Journal of Biomedical Informatics 169 (2025) 104877

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Federated causal discovery with missing data in a multicentric study on endometrial cancer

Alessio Zanga [a,b,*], Alice Bernasconi [a,c], Peter J.F. Lucas [d], Hanny Pijnenborg [e], Casper Reijnen [e], Marco Scutari [f], Anthony C. Constantinou [g]

[a] Models and Algorithms for Data and Text Mining Laboratory (MADLab), Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Milan, Italy
[b] Data Science and Advanced Analytics, F. Hoffmann - La Roche Ltd, Basel, Switzerland
[c] Evaluative Epidemiology Unit, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[d] University of Twente, Enschede, The Netherlands
[e] Radboud University Medical Center, Nijmegen, The Netherlands
[f] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland
[g] Bayesian AI Research Lab, Machine Intelligence and Decision Systems (MInDS) Research Group, School of Electronic Engineering and Computer Science (EECS), Queen Mary University of London (QMUL), London, United Kingdom

ARTICLE INFO

*Keywords:*
Federated learning
Multiple sources
Missing data

ABSTRACT

**Objectives:** Establishing causal dependencies is crucial in applied domains, such as medicine and healthcare, where decision-making must be explainable. In these settings, small sample sizes and missing data call for federated approaches to maximise the amount of information we can use.
**Methods:** We propose a novel federated causal discovery algorithm capable of pooling information from multiple sources with heterogeneous missing data to learn a graph representing cause–effect relationships. In particular, we learn a causal graph on a centralised server while taking into account both prior knowledge and missingness mechanism specific to each client.
**Results:** We applied the proposed algorithm to synthetic data and real-world data from a multicentric study on endometrial cancer, validating the obtained causal graph through quantitative analyses and a clinical literature review.
**Conclusion:** Our approach learns an accurate model despite data missing not-at-random.

## 1. Introduction

Causal discovery aims to learn a casual graph representing the underlying data generating mechanism [1,2], which is a crucial requirement in causal inference [3,4]. Tackling this issue is increasingly relevant in many fields, such as economics [5,6], psychology [7,8] and medicine [9–11]. Federated learning consists in performing distributed queries across multiple data sources and aggregating the partial results to obtain the final overall model. It is an effective solution [12–14] when we cannot pool data due to privacy policies, data regulations and integration costs.

Federated causal discovery assumes that data are complete: no missing values exist in any data source. To our knowledge, the literature has not investigated this task when data are incomplete. Causal discovery in the presence of missing data has its own set of challenging issues: not only it requires to model the data generating mechanism but also the missingness mechanism [2,15] which describes how the data are missing and why [16,17, see Section 3]. Simultaneously modelling the missingness mechanisms in multiple sources makes federated learning from incomplete data substantially different from causal discovery in a single source. Accounting for the specific missingness patterns in each data source allows for reducing bias when learning the data generating mechanism.

In this paper, we:

- Propose a novel federated causal discovery algorithm capable of dealing with missing data with different missingness mechanisms in individual sources.
- Evaluate the interaction between aggregation techniques and scoring criteria for federated causal discovery for small sample sizes.

* Corresponding author at: Models and Algorithms for Data and Text Mining Laboratory (MADLab), Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Milan, Italy.
*E-mail address:* alessio.zanga@unimib.it (A. Zanga).

https://doi.org/10.1016/j.jbi.2025.104877
Received 14 October 2024; Received in revised form 30 June 2025; Accepted 7 July 2025
Available online 22 July 2025

**Table 1**
Statement of significance.

| **Problem or Issue** |
| --- |
| Missing values introduce bias during the estimation process, especially in multiple data source settings. |
| **What is Already Known** |
| Although prior works investigated either the estimation from multiple data sources or the missing bias, none addressed these issues jointly. |
| **What This Paper Adds** |
| This study aims to relax the assumption that values are missing for the same causes across different data sources, which is an unrealistic assumption due to how data are collected in clinical practice. Additionally, this paper presents an application of the proposed approach to a multicentric clinical study on endometrial cancer, highlighting its effectiveness compared to existing solutions. |

- Simulate different scenarios to assess the impact of violations of the global missingness mechanism assumption when multiple sources are available, varying sample size, missingness type and aggregation method.
- Analyse a real-world, multicentric study on endometrial cancer (EC) involving multiple oncological clinics part of the European Network for Individualised Treatment of Endometrial Cancer (ENITEC) study and the PIpelle prospective ENDOmetrial carcinoma (PIPENDO) study.

Federated learning helps us address limited sample sizes and data missingness. Learning a causal graph capable of disentangling the interplay between the administered treatments and other covariates can support clinicians in choosing the optimal treatment for each patient, maximising the chance of survival and reducing the risk of relapse. Overall, federated causal discovery offers new insights into how these factors interact with the observed variables (see Table 1). Our experimental findings are discussed in light of previous work on this study [18–20].

The remainder of this article is organised as follows: Section 2 reports the state-of-the-art for causal discovery w.r.t. federated learning and missing data; Section 3 introduces the existing methodology, Section 4 explains the proposed approach; Section 5 reports the experiments performed to evaluate the proposed approach; Section 6 discusses the experimental findings in detail; Section 7 summarises the conclusions from previous sections. We include supplementary material in Appendix A.

## 2. Related works

Causal discovery has attracted renewed interest from the scientific community, industry, and government because of regulatory requirements on transparency and explainability. In the context of federated learning, [12] proposed the *Regret-based Federated Causal Discovery (RFCD)* algorithm to construct a causal graph from data sources by collecting only regret values from each client. This capability is essential in privacy-sensitive settings, where sharing client models could reveal information about the client data. The same authors proposed an improvement to RCFD, called *PERI* [13], that leverages distributed min–max regret optimisation. This technique provides a consistent scoring criterion that can perform causal discovery in a privacy-preserving way when paired with a score-based algorithm. [14] proposed the *Federated Causal Structure Learning (FedCSL)* algorithm that tackles the problem using a client-to-server learning strategy to distribute computations across clients. This approach is coupled with a novel weighted aggregation strategy, allowing partial score pooling without relying on an encryption layer. Authors in [21] introduced Federated GES (FedGES), a novel federated learning approach tailored to structure learning with the Greedy Equivalence Search (GES) algorithm. FedGES enhances privacy by exchanging only evolving network structures rather than model parameters or raw data. It iteratively integrates partial models from clients through structural fusion, enabling collaborative structure development while preserving data confidentiality. In the context of distributed causal discovery, [22] proposed Distributed Annealing on Regularised Likelihood Score (DARLS), a federated method for learning causal graphs across multiple clients. DARLS simulates an annealing process to explore the space of topological sorts, using distributed optimisation to determine the optimal graphical structure. Multiple rounds of communication between local clients and a central server refine the estimation, ensuring convergence to the solution an oracle would obtain with full data access. [23] introduced the FedCausal algorithm, a strategy designed to learn a unified global causal graph from heterogeneous, decentralised data. FedCausal employs a global optimisation formulation to aggregate client-specific causal graphs while enforcing acyclicity constraints without exposing local data, unifying local and global optimisation, and offering a flexible and scalable solution. Finally, [24] proposed FedCASL, a method based on continuous bi-level optimisation where clients and servers iteratively refine the causal structure while exchanging only model parameters. By incorporating a carefully designed sparse penalty term, FedCASL guides the optimisation towards a more interpretable and accurate causal graph under acyclicity constraints.

As for handling missing data, the work of [16,17] on recovering the joint missing data distribution allowed for the derivation of new causal discovery techniques. For instance, the *Missing Values PC (MVPC)* [15] algorithm deals with incomplete data following different missingness mechanisms by estimating the associated conditional independence statements to construct the causal graph. The *Hill-Climbing with Adaptive Inverse Probability Weighting (HC-aIPW)* score-based procedure proposed by [25] performs a greedy search using pairwise deletion and inverse probability weighting to reduce the bias caused by missing values. Authors in [26] introduced MissDAG, a general framework for causal discovery from incomplete data. MissDAG operates under the assumptions of ignorable missingness and identifiable additive noise models (ANMs), maximising the expected likelihood of observed data within an expectation-maximisation (EM) framework. When closed-form posterior distributions are unavailable, the likelihood is approximated using Monte Carlo EM. Extending this line of research, [27] explored the effectiveness of additive noise models for causal discovery in the presence of self-masking missingness. Their work investigates the identification problem of learning causal graphs under different missingness mechanisms, demonstrating that the traditional no self-masking missingness assumption can be appropriately relaxed.

Still, no causal discovery algorithm can deal with missing data and multiple distributed data sources. Applying existing solutions independently to each data source would result in biased estimates due to the unlikely assumption that the missingness pattern remains constant across each source.

## 3. Preliminaries

### 3.1. Representing missing data with graphs

Probabilistic graphical models [28] are probabilistic models that represent the joint probability distribution of a vector of random variables using graphs. We denote a graph $\mathcal{G}$ as a pair of sets $(\mathbf{V}, \mathbf{E})$, with $\mathbf{V}$ the set of vertices and $\mathbf{E}$ the set of edges. The parents $\boldsymbol{\Pi}_i^{\mathcal{G}}$ of a vertex $V_i$ in a graph $\mathcal{G}$ are the vertices with an edge into $V_i$.

**Definition 1** (*Probabilistic Graphical Model*)**.** A *probabilistic graphical model* (PGM) is a pair $(\mathcal{G}, \mathbf{X})$, where $\mathcal{G}$ is a graph and $\mathbf{X}$ is a random vector s.t. each vertex $V_i \in \mathbf{V}$ is associated to a random variable $X_i \in \mathbf{X}$. The graph $\mathcal{G}$ is a *structure* over the *joint probability distribution* $P(\mathbf{X})$.

Since vertices in $\mathbf{V}$ correspond to random variables in $\mathbf{X}$, we can use $V_i$ and $X_i$ interchangeably. PGMs are particularly effective when it comes to encoding dependence and independence statements of the joint probability distribution $P(\mathbf{X})$ directly into the graph $\mathcal{G}$.

**Definition 2** (*Independence Map, Dependence Map & Perfect Map*)**.** Let $\mathcal{G}$ be a graph and $P(\mathbf{X})$ be a joint probability distribution. Then, $\mathcal{G}$ is:

$$\begin{aligned}
\text{an independence map (I-Map) if:} \quad & V_i \perp\!\!\!\perp_{\mathcal{G}} V_j \mid V_k \implies X_i \perp\!\!\!\perp_P X_j \mid X_k, \\
\text{a dependence map (D-Map) if:} \quad & V_i \perp\!\!\!\perp_{\mathcal{G}} V_j \mid V_k \impliedby X_i \perp\!\!\!\perp_P X_j \mid X_k, \\
\text{a perfect map (P-Map) if:} \quad & V_i \perp\!\!\!\perp_{\mathcal{G}} V_j \mid V_k \iff X_i \perp\!\!\!\perp_P X_j \mid X_k,
\end{aligned}$$

with $\perp\!\!\!\perp_{\mathcal{G}}$ and $\perp\!\!\!\perp_P$ denoting graphical and probabilistic independence respectively.

While $\perp\!\!\!\perp_P$ denotes the usual probabilistic independence, $\perp\!\!\!\perp_{\mathcal{G}}$ refers to *graphical separation*. This property allows querying $\mathcal{G}$ to verify the validity of a given independence statement, which arises from graphical separation as defined by *d-separation* (short for "directed separation").

