intolerable pain.
5	Potential odds ratio information for sensitivity analysis directly include the level of measured variables, e.g., in different branches of the diagnostic flowcharts. For instance, one may ask how later specific measurements may change if the results of the primary tests flip (reflecting the unavailable sub-population).
6	Potential odds ratio information for sensitivity analysis can be found by analyzing healthcare facility protocols for specific measurements. Since these rules are justified based on extensive research, informed sensitivity analysis might be possible via public health works that analyze such protocols.
7	Potential odds ratio information for sensitivity analysis include differences in the level of measurement for situations when the availability of resources changes, e.g., comparing the waiting line for a medical test or number of admissions.
8, 10	Due to complete-randomness, no specific sensitivity parameter can be conceived in general for this scenario.
9	The sensitivity parameters have interpretations similar to the complete non-visit scenario, except that the data scientists have induced omission under this scenario. Therefore, the sensitivity parameters and ranges may be obtained from the original dataset.

Appendix D. Simulation

Appendix D.1. Cardiovascular Diseases (CVD) m-graph

Bakhtiyari et al. [46] introduces a discovered causal DAG for CVD as depicted in Figure A1. Even though the simulation study design can be done via any causal DAG structure, we decided to add a medical context so that inducing different missingness scenarios becomes more meaningful, which is in line with the concern of this paper. Nevertheless, for simplicity, we chose one variable per each class of variables in Figure A1: (1) BMI from class A - "Overweight, General Obesity, Visceral adiposity," (2) blood pressure (BP) from class M - "Cardiometabolic risk factors," and (3) age from class C - "Age, sex, smoking, educational status, physical activity, family history of CVD, medical treatment." Also, we chose not to induce any class U node - latent confounders to keep the identification straightforward.

 doi:10.20944/preprints202404.0429.v1

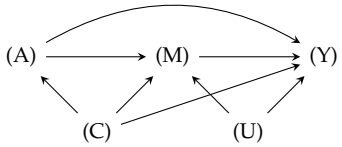


Figure A1. Causal DAG for cardiovascular disease in [46]. (A) Overweight, General Obesity, Visceral adiposity, (M) Cardiometabolic risk factors, (C) Age, sex, smoking, educational status, physical activity, family history of CVD, medical treatment, (U) unmeasured confounders, (Y) Cardiovascular disease.

Appendix D.2. m-graph Simulation

The m-graph structures for three cases were selected according to the analysis for inquiry 1 in Table A1. We used the PARCS Python package for causal simulation [64]. Using PARCS, we can partially specify a DAG as the data generation process for causal simulation and determine a search space to (uniform) randomly select the DAG parameters for different simulation iterations [47]. PARCS performs simulation with a DAG *description* object and a *guideline* for partial randomization. The configurations used for simulation in this paper are presented in Code 1, 2, 3, and 4.

Code 1: case 1 description

```
# counterfactual subgraph
## nodes
age: normal(?), correction[]
bmi: normal(?), correction[]
bp: normal(?), correction[]
cvd: bernoulli(?), correction[]

## edges
age->bmi: random
age->bp: random
bmi->bp: random
age->cvd: random
bmi->cvd: random
bp->cvd: random

# missing subgraph
R_age: bernoulli(?), correction[target_mean=0.7]
R_bmi: bernoulli(?), correction[target_mean=0.7]
R_bp: bernoulli(?), correction[target_mean=0.7]
R_cvd: bernoulli(?), correction[target_mean=0.7]
```

Code 2: case 2 description

```
# counterfactual subgraph
## nodes
age: normal(?), correction[]
bmi: normal(?), correction[]
bp: normal(?), correction[]
cvd: bernoulli(?), correction[target_mean=0.6]

# missing subgraph
R_all: bernoulli(?), correction[target_mean=0.7]
R_bmi: bernoulli(?), correction[target_mean=0.7]
R_age: bernoulli(?), correction[target_mean=0.7]
R_bp: bernoulli(?), correction[target_mean=0.7]

## edges
age->bmi: random
age->bp: random
bmi->bp: random
age->cvd: random
bmi->cvd: random
bp->cvd: random

# mechanism
cvd->R_all: random
```

Code 3: case 3 description

```
# counterfactual subgraph
## nodes
age: uniform(mu_=0.5, diff_=0.3)
bmi: uniform(mu_=0.5, diff_=0.3)
bp: normal(?), correction[]
cvd: bernoulli(?), correction[]

## edges
age->bp: random
bmi->bp: random
age->cvd: random
bmi->cvd: random
bp->cvd: random

# missing subgraph
R_bmi: bernoulli(?), correction[target_mean=0.7]
R_age: bernoulli(?), correction[target_mean=0.7]
R_bp: bernoulli(p_=0.5age+0.5bmi), correction[]

# mechanism
bmi->R_bp: random
age->R_bp: random
```