**Definition 3** (*d-separation*)**.** Let $\mathcal{G}$ be a directed acyclic graph (DAG) and let $\{X, Y\}$ and $\mathbf{Z}$ be two disjoint subsets of $\mathbf{V}$. Then, an undirected path from $X$ to $Y$ is said to be d-separated by $\mathbf{Z}$, denoted as $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Z}$, if it contains:

- a fork $V_i \leftarrow V_j \rightarrow V_k$ or a chain $V_i \rightarrow V_j \rightarrow V_k$ so that $V_j$ is in $\mathbf{Z}$, or
- a collider $V_i \rightarrow V_j \leftarrow V_k$ so that $V_j$, or any descendant of it, is not in $\mathbf{Z}$.

This definition of d-separation extends to sets of variables to express complex independencies.

**Definition 4** (*General d-separation*)**.** Let $\mathcal{G}$ be a DAG and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three disjoint subsets of $\mathbf{V}$. Then, $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$, denoted as $\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$, if every undirected path from $\mathbf{X}$ to $\mathbf{Y}$ is d-separated by $\mathbf{Z}$:

$$\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z} \iff X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Z} \quad \forall (X, Y) \in \mathbf{X} \times \mathbf{Y}$$

While $\mathcal{G}$ is defined as an I-map of $P$, we often treat it as a P-map when $\mathcal{G}$ is learnt from data. Therefore, we can simplify the notation of $\perp\!\!\!\perp_{\mathcal{G}}$ and $\perp\!\!\!\perp_P$ and use $\perp\!\!\!\perp$ directly. However, to encode *causal dependencies* as a graph, we need to explicitly model the *data generating mechanism*.

**Assumption 1** (*Causal Edge Assumption*)**.** Let $\mathcal{G}$ be a DAG, $\mathbf{X}$ be a random vector and $\mathbf{F}$ a set of functions. Then, the values assigned to each variable $X_i \in \mathbf{X}$ is completely determined by the function $f_i \in \mathbf{F}$ given its parents $\boldsymbol{\Pi}_i^{\mathcal{G}}$:

$$X_i := f_i\left(\boldsymbol{\Pi}_i^{\mathcal{G}}\right) \qquad \forall X_i \in \mathbf{X}$$

with $:=$ the assignment operator.

Assumption 1 establishes a functional dependency graph representing the data-generating distribution.

**Definition 5** (*Causal Graph*)**.** A causal graph $\mathcal{G}$ [29] is a graph in which Assumption 1 holds.

In turn, causal graphs imply a formal definition of cause and effect.

**Definition 6** (*Causes & Effects*)**.** Let $\mathcal{G}$ be a causal graph. Then, for each directed edge $V_i \rightarrow V_j \in \mathbf{E}$, $V_i$ is said to be a *cause* of $V_j$, whereas $V_j$ is an *effect* of $V_i$. If $V_i$ is a cause of $V_j$ and $V_j$ is a cause of $V_k$, then $V_i$ is a cause of $V_k$.

While causal graphs effectively describe why particular values are present (the data generating mechanism), they are not semantically adequate to express *why values are missing*; that is, the *missingness mechanism*. According to Rubin's classification [30], they can be:

- Missing Completely At Random (MCAR): the probability of a variable being missing is independent of both observed and unobserved variables,
- Missing At Random (MAR): the probability of a variable being missing is independent of the unobserved variables given the observed variables,
- Missing Not At Random (MNAR): neither MCAR nor MAR.

Missingness graphs [17] extend causal graphs to represent the missingness mechanism.

**Definition 7** (*Missingness Graph*)**.** A missingness graph $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ is a causal graph where the vertices in $\mathbf{V}$ are partitioned into five disjoint subsets:

$$\mathbf{V} = \mathbf{O} \cup \mathbf{U} \cup \mathbf{M} \cup \mathbf{S} \cup \mathbf{R}$$

where:

- $\mathbf{O}$ is the set of the *fully observed* variables, that is, variables with no missing values,
- $\mathbf{U}$ is the set of *fully unobserved* variables, also called the *latent* variables,
- $\mathbf{M}$ is the set of the *partially observed* variables, that is, the variables with at least one missing value,
- $\mathbf{S}$ is the set of the proxy variables, that is, the variables that are observed,
- $\mathbf{R}$ is the set of the *missingness indicators* such that :

$$S_i := f_i(M_i, R_i) = \begin{cases} m_i & \text{if} \quad r_i = 0, \\ ? & \text{if} \quad r_i = 1. \end{cases}$$

with $m_i$ the observed value of $M_i$ and " ? " a placeholder for the missing value.

Different missingness mechanisms correspond to different d-separations and thus independence statements encoded by the missingness graph, shown in Fig. 1: MCAR implies $\mathbf{O} \cup \mathbf{U} \cup \mathbf{M} \perp\!\!\!\perp \mathbf{R}$, MAR implies $\mathbf{U} \cup \mathbf{M} \perp\!\!\!\perp \mathbf{R} \mid \mathbf{O}$. In contrast, MNAR does not imply either statement. Note that $R_3$ has no parents in $\mathcal{G}_{\text{MCAR}}$, while has only fully observed parents in $\mathcal{G}_{\text{MAR}}$. The key difference between MCAR/MAR and MNAR is that $V_3$ itself is a parent of $R_3$ in $\mathcal{G}_{\text{MNAR}}$.

### 3.2. Causal discovery with missing data

When $\mathcal{G}$ is unknown, we can learn it from data and prior knowledge via *causal discovery* [1,2]. Formally, let $\boldsymbol{\mathcal{G}}$ be the set of graphs defined over the variables $\mathbf{V}$ of a dataset $D$ and $\mathcal{G}^* \in \boldsymbol{\mathcal{G}}$ be the *true but unknown* graph of the generating model of $D$.

**Definition 8** (*Causal Discovery Problem*)**.** The causal discovery problem [31] consists in recovering the *true* graph $\mathcal{G}^*$ from the set of graphs $\boldsymbol{\mathcal{G}}$ given the dataset $D$.

The cardinality of $\mathcal{G}$ grows exponentially with the number of vertices [32]. The Hill-Climbing (HC) [33,34] algorithm is one of the simplest and most computationally efficient algorithms to search it. HC traverses $\mathcal{G}$ looking for the graph $\mathcal{G}^*$ that maximises the goodness of a DAG $\mathcal{G}$ in modelling the data generating mechanism of $\mathcal{D}$, an objective function called *scoring criterion* $\mathcal{S}$:

$$\mathcal{G}^* = \arg\max_{\mathcal{G}\in\mathcal{G}} \mathcal{S}(\mathcal{G} \mid \mathcal{D}). \tag{1}$$

We can estimate $\mathcal{S}$ efficiently if the scoring criterion is *decomposable* into a local score for each vertex:

$$\mathcal{S}(\mathcal{G} \mid \mathcal{D}) = \sum_{i=0}^{|\mathbf{V}|} \mathcal{S}(V_i \mid \boldsymbol{\Pi}_i, \mathcal{D}).$$

In practice, several decomposable scoring criteria are *penalised log-likelihoods* of the form:

$$\underbrace{\mathcal{S}(V_i \mid \boldsymbol{\Pi}_i, \mathcal{D})}_{\text{Scoring Criterion}} = \underbrace{\log P(V_i \mid \boldsymbol{\Pi}_i, \Theta_i)}_{\text{Log-Likelihood}} - \underbrace{f(V_i \mid \boldsymbol{\Pi}_i, \Theta_i)}_{\text{Penalty}},$$

where $\Theta_i$ is the local conditional probability distribution parameters induced by the variables $(V_i, \boldsymbol{\Pi}_i)$ and $f$ is a non-negative function penalising model complexity. For this reason, we often use an $f$ from the model selection literature [35], see 2.

If $\mathcal{D}$ is complete, we can estimate $P(V_i \mid \boldsymbol{\Pi}_i)$ from all the samples in $\mathcal{D}$ because they contain no missing values. If $\mathcal{D}$ is incomplete, we can discard samples containing missing values via *list-wise deletion*, removing samples that contain at least one missing value when estimating all local distributions, or *pair-wise deletion*, removing samples that contain missing values in $\{V_i\} \cup \boldsymbol{\Pi}_i$ when estimating the corresponding local distribution. Deletion itself relies on the assumption that the local pair-wise deleted data distribution is an unbiased estimate of the local complete data distribution:

$$\underbrace{P(V_i \mid \boldsymbol{\Pi}_i)}_{\substack{\text{Local Complete}\\ \text{Data Distribution}}} \overset{?}{=} \underbrace{P(V_i \mid \boldsymbol{\Pi}_i, R_i = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})}_{\substack{\text{Local Pair-Wise Deleted}\\ \text{Data Distribution}}} \tag{2}$$

**Algorithm 1** Learn the parents of the missingness indicators.

**Input:** A dataset $\mathcal{D}$ and a conditional independence test $\perp\!\!\!\perp_P$.
**Output:** A map $\boldsymbol{\Pi}_{\mathbf{R}}$ with $R_i$ as keys and $\boldsymbol{\Pi}_{R_i}$ as values.

**procedure** ParentsMissingIndicators( $\mathcal{D}, \perp\!\!\!\perp_P$ )
  $\boldsymbol{\Pi}_{\mathbf{R}} \leftarrow \varnothing$ ▷ Allocate the candidate parents.
  **for** $V_i \in \mathbf{M}$ **do** ▷ Iterate over the partially observed variables.
    $\boldsymbol{\Pi}_{R_i} \leftarrow \mathbf{V} \setminus \{V_i\}$ ▷ Initialise the candidate parents.
    $k \leftarrow 0$ ▷ Initialise the cardinality of the conditioning set.
    **while** $|\boldsymbol{\Pi}_{R_i}| > k$ **do** ▷ While there are still variables...
      $\boldsymbol{\Pi}'_{R_i} \leftarrow \boldsymbol{\Pi}_{R_i}$ ▷ Copy the candidate's parents.
      **for** $V_j \in \boldsymbol{\Pi}'_{R_i}$ **do** ▷ For each candidate parent...
        **for** $\mathbf{S} \in C(\boldsymbol{\Pi}'_{R_i} \setminus \{V_j\}, k)$ **do** ▷ For each set ...
          **if** $(R_i \perp\!\!\!\perp_P V_j \mid \mathbf{S}, R_j = 0, \mathbf{R}_{\mathbf{S}} = \mathbf{0})$ **then**
            $\boldsymbol{\Pi}_{R_i} \leftarrow \boldsymbol{\Pi}_{R_i} \setminus \{V_j\}$ ▷ ...remove the parent.
            **break** ▷ Exit the inner loop.
      $k \leftarrow k + 1$ ▷ Increment the cardinality.
    $\boldsymbol{\Pi}_{\mathbf{R}} \leftarrow \boldsymbol{\Pi}_{\mathbf{R}} \cup \boldsymbol{\Pi}_{R_i}$ ▷ Store the parents of $R_i$.
  **return** $\boldsymbol{\Pi}_{\mathbf{R}}$ ▷ Return the parents of each missingness indicator.

**Algorithm 2** Learn the causal graph that maximises the scoring criterion.

**Input:** A dataset $\mathcal{D}$, a scoring criterion $\mathcal{S}$ and the parents $\boldsymbol{\Pi}_{\mathbf{R}}$.
**Output:** A causal graph $\mathcal{G}$.

**procedure** HC-aIPW( $\mathcal{D}, \mathcal{S}, \boldsymbol{\Pi}_{\mathbf{R}}$ )
  $\delta \leftarrow +\infty$ ▷ Initialise the delta score.
  $\mathcal{G} \leftarrow \varnothing$ ▷ Initialise empty graph.
  $\mathcal{G} \leftarrow \{\mathcal{G}\}$ ▷ Initialise the set of already visited graphs.
  **while** $\delta \neq 0$ **do** ▷ While the delta score is increasing...
    $\delta \leftarrow 0$ ▷ Set delta score to zero.
    $\mathcal{H} \leftarrow \mathcal{G}$ ▷ Set the current best graph.
    **for** $\mathcal{G}' \in \mathbf{Ne}(\mathcal{G}) \setminus \mathcal{G}$ **do** ▷ For each candidate graph...
      $\mathbf{Z} \leftarrow$ *Necessary variables* $\mathbf{W}$ ▷ See [25] for lines 8–10.
      **if** $\boldsymbol{\Pi}_{\mathbf{R}_\mathbf{Z}} \cap \mathbf{M} = \varnothing$ **then**
        $\mathbf{Z} \leftarrow$ *Sufficient variables* $\mathbf{U}$
      $\mathcal{D}' \leftarrow$ *Pair-wise deletion* on $\mathcal{D}$ w.r.t. $\mathbf{Z}$ ▷ Delete missing.
      $\boldsymbol{\beta} \leftarrow$ Eq. (3) on $\mathcal{D}'$ and $\boldsymbol{\Pi}_{\mathbf{R}}$ ▷ Estimate IPWs.
      $\delta' \leftarrow \mathcal{S}(\mathcal{G}' \mid \mathcal{D}', \boldsymbol{\beta}) - \mathcal{S}(\mathcal{G} \mid \mathcal{D}', \boldsymbol{\beta})$ ▷ Compute the new delta.
      **if** $\delta' > \delta$ **then** ▷ If the new delta is higher...
        $\delta \leftarrow \delta'$ ▷ ...update the current delta.
        $\mathcal{H} \leftarrow \mathcal{G}'$ ▷ ...and update the current graph.
    **if** $\delta > 0$ **then** ▷ If the best delta score is positive...
      $\mathcal{G} \leftarrow \mathcal{H}$ ▷ ...update the best graph.
      $\mathcal{G} \leftarrow \mathcal{G} \cup \{\mathcal{H}\}$ ▷ ...update the already visited graphs.
  **return** $\mathcal{G}$ ▷ Return graph with the highest score.

In general, Eq. (2) holds under MCAR but not under MAR/MNAR [36]. However, [15,16,25] have used missingness graphs to show that, when the parents of the missingness indicators $\boldsymbol{\Pi}_{\mathbf{R}}$ are known, $P(\mathbf{V})$ can be recovered from missing data as:

$$P(\mathbf{V}) = \underbrace{P(\mathbf{V} \mid \mathbf{R} = \mathbf{0})}_{\substack{1.\ \text{List-Wise Deleted}\\ \text{Data Distribution}}} \cdot \underbrace{\frac{P(\mathbf{R} = \mathbf{0})}{\prod_{i=0}^{|\mathbf{V}|} P(R_i = 0 \mid \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})}}_{\substack{2.\ \text{Missingness Indicators Distribution}}} \cdot \prod_{i=0}^{|\mathbf{V}|} \underbrace{\frac{P(\boldsymbol{\Pi}_{R_i} \mid \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})}{P(\boldsymbol{\Pi}_{R_i} \mid R_i = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})}}_{\substack{3.\ \text{Inverse Probability Weights}}} \tag{3}$$

where $\boldsymbol{\Pi}_{R_i}$ are the parents of the missingness indicator $R_i$ and $\mathbf{R}_{\boldsymbol{\Pi}_{R_i}}$ is the set of missingness indicators of $\boldsymbol{\Pi}_{R_i}$. Eq. (3) decomposes the joint probability distribution $P(\mathbf{V})$ into three terms:

1. The list-wise deleted data distribution.
2. The missingness indicators distribution, which is the probability $P(\mathbf{R} = 0)$ of the complete sample scaled by probability $P(R_i = 0 \mid \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})$ of each variable to be complete given its parents.
3. The inverse probability weights (IPWs) that account for $\boldsymbol{\Pi}_{R_i} \neq \varnothing$ in MAR/MNAR.

Since the objective is to maximise $\sum_{i=0}^{|\mathbf{V}|} \log P(V_i \mid \boldsymbol{\Pi}_i)$ and the distribution of the missingness indicators is fixed, we can approximate Eq. (3) following the graph $\mathcal{G}$:

$$\begin{aligned} P(\mathbf{V}) &\propto P(\mathbf{V} \mid \mathbf{R} = \mathbf{0}) \cdot \prod_{i=0}^{|\mathbf{V}|} \frac{P(\boldsymbol{\Pi}_{R_i} \mid \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})}{P(\boldsymbol{\Pi}_{R_i} \mid R_i = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})} \\ &= \prod_{i=0}^{|\mathbf{V}|} \left[ P(V_i \mid \boldsymbol{\Pi}_i, R_i = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0}) \cdot \frac{P(\boldsymbol{\Pi}_{R_i} \mid \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})}{P(\boldsymbol{\Pi}_{R_i} \mid R_i = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})} \right] \\ &= \prod_{i=0}^{|\mathbf{V}|} \left[ P(V_i \mid \boldsymbol{\Pi}_i, R_i = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0}) \cdot \beta_i \right]. \end{aligned}$$

The local pair-wise deleted data distribution re-weighted with IPWs is an unbiased approximation of the local complete data distribution:

$$\underbrace{P(V_i \mid \boldsymbol{\Pi}_i)}_{\substack{\text{Local Complete}\\ \text{Data Distribution}}} \propto \underbrace{P(V_i \mid \boldsymbol{\Pi}_i, R_i = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0})}_{\substack{\text{Local Pair-Wise Deleted}\\ \text{Data Distribution}}} \cdot \underbrace{\beta_i}_{\text{IPW}}. \tag{4}$$

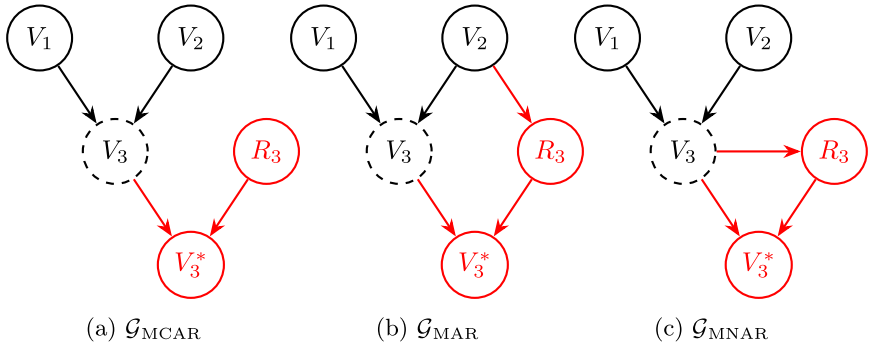


**Fig. 1.** Missingness graphs for missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), respectively.

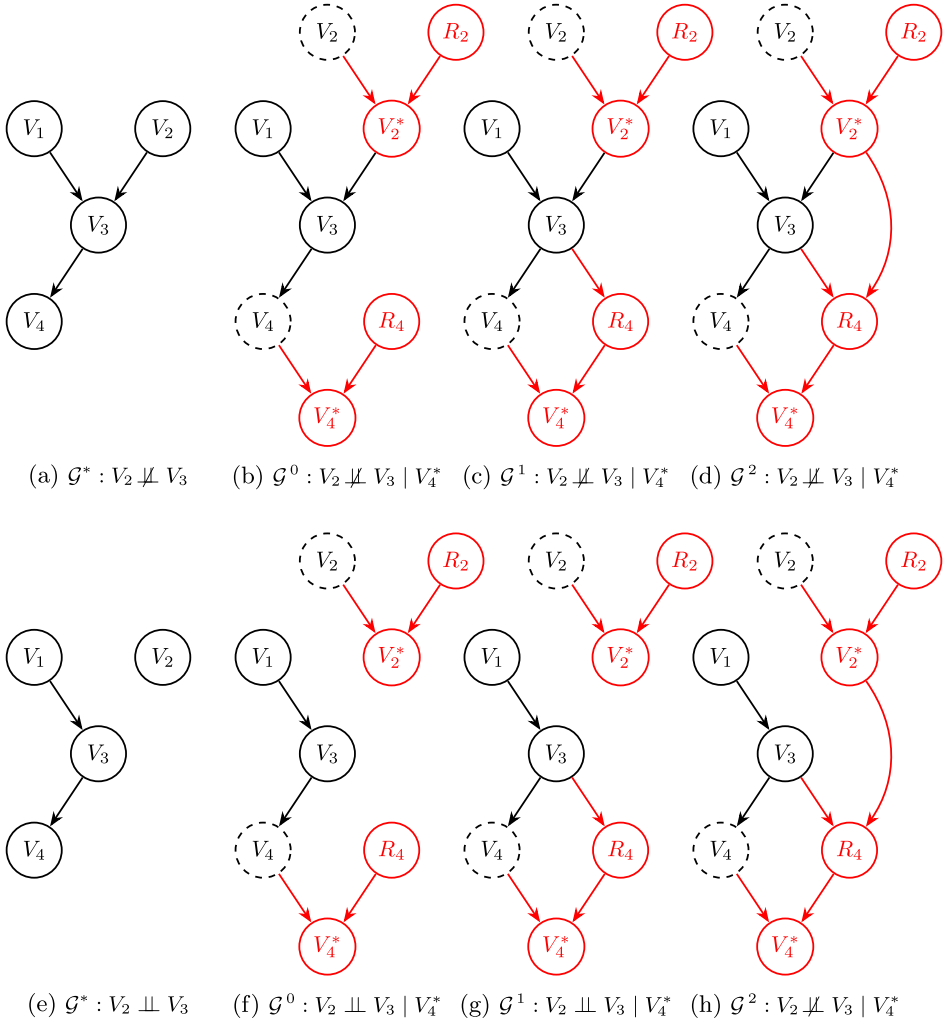


**Fig. 2.** The parents of $R_4$ change and conditioning on $R_4$ impact the independence statements.

Thanks to this unbiased estimate of $\mathcal{S}$, causal discovery in the presence of missing data simplifies to:

1. Learn the parents of the missingness indicators $\boldsymbol{\Pi}_{\mathbf{R}}$ from $\mathcal{D}$: [15] proposed to test each missingness indicator $R_i$ for independence against other variables $V_j$ conditionally on a subset of $\mathbf{V}$. Algorithm 1 efficiently performs this procedure.
2. Search the graph $\mathcal{G}^*$ that maximises $\mathcal{S}$ using Eq. (4) and $\boldsymbol{\Pi}_{\mathbf{R}}$: Algorithm 2 from [25] adapts HC to perform such procedure.