Code 4: randomization guideline

```
nodes:
  bernoulli:
    p_: [[f-range, -1, 1], [f-range, -7, -3, 3, 7], [f-range, -7, -3, 3, 7]]
  normal:
    mu_: [[f-range, -7, -1, 1, 7], [f-range, -7, -1, 1, 7], [f-range, -7, -1, 1, 7]]
    sigma_: [[f-range, -7, -1, 1, 7], [f-range, -7, -1, 1, 7], [f-range, -7, -1, 1, 7]]
edges:
  identity: none
```

Appendix D.3. Objectives

We estimated the counterfactual mean blood pressure under different scenario constellations for the first experiment. In the second experiment, while correct ML training in the presence of missing data is a crucial topic, we were interested in model performance evaluation, regardless of how suitable the model is. Performance evaluation in the presence of missing data is an under-explored yet critical topic, even more crucial than model training. The reason is that bias in model training (with correct evaluation) leads to poor training in the worst-case scenario. A poor performance evaluation leads to false promises about a model's suitability for deployment, which may turn out catastrophically.

 doi:10.20944/preprints202404.0429.v1

Table A8. Code specifications for the models and methods.

Specification	Description
	Classifier
Model	Logistic Regression
Software	Sklearn v1.4.1, using `linear_model.LogisticRegression`
Parameters	Sklearn default parameters
	Missforest imputer
Software	Sklearn v1.4.1, using `impute.IterativeImputer` and `ensemble.RandomForestRegressor`
Parameters	Sklearn default parameters for both objects
	Propensity score model
Model	Logistic Regression
Software	Sklearn v1.4.1, using `linear_model.LogisticRegression`
Parameters	Sklearn default parameters

We trained a logistic regression model according to the specifications of each case and evaluated it three times using the CCA, Missforest, and IPW methods. Code specifications for the classifier and missing data methods are presented in Table A8.

Appendix D.4. Identification

For case 1, we have $R \perp\!\!\!\perp X$, therefore we have $\text{PS}(x) = 1, \forall x \in X$. For case 2, we write the propensity score as

$$p_{\pi_{\text{init}}}(R_{\text{bp}} = 1|\text{Age}, \text{BMI}, \text{BP}, \text{CVD}). \tag{A6}$$

The m-graph in Figure 5b gives $R_{\text{bp}} \perp\!\!\!\perp \text{Age}, \text{BMI}, \text{BP}|\text{CVD}$. Therefore, the weight in (A6) is simplified as

$$p_{\pi_{\text{init}}}(R_{\text{bp}} = 1|\text{CVD}). \tag{A7}$$

More specifically, we have two scores for CVD $= 0$ and 1:

$$\begin{aligned} \text{PS}_1 &= p(R = 1|\text{CVD} = 1), \\ \text{PS}_0 &= p(R = 1|\text{CVD} = 0), \end{aligned}$$

or normalized for PS_1 as

$$\begin{aligned} \text{PS}_1 &= 1, \\ \text{PS}_0 &= \frac{O(CVD|R = 1)}{O(CVD)}, \end{aligned} \tag{A8}$$

The numerator of PS_0 in equation A8 can be estimated from the available data. However, the denominator queries the odds of CVD in the entire population, including the censored patients. This means that the estimand is unidentifiable due to the self-masking edge $\text{CVD} \to R_{\text{CVD}}$.

Regardless of the apparent dead-end, we can proceed with the analysis if population-wide summary statistics about CVD are available. Such information can be possibly found in public health research literature. For the sake of simulation, we query the odds from the ground truth dataset and employ it in the IPW estimation, knowing that, in reality, such information must be sought outside the available dataset.

 doi:10.20944/preprints202404.0429.v1

For case 3, we have $R \perp\!\!\!\perp \text{BP}, \text{CVD} | \text{Age}, \text{BMI}$. Hence we write the propensity score estimand as $p_{\pi_{\text{init}}}(R|\text{Age}, \text{BMI})$. This model, however, is expressed in terms of counterfactual variables. According to the m-graph in Figure 5c, we factorize R as follows:

$$p(R|\text{Age}, \text{BMI}) = p(R_{\text{Age}})p(R_{\text{Age}})p(R_{\text{CVD}})p(R_{\text{BP}}|\text{Age}, \text{BMI}). \tag{A9}$$