### 3.3. Source-dependant missingness mechanism

So far, we tackled the causal discovery problem where both $\mathcal{D}$ and $\boldsymbol{\Pi}_{\mathbf{R}}$ were unique. In federated learning we usually have multiple sources $\boldsymbol{\mathcal{D}} = \{\mathcal{D}^0, \mathcal{D}^1, \dots, \mathcal{D}^j, \dots, \mathcal{D}^{n-1}\}$ induced from the same underlying causal graph $\mathcal{G}^*$. Nonetheless, there could be multiple associated missingness graphs.

**Example 1** (*One Causal Graph, Multiple Missingness Graphs*)**.** Consider the graph reported in Fig. 2. For each row, the first graph on the left is the true causal graph $\mathcal{G}^*$, while the other graphs $\{\mathcal{G}^0, \mathcal{G}^1, \mathcal{G}^2\}$ are different missingness graphs associated to the same $\mathcal{G}^*$. In particular, $\mathcal{G}^0$ represents the typical MCAR mechanism where the missingness indicator $R_4$ has no parents, while $\mathcal{G}^1$ and $\mathcal{G}^2$ describe a MAR mechanism where $R_4$ does have parents. Each graph $\mathcal{G}^j$ describes a data generating mechanism that, in turn, induces a dataset $\mathcal{D}^j$ and a sample probability distribution $P^j$. Following Eq. (2), we observe:

$$P^*(V_4 \mid \boldsymbol{\Pi}_4) = P^0(V_4 \mid \boldsymbol{\Pi}_4, R_4 = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_4}} = \mathbf{0})$$

and it holds true also for $P^1$ since $\mathbf{R}_{\boldsymbol{\Pi}_{R_4}} = \varnothing$ in $\mathcal{G}^0$ and $\mathcal{G}^1$, but not for $P^2$ due to $\mathbf{R}_{\boldsymbol{\Pi}_{R_4}} = \{R_2\}$ in $\mathcal{G}^2$:

$$P^*(V_4 \mid \boldsymbol{\Pi}_4) \neq P^2(V_4 \mid \boldsymbol{\Pi}_4, R_4 = 0, \mathbf{R}_{\boldsymbol{\Pi}_{R_4}} = \mathbf{0})$$

Hence, in order to obtain an unbiased estimate of $P^*$ from a given $P^j$ using IPW as in Eq. (4), we need to model individual $\mathbf{R}^j_{\boldsymbol{\Pi}_{R_4}}$ for each data source $\mathcal{D}^j$, leading to multiple missingness graphs $\mathcal{G}^j$.

Therefore, even if a single causal graph exists, the data collection in each data source may be affected by different missingness biases, which we must model in the missingness graphs along with the respective causes.

## 4. Federated causal discovery via mixture distributions

We will formally state the federated causal discovery problem and propose a novel algorithm.

**Definition 9** (*Federated Causal Discovery Problem*)**.** The federated causal discovery problem [14] consists in recovering the *true* graph $\mathcal{G}^*$ from the set of graphs $\boldsymbol{\mathcal{G}}$ given the set of datasets $\boldsymbol{\mathcal{D}}$.

Recovering the global probability distribution $P(\mathbf{V})$ is a non-trivial task when multiple sources $\mathcal{D}^j$ with source-dependant missingness mechanism $\boldsymbol{\Pi}^j_{\mathbf{R}}$ are present. A possible solution is to model the global distribution using a mixture of local distributions:

$$\underbrace{\boldsymbol{\mathcal{D}} = \{\mathcal{D}^0, \mathcal{D}^1, \dots, \mathcal{D}^{n-1}\}}_{\text{Multiple Sources}}, \quad \underbrace{\boldsymbol{\omega} : \omega^j \geq 0, \textstyle\sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \omega^j = 1}_{\text{Mixture Weights}}, \quad \underbrace{P(\mathbf{V}) = \textstyle\sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \omega^j P^j(\mathbf{V})}_{\text{Mixture Distribution}}$$

This approach allows us to model the contribution of each source to the final global distribution flexibly by specifying the mixture weights $\boldsymbol{\omega}$, see 3. To solve the problem stated in Definition 9, we must provide the associated optimisation problem we want to solve. We recast Eq. (1) to allow for multiple sources setting via Eq. (5):

$$\underbrace{\mathcal{G}^* = \arg\max_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} P(\mathcal{G} \mid \boldsymbol{\mathcal{D}})}_{\text{Federated Causal Discovery}} = \overbrace{\arg\max_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \underbrace{\sum_{i=0}^{|\mathbf{V}|} \left[\log \omega_i^j + \log P^j(V_i \mid \boldsymbol{\Pi}_i)\right]}_{\text{Client-side Estimation}}}^{\text{Server-side Optimisation}}. \tag{5}$$

Refer to Appendix A for a detailed derivation. Eq. (5) has two parts:

- Server-side optimisation - Finding the graph $\mathcal{G}^*$ that maximises the objective function across all data sources in $\boldsymbol{\mathcal{D}}$ by aggregating the partial results computed by each client in a centralised server.
- Client-side estimation - Estimating the mixture component of a specific client independently, without transferring any data to the server.

Finally, when missing data are present, we can plug Eq. (4) into Eq. (5) to obtain an unbiased estimate of the objective function:

$$\arg\max_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \sum_{i=0}^{|\mathbf{V}|} \left[\log \omega_i^j + \log P^j(V_i \mid \boldsymbol{\Pi}_i)\right] =$$
$$\arg\max_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \sum_{i=0}^{|\mathbf{V}|} \left[\log \omega_i^j + \log P^j(V_i \mid \boldsymbol{\Pi}_i, R_i^j = 0, \mathbf{R}^j_{\boldsymbol{\Pi}_{R_i}} = \mathbf{0}) + \log \beta_i^j\right]. \tag{6}$$

To solve this optimisation problem, we propose the novel Fed-HC-aIPW algorithm, reported in Algorithm 3, that extends Algorithm 2 to a client–server configuration. Algorithm 3 is divided into four stages:

1. Server-side initialisation - Allocates the resources on the server, namely, the graph $\mathcal{G}$, the delta score $\delta$ and a cache $\boldsymbol{C}$, which minimises the number of evaluations on each data source, reducing the latency and the overall computational burden.
2. Client-side initialisation - Learn the parents of the missingness indicators $\boldsymbol{\Pi}^j_{\mathbf{R}}$ for each data source $\mathcal{D}^j$ to model the local missingness mechanisms.
3. Initial evaluation - Evaluate the initial graph by aggregating the local $\delta^j_i$.
4. Solve optimisation - Find the graph that maximises the score by evaluating *the neighbours* of $\mathcal{G}$, that is, all the graphs $\mathcal{G}'$ obtained by adding, removing or reversing an edge w.r.t. $\mathcal{G}$ and $\mathcal{K}$.

Key computational information regarding Algorithm 3:

- Lines in red, namely 6, 9–12 and 21–24, are executed on the clients and sent to the server to perform the aggregation.
- The cache $\boldsymbol{C}$ avoids redundant local computation and, in turn, the communication effort between the server and the clients.
- The neighbours $\mathbf{Ne}(\mathcal{G}, \mathcal{K})$ are the candidate graphs that can be reached by adding, deleting or reversing an edge in the current best graph $\mathcal{G}$. The prior knowledge $\mathcal{K}$ applies additional constraints to the exploration space. Such constraints are usually in the form of forbidden and required edges elicited from domain knowledge experts.
- The set $\mathbf{V}'$ identifies the local scores that must be computed. If an edge $V_i \leftarrow V_j$ is added or deleted from $\mathcal{G}$, then it will change just one parent set at the time, triggering the computation of the modified $\boldsymbol{\Pi}_i^{\mathcal{G}'}$. On the contrary, if the same edge is reversed, we must recompute the local score for the parent sets $\boldsymbol{\Pi}_i^{\mathcal{G}'}$ and $\boldsymbol{\Pi}_j^{\mathcal{G}'}$.
- Eq. (6) can be extended to the broader family of penalised log-likelihood functions by adding a penalty term based on the pairwise deleted dataset $D_i^j$. Hence, lines 11 and 23 refer to the scoring criterion $\mathcal{S}$. See 2.
- We provide asymptotic time and space complexity in Appendix B.

**Table 2**
Scoring criteria often used in the model selection context.

| Scoring criterion | Penalty function |
|---|---|
| Akaike Information Criterion (AIC) [37] | $\lvert\Theta_i\rvert$ |
| Bayesian Information Criterion (BIC) [38] | $\frac{1}{2}\cdot\log\lvert\mathcal{D}\rvert\cdot\lvert\Theta_i\rvert$ |
| AIC *with small sample size correction* (AICc) [39] | $\max\left\{\frac{\lvert\mathcal{D}\rvert}{\lvert\mathcal{D}\rvert-\lvert\Theta_i\rvert-2},1\right\}\cdot\lvert\Theta_i\rvert$ |
| BIC *with small sample size correction* (BICc) [40] | $\frac{1}{2}\cdot\max\left\{\frac{\lvert\mathcal{D}\rvert}{\lvert\mathcal{D}\rvert-\lvert\Theta_i\rvert-2},1\right\}\cdot\log\lvert\mathcal{D}\rvert\cdot\lvert\Theta_i\rvert$ |
| Hannan-Quinn Information Criterion (HQC) [41] | $\log\log\lvert\mathcal{D}\rvert\cdot\lvert\Theta_i\rvert$ |

**Algorithm 3** Learn for the causal graph that maximises the score.

**Input:** Data $\mathcal{D}$, knowledge $\mathcal{K}$, scoring $S$ and independence test $\perp\!\!\!\perp_P$.
**Output:** A causal graph $\mathcal{G}$.

**procedure** FED-HC-AIPW($\mathcal{D}, \mathcal{K}, S, \perp\!\!\!\perp_P$)
# Stage 1 - Server-side initialisation.
  $\delta \leftarrow +\infty$ ▷ Initialise the delta score.
  $\mathcal{G} \leftarrow \varnothing$ ▷ Initialise the empty graph.
  $\boldsymbol{C} \leftarrow \varnothing$ ▷ Initialise the empty cache.
# Stage 2 - Client-side initialisation.
  **for** $D^j \in \mathcal{D}$ **do** ▷ For each data source, learn the local...
    $\boldsymbol{\Pi}^j_{\mathbf{R}} \leftarrow$ Algorithm 1 $(D^j, \perp\!\!\!\perp_P)$ ▷ ... missingness mechanism.
# Stage 3 - Initial evaluation.
  **for** $V_i \in \mathbf{V}$ **do** ▷ For each vertex...
    **for** $D^j \in \mathcal{D}$ **do** ▷ For each data source...
      $D^j_i \leftarrow$ *Pair-wise deletion* on $D^j$ w.r.t. $V_i$ ▷ [25]
      $\beta^j_i \leftarrow$ Eq. (3) on $D^j_i$ and $\boldsymbol{\Pi}^j_{\mathbf{R}}$ ▷ Estimate IPWs.
      $\delta^j_i \leftarrow S(V_i \mid \varnothing, D^j_i, \beta^j_i)$ ▷ Compute the new score.
    $\boldsymbol{C}[V_i \mid \varnothing] \leftarrow \sum_{j=0}^{|\mathcal{D}|} \delta^j_i$ ▷ Aggregate scores in cache.
# Stage 4 - Solve optimisation.
  **while** $\delta > 0$ **do** ▷ While the score increases...
    $\delta \leftarrow 0$ ▷ Set the delta score to zero.
    $\mathcal{H} \leftarrow \mathcal{G}$ ▷ Set the current best graph.
    **for** $\mathcal{G}' \in \mathbf{Ne}(\mathcal{G}, \mathcal{K})$ **do** ▷ For each candidate graph...
      $\mathbf{V}' \leftarrow \{V_i \mid \boldsymbol{\Pi}^{\mathcal{G}'}_i \neq \boldsymbol{\Pi}^{\mathcal{G}}_i \quad \forall V_i \in \mathbf{V}\}$ ▷ Restrict parents.
      **for** $V_i \in \mathbf{V}'$ **do** ▷ For each vertex...
        **if** $(V_i \mid \boldsymbol{\Pi}^{\mathcal{G}'}_i) \notin \boldsymbol{C}$ **then** ▷ If the score is cached...
          **for** $D^j \in \mathcal{D}$ **do** ▷ For each data source...
            $D^j_i \leftarrow$ *Pair-wise deletion* on $D^j$ w.r.t. $V_i$ ▷ [25]
            $\beta^j_i \leftarrow$ Eq. (3) on $D^j_i$ and $\boldsymbol{\Pi}^j_{\mathbf{R}}$ ▷ IPWs.
            $\delta^j_i \leftarrow S(V_i \mid \boldsymbol{\Pi}^{\mathcal{G}'}_i, D^j_i, \beta^j_i)$ ▷ Compute the score.
          $\boldsymbol{C}[V_i \mid \boldsymbol{\Pi}^{\mathcal{G}'}_i] \leftarrow \sum_{j=0}^{|\mathcal{D}|} \delta^j_i$ ▷ Cache the score.
      $\delta' \leftarrow \sum_{i=0}^{|\mathbf{V}'|} (\boldsymbol{C}[V_i \mid \boldsymbol{\Pi}^{\mathcal{G}'}_i] - \boldsymbol{C}[V_i \mid \boldsymbol{\Pi}^{\mathcal{G}}_i])$ ▷ Compute the delta.
      **if** $\delta' > \delta$ **then** ▷ If the new delta score is higher...
        $\delta \leftarrow \delta'$ ▷ ...update the current delta score.
        $\mathcal{H} \leftarrow \mathcal{G}'$ ▷ ...update the current graph.
    $\mathcal{G} \leftarrow \mathcal{H}$ ▷ Update the best graph.
  **return** $\mathcal{G}$ ▷ Return graph with the highest score.

## 5. Experimental setup

We evaluated the proposed Algorithm 3 against a simulation study and a multicentric study on endometrial cancer, performing a grid search across:

- Scoring methods: we explored a list of candidate scoring criteria, namely `[AIC, AICC, BIC, BICC, HQC]` in Table 2, to evaluate the impact of different penalisation functions. For instance, some apply small sample size corrections to deal with distortions introduced by insufficient observations during parameter estimation.

**Table 3**
Weights aggregation methods and their formula.

| Mixture Weights | Mixture Formula |
|---|---|
| Uniform weights (GS) | $\sum_{i=0}^{\lvert\mathcal{D}\rvert} P^j(\mathbf{V})$ |
| Global weights (GW) | $\sum_{i=0}^{\lvert\mathcal{D}\rvert} f(\lvert D^j\rvert)\cdot P^j(\mathbf{V})$ |
| Local weights (LW) | $\sum_{j=0}^{\lvert\mathcal{D}\rvert}\sum_{i=0}^{\lvert\mathbf{V}\rvert} f(\lvert D^j_i\rvert)\cdot P^j(V_i \mid \Pi_i)$ |
| Local weights on *sufficient statistics* (LS) | $\sum_{j=0}^{\lvert\mathcal{D}\rvert}\sum_{i=0}^{\lvert\mathbf{V}\rvert} P^j\left(f(\lvert D^j_i\rvert)\cdot\theta_{V_i\mid\Pi_i}\right)$ |

- Mixture methods: we evaluated multiple aggregation methods, listed as `[GS, GW, LW, LS]` in Table 3, to fine-tune the granularity of the weighting scheme. We could then re-weight each local score according to the sample size of the local pair-wise deleted dataset, controlling the contribution of each partial result to the aggregated score.
- Balancing methods: we removed the missing values with `[IPW, AIPW]`. Refer to [25] for more details about the deletion process.

### 5.1. Simulation study

We conducted a simulation study generating synthetic data from the models listed in Table 4. Each data set has sample size $|D| = |\Theta|\rho$, where $|\Theta|$ is the number of parameters of the model and $\rho$ a sample ratio coefficient in $[0.1, 0.2, 0.5, 1.0, 2.0, 5.0]$. For each sample ratio, we generated a training set and a test set for in-sample and out-of-sample evaluation.

Moreover, we simulated the effect of MCAR, MAR and MNAR by generating missingness masks using the experimental setup from [25]. We repeated the data generation process five times, changing the sample size $|D|$ with a correction factor $c \in [0.5, 1.5]$ to obtain a 1-to −5 server-clients configuration. Overall, this approach simulates the common scenario where data are spread across multiple sources with both different sample sizes and missingness mechanisms. As a baseline, we also pooled clients' data into a single dataset, violating the assumption of a global missingness mechanism (which we relax in our federated proposal). Following the grid search described above, we applied Algorithm 3 and computed the F1 metric between the learned and true graphs of the selected models. Results are shown in Fig. 3. The source code can be found here.

### 5.2. Multicentric study on endometrial cancer

We analysed a case study on endometrial cancer (EC) involving the 19 gynaecological oncological clinics that are part of the European Network for Individualised Treatment of Endometrial Cancer (ENITEC) and the PIpelle prospective ENDOmetrial carcinoma (PIPENDO) study. EC is a cancer of the endometrium of the uterus. Approximately 90.000 patients die each year due to EC [42], calling for more research on personalised EC treatments. In this context, pelvic and para-aortic lymph node metastases (LNM) are among the most important prognostic factors for choosing adjuvant treatment and improving survival in node-positive EC. Approximately 10% of endometrial cancer patients present lymph node metastases at diagnosis according to clinical literature [42]. Clinical experts selected the variables that they considered

**Table 4**
Summary statistics of the reference models sorted by $|\Theta|$. The average Markov blanket (Avg. MB) and degree (Avg. Deg.) are computed as in [28].

| Model | Parameters | Vertices | Edges | Avg. M.B. | Avg. Deg. |
|---|---|---|---|---|---|
| ALARM | 509 | 37 | 46 | 3.51 | 2.49 |
| WIN95PTS | 574 | 76 | 112 | 5.92 | 2.95 |
| INSURANCE | 1008 | 27 | 52 | 5.19 | 3.85 |
| HAILFINDER | 2656 | 56 | 66 | 3.54 | 2.36 |

**Table 5**
Variables in the data sources. Variables above the horizontal line are measured preoperatively, and those below the gap postoperatively.

| Variable | Abbreviation | Tier |
|---|---|---|
| Gynaecological clinic | Hospital | 1 |
| Preoperative cervical cytology | Cytology | 1 |
| Preoperative tumour grade | PreoperativeGrade | 0 |
| Cancer Antigen 125 serum levels | CA125 | 1 |
| CT or MRI diagnostic imaging | CTMRI | 1 |
| Estrogen receptor levels | ER | 1 |
| Progesterone receptor levels | PR | 1 |
| L1 cell adhesion molecule | L1CAM | 1 |
| p53 tumour suppressor gene | p53 | 1 |
| Platelets in blood | Platelets | 1 |
| Postoperative tumour grade | PostoperativeGrade | 2 |
| Lymphovascular space invasion | LVSI | 2 |
| (Abdominal) lymph node metastases | LNM | – |
| Tumour invasion of myometrium | MyometrialInvasion | 2 |
| Treatment by chemotherapeutic drugs | Chemotherapy | 2 |
| Treatment by radiation | Radiotherapy | 2 |
| Recurrence of the tumour | Recurrence | 3 |
| Survival of at least $i$ years | Survival$i$yr, $i \in \{1,3,5\}$ | 4 |

most important for predicting survival and the presence of LNMs [18–20,43].

Table 5 reports variables collected at the different gynaecological clinics where the patients were treated: the cytology of the cervix uteri, the preoperative tumour grade, the postoperative tumour grade (after pathological examination of the tumour tissue obtained after surgical removal of the uterus), treatment by chemotherapy or radiotherapy, lymphovascular space invasion (that is, whether there is tumour growth into the lymph or blood vessels), the levels of estrogen and progesterone in the blood, the presence of lymph node metastasis according to CT or MRI imaging, the CA125 tumour marker, L1CAM (an intracellular protein that promotes tumour cell motility), the p53 tumour suppressor gene, the number of platelets, presence of lymph node metastases, recurrence of the tumour, and lastly the survival before and after 1, 3, and 5 years. The tumour markers, such as p53, CA125, L1CAM, estrogen and progesterone levels, are thought to offer causal prognostic information about tumour cell behaviour and thus tumour in-growth, metastases, recurrence, and survival.

To incorporate our prior knowledge of the temporal order of the variable measurements, we assigned each variable a tier that determines if the value assignment of a given variable happens before or after another one.

Following [11], for each combination of scoring method, mixture method and balancing method, we perform four different analyses:

1. *Single-source analysis.* To evaluate the impact of the heterogeneity and sample sizes of the data sources, we applied Algorithm 3 to each source separately instead of using it as a federated algorithm. Fig. 4 represent two selected instances of this exploratory analysis.
2. *Inference-based analysis.* For each learned causal graph, we computed the Node-Average Likelihood (NAL) [44] as an estimate of the average likelihood for each node based on the incomplete data. Fig. 5 reports the in-sample and out-of-sample NAL for each simulation scenario.
3. *Sensitivity analysis.* We estimated the average confidence by applying conditional independence testing to test whether we can remove each edge in each graph. Lower p-values translate to higher confidence that the arc should be retained, so we show the complement to 1 of the average across all p-values in each graph in Fig. 5.
4. *Predictive-based analysis.* LNM is clinically relevant when evaluating different treatment strategies. Therefore, predicting its status in early-stage patients is essential for personalised treatment. Fig. 6 report associated the area under the curve (AUC) [45].

Except for the single-source analysis, we performed a 10-fold cross-validation stratified on the Hospital data source identifier to obtain in-sample and out-of-sample estimates of our metrics while guaranteeing a representative patient case mix across each fold. The final results are min–max scaled Figs. 5 and 6 to facilitate comparisons across different analyses.

### 5.3. Unavailability of reference baselines

As for the baselines, none of the federated causal discovery algorithms cited in Section 2 can handle missing data. For the non-federated ones, we resorted to evaluating the existing options by pooling data together. The work from [27] does not provide code to run the proposed algorithm. The MissDAG [26] algorithm assumes ignorable missingness, which is incompatible with the fundamental assumption of non-ignorable missingness. The MVPC [15] algorithm implementation supports binary variables only, which are not representative of the current experimental setting. Ultimately, the only available baseline is the HC-aIPW [25] algorithm, which we extended to allow for prior knowledge and reported in the following section. This lack of direct comparability with existing methods limits the validity of the proposed approach.