Equation A9 shows that the propensity score for both estimands can be estimated as

$$p_{\pi_{\text{init}}}(R_{\text{BP}}|\text{Age}^*, \text{BMI}^*, R_{\text{BMI}} = 1, \text{Age} = 1). \tag{A10}$$

References

1. Schafer, J.L.; Graham, J.W. Missing data: our view of the state of the art. Psychological methods **2002**, 7, 147.
2. Mohan, K.; Pearl, J.; Tian, J. Graphical Models for Inference with Missing Data. In Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States; Burges, C.J.C.; Bottou, L.; Ghahramani, Z.; Weinberger, K.Q., Eds., 2013, pp. 1277–1285.
3. Penny, K.I.; Atkinson, I. Approaches for dealing with missing data in health care studies. Journal of clinical nursing **2012**, 21, 2722–2729.
4. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Annals of internal medicine **2015**, 162, 55–63.
5. Le, T.D.; Beuran, R.; Tan, Y. Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare. In Proceedings of the 10th International Conference on Knowledge and Systems Engineering, KSE 2018, Ho Chi Minh City, Vietnam, November 1-3, 2018; Thuy, N.T.; Tojo, S.; Hanh, T.; Nguyen, M.L.; Phuong, T.M.; Bao, V.N.Q., Eds. IEEE, 2018, pp. 247–251. https://doi.org/10.1109/KSE.2018.8573344.
6. Lee, K.J.; Tilling, K.M.; Cornish, R.P.; Little, R.J.; Bell, M.L.; Goetghebeur, E.; Hogan, J.W.; Carpenter, J.R.; et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. Journal of clinical epidemiology **2021**, 134, 79–88.
7. Haneuse, S.; Arterburn, D.; Daniels, M.J. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. JAMA Network Open **2021**, 4, e210184–e210184.
8. Tsvetanova, A.; Sperrin, M.; Peek, N.; Buchan, I.; Hyland, S.; Martin, G. Inconsistencies in handling missing data across stages of prediction modelling: a review of methods used. In Proceedings of the 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI). IEEE, 2021, pp. 443–444.
9. Ayilara, O.F.; Zhang, L.; Sajobi, T.T.; Sawatzky, R.; Bohm, E.; Lix, L.M. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health and quality of life outcomes **2019**, 17, 1–9.
10. Phung, S.; Kumar, A.; Kim, J. A deep learning technique for imputing missing healthcare data. In Proceedings of the 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2019, pp. 6513–6516.
11. Verma, H.; Kumar, S. An accurate missing data prediction method using LSTM based deep learning for health care. In Proceedings of the Proceedings of the 20th international conference on distributed computing and networking, 2019, pp. 371–376.
12. Ismail, A.R.; Abidin, N.Z.; Maen, M.K. Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. Journal of Robotics and Control (JRC) **2022**, 3, 143–152.
13. Lee, K.J.; Carlin, J.B.; Simpson, J.A.; Moreno-Betancur, M. Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. International Journal of Epidemiology **2023**, p. dyad008.
14. Wells, B.J.; Chagin, K.M.; Nowacki, A.S.; Kattan, M.W. Strategies for handling missing data in electronic health record derived data. Egems **2013**, 1.
15. Mirkes, E.M.; Coats, T.J.; Levesley, J.; Gorban, A.N. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. Computers in biology and medicine **2016**, 75, 203–216.