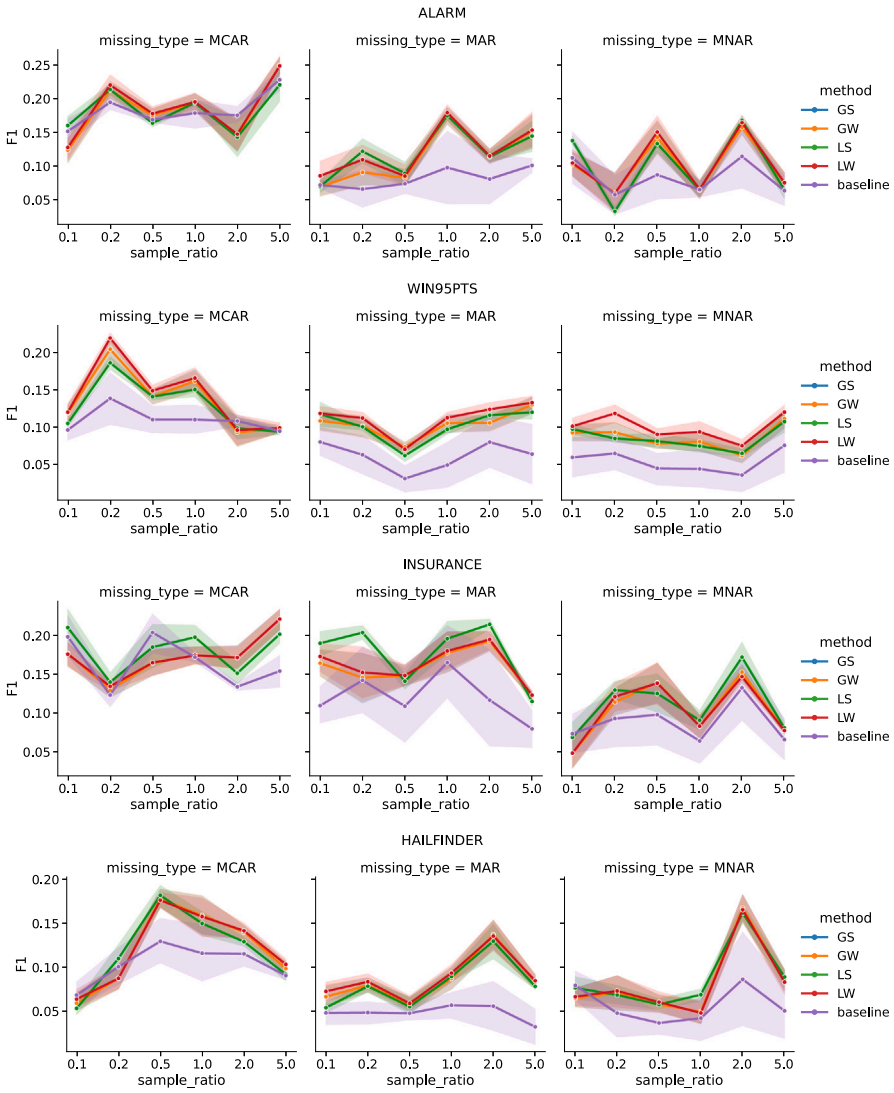


**Fig. 3.** Results of the simulation study. Rows and columns correspond to the reference models in Table 4 and the type of missingness mechanism, respectively. The legend reports the different mixture weights, such as uniform weights (GS), global weights (GW), local weights (LW) and local weights on sufficient statistics (LS), as defined in Table 3. Higher values correspond to better models.

## 6. Results and discussion

### 6.1. Simulation study

The results of the simulation study are summarised in Fig. 3. We reported the F1 score obtained by comparing the edges of the learned causal graph to the edges of the reference causal graph under different missingness mechanisms. As we can see, the proposed method can recover the underlying causal graph better than the baseline. The difference is statistically significant in most simulated scenarios but more marked in the MAR and MNAR settings. In MCAR, the parent sets of the missingness indicators are always empty, making them indistinguishable across the clients. Hence, the IPW weights are constant, but the mixture weights might differ due to a frequency of client-specific missingness for some variables. However, the parent sets of the missingness indicators can contain different variables for each client in MAR and MNAR: IPW weights then vary across clients, leading to significant differences compared to the baseline. This discrepancy is more pronounced for complex reference models, where more variables are subject to different missingness mechanisms. Interestingly, the aggregation method does not seem to affect the results, which could imply that learning local parent sets for the missingness indicators contributes the most to the unbiased estimation of the objective function.

### 6.2. Multicentric study on endometrial cancer

The exploratory single-source evaluation highlighted several limitations of the non-federated approach. For instance, the data from the 5th Hospital result in a graph in which all variables except `LVSI` are connected. However, the data from the 9th Hospital result in a graph with multiple, disconnected components (Fig. 4). In particular, the survival nodes are disconnected from the rest of the graph, which is inconsistent with the causal interpretation supported by the prior knowledge elicited from clinicians. These differences between data sources can be explained by the upper bound on the size of parent sets for penalised log-likelihood scores and small samples [46]. Missing values and pair-wise deletion further exacerbate this issue.

As for the federated evaluation, Fig. 5 shows that aggregation and scoring impact the average NAL in the inference-based analysis. In particular, `LS` outperforms other aggregation methods along with `HQC`, which also has stronger theoretical guarantees for model selection [35]. We do not observe any significant difference between different balancing methods. The combination of these two effects is evident in Fig. 6,

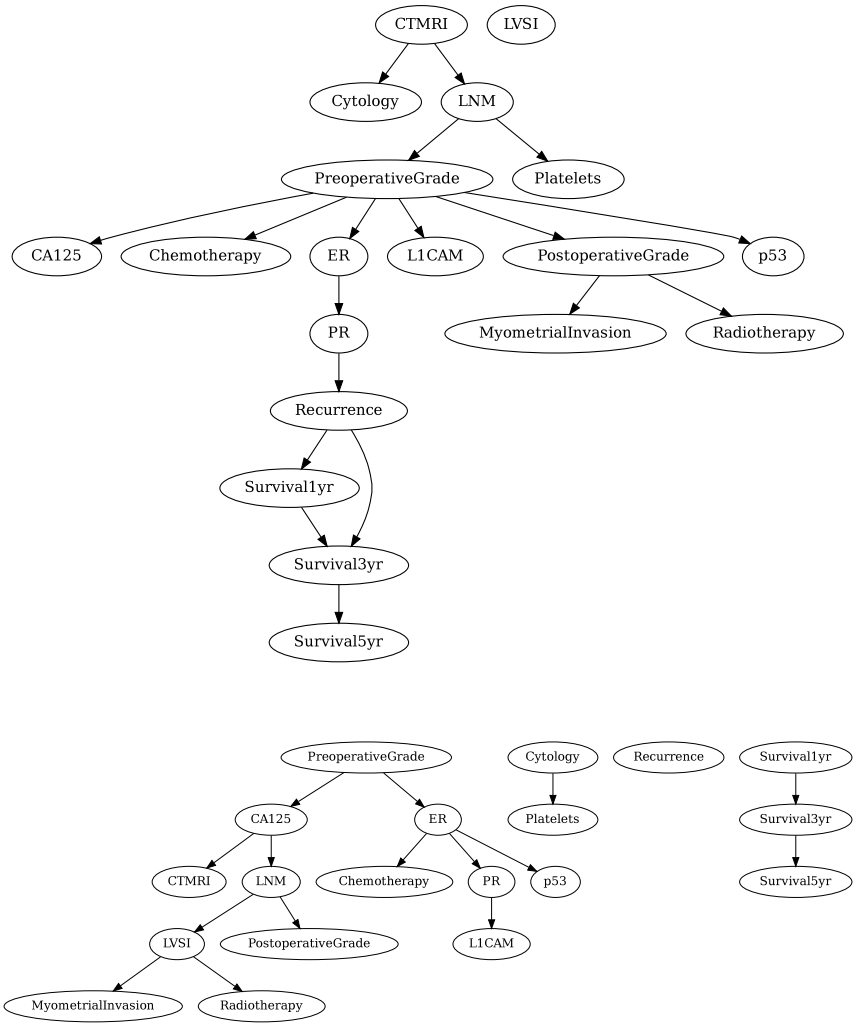


**Fig. 4.** Single-source causal graph learnt from the 5th and 9th Hospital only.

where the models that achieve both high in-sample and out-of-sample AUC are obtained by combining LS and HQC.

Finally, the estimated average confidence in Fig. 5 is an approximate indicator of the quality of the learned graphs. The difference between the in-sample (60% to 80%) and out-of-sample (20% to 40%) confidence is expected and arises from the size disparity between training and test folds in 10-fold cross-validation. Roughly, we can interpret it as an indicative estimate of how many edges are well-supported by the out-of-sample folds, that is, of how many of those edges we would expect to learn correctly. Interestingly, LS and HQC result in models with higher scores but lower average confidence, as observed in other real-world case studies [11].

### 6.3. Validation of the causal relationships

In our application of causal discovery, the obtained graph summarises the causal relationship learned by combining data and prior knowledge. Each edge points from a cause to its effect, defining the data-generating mechanism. In this case, validating the causal relationships means verifying that the learned graph matches the underlying data-generating mechanism.

In the case of synthetic experiments, we generated the data by sampling directly from the reference models reported in Table 4 and removing the data as in [25]. Thus, the true graph is given by the graph of each reference model.

In real-world applications, the true graph is usually not available. Here, causal discovery is primarily used to build a representation of the multivariate interactions between the observed variables and uncover previously unknown causal relationships.

Without ground truth, causal graphs learned from observational data can be validated via prior knowledge. For instance, involving experts, such as clinicians, during the validation of the model is a crucial part of the knowledge elicitation. In fact, the actual validation of a causal model strictly depends on the specific problem and the set of assumptions underlying the learning step.

Still, when prior knowledge is not available, some high-level quantitative validation pipelines can be found in the literature [11]:

- Model averaging - This step involves learning multiple models by applying bootstrap resampling and averaging them to estimate the confidence in the obtained edges as a form of non-parametric sensitivity analysis.
- Inference and prediction - The learned model can be validated by evaluating observational, interventional and counterfactual queries.
- Case study knowledge - Literature reviews provided by experts can be used to validate the causal claims entailed by the graph. This process is usually carried out in a multidisciplinary setting, where experts provide a set of statements that the model must satisfy.

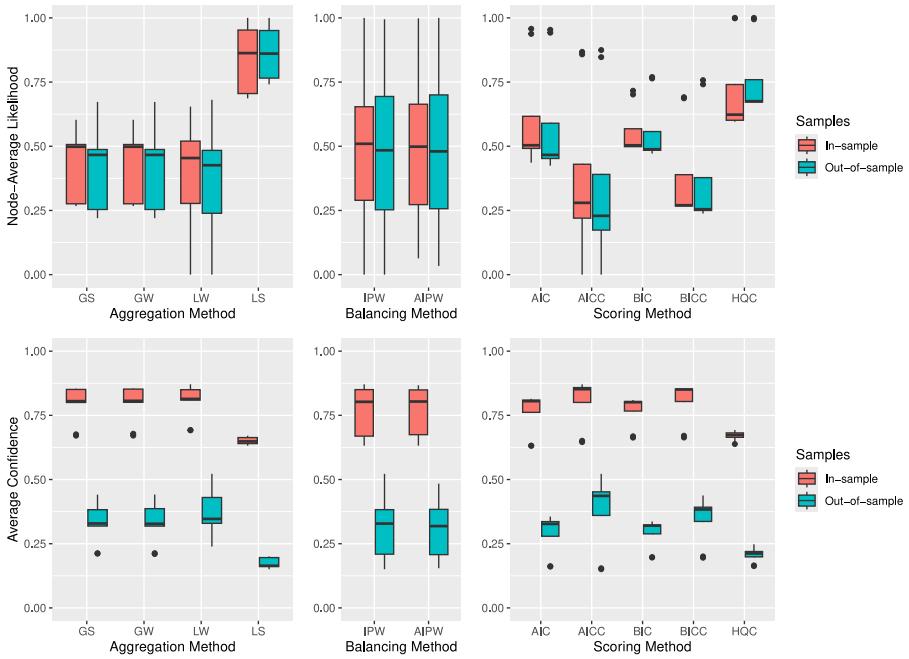


**Fig. 5.** In-sample and out-of-sample NAL and confidence by aggregation, balancing and scoring method. Higher values correspond to better models.

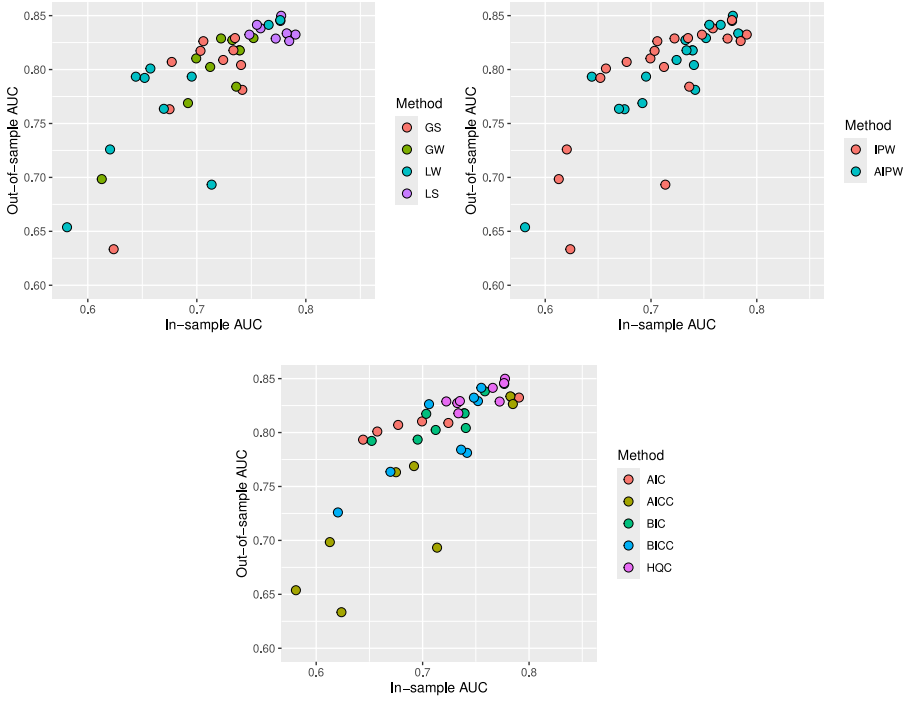


**Fig. 6.** In-sample and out-of-sample AUC by aggregation, balancing and scoring method.

In our case study on endometrial cancer, we provided inference and predictive-based validation, as in Figs. 4 to 6. Extensive validation of the causal claims using experts' knowledge will be done in future work.

### 6.4. Limitations of the proposed approach

Although the proposed approach overcomes the assumption that every data source has the same missingness mechanism, in real-world use cases it may be that others factors limits its applicability.

An implicit assumption is that every client must collect the same set of variables to be able to estimate the local scores. This is a common assumption in the federated algorithms cited in Section 2. Some non-federated causal discovery studies investigated the case where each source might have a different set of variables, a setting called "non-identical variables sets" or "partial overlapping variables" [47–49].

Another limitation, closely related to the previous, is given by the assumption of causal sufficiency [31] that unobserved variables do not impact the data-generating mechanism under study. Such an assumption may be appropriate in controlled environments, where external factors can be treated as noise factors, but is trivially false in real-world case studies, where partial observability is the norm.

## 7. Conclusions

In this paper, we proposed the first federated causal discovery algorithm for learning a global causal graph that accounts for the local missing data distributions. We extended existing causal discovery algorithms for missing data to the federated learning setting, relaxing the underlying assumption of a global missing mechanism. Unlike state-of-the-art approaches, we can thus model multiple sources and their local missingness mechanisms independently.

We performed a simulation study by generating synthetic data from well-studied reference models to investigate the properties of the proposed method. We assessed the impact of violating the global missing mechanism assumption by comparing our aggregation methods against the naïve approach of pooling data together. Results show a significant improvement in our ability to recover the underlying causal graph when this assumption is relaxed, especially for the more complex reference models.

We explored a case study on endometrial cancer involving multiple gynaecological oncological clinics part of the European Network for Individualised Treatment of Endometrial Cancer (ENITEC) study and the PIpelle prospective ENDOmetrial carcinoma (PIPENDO) study. We evaluated the proposed approach with clinicians against single-source analysis, inference and predictive-based analyses and an overall sensitivity analysis. Future work involves extensive validation of the causal claims using experts' knowledge.

Still, the current approach has limitations that hinder the applicability of federated causal discovery to real-world scenarios. For instance, it would be interesting to explore settings where data sources do not share the same set of observed variables or show significant distribution shifts due to selection bias.

## CRediT authorship contribution statement

**Alessio Zanga:** Writing – original draft, Software, Methodology, Conceptualization. **Alice Bernasconi:** Writing – review & editing. **Peter J.F. Lucas:** Validation, Resources, Data curation. **Hanny Pijnenborg:** Validation, Resources, Data curation. **Casper Reijnen:** Validation, Resources, Data curation. **Marco Scutari:** Writing – review & editing. **Anthony C. Constantinou:** Writing – review & editing, Project administration.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alessio Zanga reports financial support was provided by F Hoffmann-La Roche Ltd. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

Alessio Zanga is funded by F. Hoffmann-La Roche Ltd.

## Appendix A. Derivations for Eq. (5)

$$
\begin{aligned}
\mathcal{G}^* &= \arg\max_{\mathcal{G}\in\mathcal{G}} P(\mathcal{G} \mid \boldsymbol{\mathcal{D}}) = \\
&= \arg\max_{\mathcal{G}\in\mathcal{G}} \sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \left[\omega^j \cdot P(\mathcal{G} \mid \mathcal{D}^j)\right] = \\
&= \arg\max_{\mathcal{G}\in\mathcal{G}} \sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \prod_{i=0}^{|\mathbf{V}|} \left[\omega_i^j \cdot P^j(V_i \mid \boldsymbol{\Pi}_i)\right] = \\
&= \arg\max_{\mathcal{G}\in\mathcal{G}} \sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \log \prod_{i=0}^{|\mathbf{V}|} \left[\omega_i^j \cdot P^j(V_i \mid \boldsymbol{\Pi}_i)\right] = \\
&= \arg\max_{\mathcal{G}\in\mathcal{G}} \sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \sum_{i=0}^{|\mathbf{V}|} \log \left[\omega_i^j \cdot P^j(V_i \mid \boldsymbol{\Pi}_i)\right] = \\
&= \arg\max_{\mathcal{G}\in\mathcal{G}} \sum_{j=0}^{|\boldsymbol{\mathcal{D}}|} \sum_{i=0}^{|\mathbf{V}|} \left[\log \omega_i^j + \log P^j(V_i \mid \boldsymbol{\Pi}_i)\right]
\end{aligned}
$$

## Appendix B. Asymptotic time and space complexity

We provide asymptotic time and space complexity with big $\mathcal{O}$ notation for the worst-case scenario assuming:

- $\mathcal{G}$ is a $n \times n$ adjacency matrix,
- $\boldsymbol{\mathcal{D}}$ contains $p$ clients, in which the client with the maximum number of observations contains $m$ samples for $n$ random variables.
- $\boldsymbol{\mathcal{K}}$ is a $n \times n$ matrix representing the forbidden and required edges provided by prior knowledge.
- $\mathcal{S}$ and $\perp\!\!\!\perp_P$ are the scoring criterion and the conditional independence test for categorical random variables, respectively.

The asymptotic time complexity is $\mathcal{O}(p \cdot m \cdot n \cdot 2^n \cdot (n + a_n))$, given by:

- Stage 1 - The server-side initialisation sets the initial values of the global solution, taking $\mathcal{O}(1)$ time.
- Stage 2 - The client-side initialisation requires to compute the parents of the missingness indicators $\boldsymbol{\Pi}_{\mathbf{R}}^j$ for each client and each variable with missing values. Since we assumed we are testing for conditional independence with categorical random variables, the sufficient statistics for $V_i \perp\!\!\!\perp_P V_j \mid \mathbf{S}$ include the computation of the conditional counts $\mathbf{N}_{i,j|\mathbf{k}}$. The counts depend on (i) the number of categories of each random variable and (ii) the cardinality of the conditioning set, scaling exponentially as $\mathcal{O}(2^n)$ in the worst case where $\mathbf{S}$ is $\mathbf{V} \setminus \{V_i, V_j\}$. The tests are then repeated for each client in $\boldsymbol{\mathcal{D}}$ and for each $(R_i, V_j)$ pair, with time complexity of $\mathcal{O}(p \cdot m \cdot n^2 \cdot 2^n)$.
- Stage 3 - Initially, the algorithm evaluates the score once for each variable and each client. Each score evaluation needs to (i) apply pair-wise deletion to obtain $\mathcal{D}_i^j$, (ii) estimate the IPWs $\beta_i^j$ and (iii) compute the score $\delta_i^j$. The pair-wise deletion takes $\mathcal{O}(m)$ time to delete the samples containing missing values. The estimation of the inverse probability weights requires to compute the third term of Eq. (3), where both the numerator and denominator depend on $\boldsymbol{\Pi}_{R_i}$, which in turn relies on the joint counts $\mathbf{N}_{\boldsymbol{\Pi}_{R_i}}$. Since in the worst case $\boldsymbol{\Pi}_{R_i}$ is $\mathbf{V} \setminus \{V_i\}$, this step takes $\mathcal{O}(m \cdot 2^n)$ time. Initialising the penalised log-likelihood score relies on the computation of the marginal counts $\mathbf{N}_i$ that takes $\mathcal{O}(m)$. Finally, summing up the contribution of each step and taking into account that we need to repeat it for each variable across and for each client, we obtain a worst-case time complexity of $\mathcal{O}(p \cdot m \cdot n \cdot 2^n)$.
- Stage 4 - The solution of the optimisation problem repeats Stage 3, increasing the cardinality of the conditioning set incrementally. In the worst-case scenario, where the solution is given by a complete DAG, we need to evaluate all the DAGs with $n$ vertices. This number is given by the following recurrence relation [50]:

$$a_n = \sum_{k-1}^{n} (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k}$$

  The presence of a cache allows us to compute the score of each $(V_i, \boldsymbol{\Pi}_i)$ pair exactly once. If we implement the cache $\boldsymbol{C}$ using a hash map, we can retrieve the scores we already computed in $\mathcal{O}(n)$ time, provided that the time to access the value is amortised constant and that to compute the key is linear in the number of vertices. The time complexity for this final step is $\mathcal{O}(p \cdot m \cdot n \cdot 2^n \cdot a_n)$.