16. Millard, L.A.; Fernández-Sanlés, A.; Carter, A.R.; Hughes, R.A.; Tilling, K.; Morris, T.P.; Major-Smith, D.; Griffith, G.J.; Clayton, G.L.; Kawabata, E.; et al. Exploring the impact of selection bias in observational studies of COVID-19: a simulation study. International Journal of Epidemiology **2023**, 52, 44–57.
17. Moreno-Betancur, M.; Lee, K.J.; Leacy, F.P.; White, I.R.; Simpson, J.A.; Carlin, J.B. Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. American journal of epidemiology **2018**, 187, 2705–2715.
18. Marino, M.; Lucas, J.; Latour, E.; Heintzman, J.D. Missing data in primary care research: importance, implications and approaches. Family Practice **2021**, 38, 199–202.
19. Sperrin, M.; Martin, G.P.; Sisk, R.; Peek, N. Missing data should be handled differently for prediction than for description or causal explanation. Journal of clinical epidemiology **2020**, 125, 183–187.
20. Le Morvan, M.; Josse, J.; Scornet, E.; Varoquaux, G. What is a good imputation to predict with missing values? Advances in Neural Information Processing Systems **2021**, 34, 11530–11540.
21. Cheng, G.; Chen, Y.C.; Smith, M.A.; Zhao, Y.Q. Handling Nonmonotone Missing Data with Available Complete-Case Missing Value Assumption. arXiv preprint arXiv:2207.02289 **2022**.
22. von Kleist, H.; Zamanian, A.; Shpitser, I.; Ahmidi, N. Evaluation of Active Feature Acquisition Methods for Time-varying Feature Settings. arXiv preprint arXiv:2312.01530 **2023**.
23. Rubin, D.B. Inference and missing data. Biometrika **1976**, 63, 581–592.
24. Zhou, Y.; Little, R.J.; Kalbfleisch, J.D. Block-conditional missing at random models for missing data. Statistical Science **2010**, 25, 517–532.
25. Nabi, R.; Bhattacharya, R.; Shpitser, I. Full law identification in graphical models of missing data: Completeness results. In Proceedings of the International Conference on Machine Learning. PMLR, 2020, pp. 7153–7163.
26. Malinsky, D.; Shpitser, I.; Tchetgen Tchetgen, E.J. Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. Journal of the American Statistical Association **2022**, 117, 1415–1423.
27. Chen, Y.C. Pattern graphs: a graphical approach to nonmonotone missing data. The Annals of Statistics **2022**, 50, 129–146.
28. Li, Y.; Miao, W.; Shpitser, I.; Tchetgen Tchetgen, E.J. A self-censoring model for multivariate nonignorable nonmonotone missing data. Biometrics **2023**, 79, 3203–3214.
29. Van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research **2007**, 16, 219–242.
30. Stekhoven, D.J.; Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics **2012**, 28, 112–118.
31. Sun, B.; Perkins, N.J.; Cole, S.R.; Harel, O.; Mitchell, E.M.; Schisterman, E.F.; Tchetgen Tchetgen, E.J. Inverse-probability-weighted estimation for monotone and nonmonotone missing data. American journal of epidemiology **2018**, 187, 585–591.
32. Tchetgen, E.J.T.; Wang, L.; Sun, B. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. Statistica Sinica **2018**, 28, 2069.
33. Carpenter, J.R.; Kenward, M.G.; White, I.R. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Statistical methods in medical research **2007**, 16, 259–275.
34. Kim, J.K.; Yu, C.L. A semiparametric estimation of mean functionals with nonignorable missing data. Journal of the American Statistical Association **2011**, 106, 157–165.
35. Franks, A.M.; Airoldi, E.M.; Rubin, D.B. Nonstandard conditionally specified models for nonignorable missing data. Proceedings of the National Academy of Sciences **2020**, 117, 19045–19053.
36. Zamanian, A.; Ahmidi, N.; Drton, M. Assessable and interpretable sensitivity analysis in the pattern graph framework for nonignorable missingness mechanisms. Statistics in Medicine **2023**, 42, 5419–5450.
37. Johnson, A.E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific data **2023**, 10, 1.
38. Gui, Q.; Jin, Z.; Xu, W. Exploring missing data prediction in medical monitoring: A performance analysis approach. In Proceedings of the 2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE, 2014, pp. 1–6.
39. Bickley, L.; Szilagyi, P.G. Bates' guide to physical examination and history-taking; Lippincott Williams & Wilkins, 2012.

40. Elovic, A.; Pourmand, A. MDCalc medical calculator app review. Journal of digital imaging **2019**, 32, 682–684.
41. Abdala, O.; Saeed, M. Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted K-nearest neighbors algorithm. In Proceedings of the Computers in Cardiology, 2004. IEEE, 2004, pp. 693–696.
42. Little, R.J.; Rubin, D.B. Statistical analysis with missing data; Vol. 793, John Wiley & Sons, 2019.
43. Shpitser, I.; Mohan, K.; Pearl, J. Missing Data as a Causal and Probabilistic Problem. In Proceedings of the Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands; Meila, M.; Heskes, T., Eds. AUAI Press, 2015, pp. 802–811.
44. Nawaz, S.; Cleveland, T.; Gaines, P.; Chan, P. Clinical risk associated with contrast angiography in metformin treated patients: a clinical review. Clinical radiology **1998**, 53, 342–344.
45. Liu, C. Bartlett's Decomposition of the Posterior Distribution of the Covariance for Normal Monotone Ignorable Missing Data. Journal of Multivariate Analysis **1993**, 46, 198–206.
46. Bakhtiyari, M.; Kazemian, E.; Kabir, K.; Hadaegh, F.; Aghajanian, S.; Mardi, P.; Ghahfarokhi, N.T.; Ghanbari, A.; Mansournia, M.A.; Azizi, F. Contribution of obesity and cardiometabolic risk factors in developing cardiovascular disease: a population-based cohort study. Scientific reports **2022**, 12, 1544.
47. Zamanian, A.; Mareis, L.; Ahmidi, N. Partially Specified Causal Simulations. arXiv preprint arXiv:2309.10514 **2023**.
48. Yoon, J.; Jordon, J.; Schaar, M. Gain: Missing data imputation using generative adversarial nets. In Proceedings of the International conference on machine learning. PMLR, 2018, pp. 5689–5698.
49. Jarrett, D.; Cebere, B.C.; Liu, T.; Curth, A.; van der Schaar, M. Hyperimpute: Generalized iterative imputation with automatic model selection. In Proceedings of the International Conference on Machine Learning. PMLR, 2022, pp. 9916–9937.
50. Ipsen, N.B.; Mattei, P.A.; Frellsen, J. How to deal with missing data in supervised deep learning? In Proceedings of the 10th International Conference on Learning Representations, 2022.
51. Schmier, J.K.; Halpern, M.T. Patient recall and recall bias of health state and health status. Expert review of pharmacoeconomics & outcomes research **2004**, 4, 159–163.
52. Perez Ruiz de Garibay, A.; Kortgen, A.; Leonhardt, J.; Zipprich, A.; Bauer, M. Critical care hepatology: definitions, incidence, prognosis and role of liver failure in critically ill patients. Critical Care **2022**, 26, 289.
53. Packer, C.D.; Packer, C.D. Pertinent Positives and Negatives. Presenting Your Case: A Concise Guide for Medical Students **2019**, pp. 57–71.
54. Lip, G.Y.; Nieuwlaat, R.; Pisters, R.; Lane, D.A.; Crijns, H.J. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest **2010**, 137, 263–272.
55. Apgar, V. A proposal for a new method of evaluation of the newborn infant. Anesthesia & Analgesia **1953**, 32, 260–267.
56. Zachariasse, J.M.; Seiger, N.; Rood, P.P.; Alves, C.F.; Freitas, P.; Smit, F.J.; Roukema, G.R.; Moll, H.A. Validity of the Manchester Triage System in emergency care: A prospective observational study. PloS one **2017**, 12, e0170811.
57. Limb, M. Delayed discharge: how are services and patients being affected? https://doi.org/10.1136/bmj.o118, 2022. accessed on March 27, 2024.
58. Gray, L.; Taub, N.; Khunti, K.; Gardiner, E.; Hiles, S.; Webb, D.; Srinivasan, B.; Davies, M. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. Diabetic medicine **2010**, 27, 887–895.
59. Falcoz, P.E.; Conti, M.; Brouchet, L.; Chocron, S.; Puyraveau, M.; Mercier, M.; Etievent, J.P.; Dahan, M. The Thoracic Surgery Scoring System (Thoracoscore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery. The Journal of Thoracic and Cardiovascular Surgery **2007**, 133, 325–332.
60. Aguirre, U.; García-Gutiérrez, S.; Romero, A.; Domingo, L.; Castells, X.; Sala, M.; Group, C.S. External validation of the PREDICT tool in Spanish women with breast cancer participating in population-based screening programmes. Journal of Evaluation in Clinical Practice **2019**, 25, 873–880.
61. Wishart, G.C.; Azzato, E.M.; Greenberg, D.C.; Rashbass, J.; Kearins, O.; Lawrence, G.; Caldas, C.; Pharoah, P.D. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Research **2010**, 12, 1–10.

 doi:10.20944/preprints202404.0429.v1

62. Fernando, S.M.; Fox-Robichaud, A.E.; Rochwerg, B.; Cardinal, P.; Seely, A.J.; Perry, J.J.; McIsaac, D.I.; Tran, A.; Skitch, S.; Tam, B.; et al. Prognostic accuracy of the Hamilton Early Warning Score (HEWS) and the National Early Warning Score 2 (NEWS2) among hospitalized patients assessed by a rapid response team. Critical care **2019**, 23, 1–8.
63. Blatchford, O.; Murray, W.R.; Blatchford, M. A risk score to predict need for treatment for uppergastrointestinal haemorrhage. The Lancet **2000**, 356, 1318–1321.
64. Zamanian, A. FraunhoferIKS/parcs: v1.0.0, 2023. https://doi.org/10.5281/zenodo.8322067.