The asymptotic space complexity is $\mathcal{O}(n^2 + p \cdot 2^n)$, given by:

- Stage 1 - The server-side initialisation allocates memory for the delta score, the graph and the cache. The graph requires $\mathcal{O}(n^2)$ allocations for the adjacency matrix. The cache is initially empty and will grow linearly with the number of score evaluations.
- Stage 2 - The client-side initialisation allocates memory for the parents of the missingness indicators $\boldsymbol{\Pi}_{\mathbf{R}}^{j}$ and the independence test $\perp\!\!\!\perp_P$. Each client keeps track of the sets of parents of each missingness indicator, requiring $\mathcal{O}(n^2)$ allocations. As for the time complexity, we need to compute the conditional counts $\mathbf{N}_{i,j|\mathbf{k}}$, with $\mathcal{O}(p \cdot 2^n)$ allocations each time we perform a test.
- Stage 3 - The initial evaluation allocates memory for the scoring criterion. Similarly to the time complexity, we need to compute the joint counts $\mathbf{N}_{\boldsymbol{\Pi}_{R_i}}$, obtaining a worst-case space complexity of $\mathcal{O}(p \cdot 2^n)$.
- Stage 4 - Finally, the solution of the optimisation problem repeats Stage 3, allocating $\mathcal{O}(p \cdot 2^n)$ for each evaluation. In the worst-case scenario, the cache $\boldsymbol{C}$ grows to $\mathcal{O}(2^n)$ for each client.

While both asymptotic time and space complexity are more than exponential, there are some practical considerations that we must take into account:

- The worst-case complexity is in line with other solutions present in the literature [34].
- The number of parameters strictly depends on the type of probability distribution. For instance, the parameters and the sufficient statistics of the conditional Gaussian distribution are polynomial in size.
- These bounds take into account the *overall* complexity, both on the server and the clients. The term $p$ can be dropped if the algorithm is executed asynchronously, as in most federated scenarios.
- The terms $2^n$ and $a_n$ can be bounded by fixing the maximum number of parents for each variable to $k$, where $k$ can be derived from the penalisation term of the scoring criterion. Refer to Theorem 3 in [46].

## References


[1] Alessio Zanga, Elif Ozkirimli, Fabio Stella, A survey on causal discovery: Theory and practice, Internat. J. Approx. Reason. 151 (2022) 101–129.

[2] Neville Kenneth Kitson, Anthony C. Constantinou, Zhigao Guo, et al., A survey of Bayesian network structure learning, Artif. Intell. Rev. 56 (8) (2023) 8721–8814.

[3] Judea Pearl, Madelyn Glymour, Nicholas P. Jewell, Causal Inference in Statistics: a Primer, John Wiley & Sons, 2016.

[4] M.A. Hernán, J.M. Robins, Causal Inference: What If? CRC Press, USA, 2020.

[5] Peter Martey Addo, Christelle Manibialoa, Florent McIsaac, Exploring nonlinearity on the CO2 emissions, economic production and energy use nexus: A causal discovery approach, Energy Rep. 7 (2021) 6196–6204.

[6] Emanuele Cavenaghi, Alessio Zanga, Fabio Stella, Markus Zanker, Towards a causal decision-making framework for recommender systems, ACM Trans. Recomm. Syst. 2 (2) (2024) 1–34.

[7] Justin J. Anker, Erich Kummerfeld, Alexander Rix, et al., Causal network modeling of the determinants of drinking behavior in comorbid alcohol use and anxiety disorder, Alcohol.: Clin. Exp. Res. 43 (1) (2019) 91–97.

[8] Kathleen Miley, Piper Meyer-Kalos, Sisi Ma, et al., Causal pathways to social and occupational functioning in the first episode of schizophrenia: Uncovering unmet treatment needs, Psychol. Med. (2021) 1–9.

[9] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, et al., Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology, Sci. Rep. 10 (1) (2020) 2975.

[10] A. Bernasconi, A. Zanga, P.J.F. Lucas, et al., Towards a transportable causal network model based on observational healthcare data, in: CEUR Workshop Proceedings, vol. 3578, 2023, pp. 122–129.

[11] Anthony Constantinou, Neville K. Kitson, Yang Liu, et al., Open problems in causal structure learning: A case study of COVID-19 in the UK, Expert Syst. Appl. 234 (2023) 121069.

[12] Osman Mian, David Kaltenpoth, Michael Kamp, Regret-based federated causal discovery, in: Thuc Duy Le, Lin Liu, Emre Kıcıman, et al. (Eds.), Proceedings of the KDD'22 Workshop on Causal Discovery, in: Proceedings of Machine Learning Research, vol. 185, PMLR, 2022, pp. 61–69.

[13] Osman Mian, David Kaltenpoth, Michael Kamp, Jilles Vreeken, Nothing but regrets - privacy-preserving federated causal discovery, in: Francisco Ruiz, Jennifer Dy, Jan-Willem van de Meent (Eds.), Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 206, PMLR, 2023, pp. 8263–8278.

[14] Xianjie Guo, Kui Yu, Lin Liu, Jiuyong Li, FedCSL: A scalable and accurate approach to federated causal structure learning, Proc. AAAI Conf. Artif. Intell. 38 (11) (2024) 12235–12243.

[15] Ruibo Tu, Kun Zhang, Paul Ackermann, et al., Causal discovery in the presence of missing data, 2018.

[16] Karthika Mohan, Judea Pearl, Jin Tian, Graphical models for inference with missing data, in: Advances in Neural Information Processing Systems, 2013.

[17] Karthika Mohan, Judea Pearl, Graphical models for processing missing data, J. Amer. Statist. Assoc. 116 (534) (2021) 1023–1037.

[18] C. Reijnen, E. Gogou, L. van der Putten, et al., Development and validation of an endometrial carcinoma preoperative Bayesian network using molecular and clinical biomarkers (ENDORISK): an ENITEC collaboration study, 2019.

[19] A. Zanga, A. Bernasconi, P.J.F. Lucas, et al., Risk assessment of lymph node metastases in endometrial cancer patients: A causal approach, HC@AIxIA, in: Proceedings of the 1st Workshop on Artificial Intelligence for Healthcare, vol. 3307, 2022.

[20] Alessio Zanga, Alice Bernasconi, Peter J.F. Lucas, et al., Causal discovery with missing data in a multicentric clinical study, in: Proceedings of the 21st International Conference of Artificial Intelligence in Medicine, AIME, vol. 13897 LNAI, 2023, pp. 40–44.

[21] Pablo Torrijos, José A. Gámez, José M. Puerta, FedGES: A federated learning approach for Bayesian network structure learning, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 15244 LNAI, Springer Science and Business Media Deutschland GmbH, 2025, pp. 83–98.

[22] Qiaoling Ye, Arash A. Amini, Qing Zhou, Federated learning of generalized linear causal networks, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[23] Dezhi Yang, Xintong He, Jun Wang, et al., Federated causality learning with explainable adaptive optimization, Proc. AAAI Conf. Artif. Intell. 38 (15) (2024) 16308–16315.

[24] Chunlong Liu, Shunge Wang, Zhaojun Wang, et al., Federated causal structure learning with a bi-level optimization model, in: 14th IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, CYBER 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 708–713.

[25] Yang Liu, Anthony C. Constantinou, Greedy structure learning from data that contain systematic missing values, Mach. Learn. 111 (10) (2022) 3867–3896.

[26] Erdun Gao, Ignavier Ng, Mingming Gong, et al., MissDAG: Causal discovery in the presence of missing data with continuous additive noise models, 2022.

[27] Jie Qiao, Zhengming Chen, Jianhua Yu, et al., Identification of causal structure in the presence of missing data with additive noise model, Proc. AAAI Conf. Artif. Intell. 38 (18) (2024) 20516–20523.

[28] Daphne Koller, Nir Friedman, Probabilistic graphical models: Principles and techniques, in: Journal of Chemical Information and Modeling, The MIT Press, USA, 2009.

[29] Elias Bareinboim, Juan D. Correa, Duligur Ibelind, Thomas Icard, On Pearl's hierarchy and the foundations of causal inference, in: Probabilistic and Causal Inference: The Works of Judea Pearl, 2020, pp. 1–62.

[30] Donald D. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[31] Peter Spirtes, Clark N. Glymour, Richard Scheines, David Heckerman, Causation, prediction, and search, MIT Press, USA, 2000.

[32] Frank Harary, Edgar Palmer, Graphical enumeration, in: Graphical Enumeration, Elsevier, 1973.

[33] David Maxwell Chickering, Optimal structure identification with greedy search, J. Mach. Learn. Res. 3 (3) (2003) 507–554.

[34] Marco Scutari, Claudia Vitolo, Allan Tucker, Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation, Stat. Comput. 29 (5) (2019) 1095–1108.

[35] Gerda Claeskens, Nils Lid Hjort, Model Selection and Model Averaging, Cambridge University Press, 2001.

[36] Marco Scutari, Bayesian network models for incomplete and dynamic data, Stat. Neerl. 74 (3) (2020) 397–419.

[37] Hirotugu Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.

[38] Gideon Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (2) (1978) 461–464.

[39] Clifford M. Hurvich, Chih-Ling Tsai, Regression and time series model selection in small samples, Biometrika 76 (2) (1989) 297–307.

[40] Allan D. McQuarrie, A small-sample correction for the Schwarz SIC model selection criterion, Statist. Probab. Lett. 44 (1) (1999) 79–86.

[41] E.J. Hannan, B.G. Quinn, The determination of the order of an autoregression, J. R. Stat. Soc. Ser. B Stat. Methodol. 41 (2) (1979) 190–195.

[42] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin. 68 (6) (2018) 394–424.

[43] Casper Reijnen, Evangelia Gogou, Nicole C.M. Visser, et al., Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study, PLoS Med. 17 (5) (2020).
[44] Tjebbe Bodewes, Marco Scutari, Learning Bayesian networks from incomplete data with the node-average likelihood, Internat. J. Approx. Reason. 138 (2021) 145–160.
[45] Andrew P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159.
[46] Cassio P. de Campos, Mauro Scanagatta, Giorgio Corani, Marco Zaffalon, Entropy-based pruning for learning Bayesian networks using BIC, Artificial Intelligence 260 (2018) 42–50.
[47] Robert E. Tillman, Frederick Eberhardt, Learning causal structure from multiple datasets with similar variable sets, Behaviormetrika 41 (1) (2014) 41–64.
[48] Sofia Triantafillou, Ioannis Tsamardinos, Constraint-based causal discovery from multiple interventions over overlapping variable sets, J. Mach. Learn. Res. 16 (2015) 2147–2205.
[49] Biwei Huang, Kun Zhang, Mingming Gong, Clark Glymour, Causal discovery from multiple data sets with non-identical variable sets, in: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 2020, pp. 10153–10161.
[50] R.W. Robinson, Counting Unlabeled Acyclic Digraphs, Springer, Berlin, Heidelberg, 1977, pp. 28–43.