Federated Causal Discovery in Medicine: Trends, Opportunities, and Challenges

Niccolò Rocchi[1,2], Marco Scutari[3], Alessio Zanga[1], Radha Nagarajan[4], Fabio Stella[1]

[1] University of Milano-Bicocca, Italy.
[2] Fondazione IRCCS Istituto Nazionale dei Tumori, Italy.
[3] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Switzerland.
[4] Rady Children's Health, Orange, CA, United States.

Abstract

Exponential growth and continued digitization have accelerated the adoption of data-driven and evidence-based approaches in medicine. This includes deciphering associations, including potential causal associations, from multivariate observational biomedical data under certain implicit assumptions. Such an evidence-based approach marks the shift from classical hypothesis testing to discovery and hypothesis generation. Widespread adoption of common data models has especially accelerated collaborative approaches in medicine while transitioning from centralized to federated architectures that facilitate discovery without the explicit sharing of sensitive medical data. This perspective provides an overview of causal inference from observational data with a focus on federated causal discovery. More importantly, it outlines the trends, opportunities, and challenges of federated causal discovery in medicine. While medicine has traditionally relied on the hierarchy of evidence generated from the evidence pyramid, the ability of federated causal discovery to facilitate evidence generation collaboratively from heterogeneous sources is expected to enhance the generalizability and transportability of findings while addressing sample size considerations—a critical aspect for its successful and widespread adoption in medicine.

1 Introduction

During the last decade, an exponential growth in *multivariate* and *multimodal* biomedical data has taken place due to several factors, including the widespread adoption of digital technology and continued digitization efforts from a multitude of source systems and high-throughput assays.[1] While multivariate data typically consists of *relational* and *structured* data that can be represented in a tabular format (e.g., numerical and categorical features), multimodal data includes *non-relational* and *unstructured* data (e.g., texts and images). Multivariate data is typically stored in *data warehouses* for querying and downstream analytics. *Centralized* and *federated* warehouse architectures have been proposed for the same.[2] A brief description of these architectures, along with their pros and cons, is shown in Table 1.

There is an increasing interest in harnessing information from multivariate and multimodal observational data to improve patient outcomes in an evidence-based manner. This includes modeling associations, including causal associations from multivariate observational biomedical data, complementing traditional evidence gener-

Table 1: Advantages and disadvantages of centralized and federated architectures.

	Pros	Cons
Centralized	Centralized governance with representation from the participating institutions.	A centralized resource is a single point of failure.
	Single storage and compute infrastructure for querying and analytics.	Centralizing high-throughput data may have prohibitive bandwidth and storage costs.
	Compute architecture is straightforward compared to federated ones.	Varying compliance, lack of trust, and regulatory regimes across institutions may limit data access and permitted analyses.
Federated	Architectures are more robust to failures because of distribution.	Architectures are more complex and difficult to manage.
	No explicit data sharing, enhanced trust across the participating institutions, and reduced patient consent burden.	Data drift and uneven data distributions across institutions can impact training and the generalizability of models.
	More control across the participating institutions. Data is compliant with local governance.	Execution of queries and analytics can have relatively higher latencies in federated architectures.

ation.[3] Such a data-driven approach, combined with domain knowledge, has the potential to validate established associations while discovering novel ones for critical assessment, thus generating new hypotheses and research questions. It also has the potential to provide novel system-level insights, a precursor to developing targeted interventions, tailored treatment regimens, and disease management strategies.[4] More importantly, it can complement the hierarchy of evidence from the *evidence pyramid*, including systematic reviews and *randomized controlled trials* (RCT) used widely in medicine.[3]

Networks have proven to be especially useful abstractions in this regard,[5] with *nodes* representing the variables of interest, and *edges* their associations. These networks can be modeled across multiple scales (e.g., molecular, clinical, or demographic data) of varying granularity and resolution,[6] as well as from cross-sectional and longitudinal profiles. While longitudinal profiles capture the explicit temporal evolution of multivariate processes, they are usually challenging to generate from economic and stationarity standpoints. The latter demands controlling a number of factors so that the statistical properties are preserved as a function of time.[7] Not surprisingly, cross-sectional profiles that interrogate the multivariate process in a chosen window of time, in conjunction with replicate measurements, have been prevalent. While earlier attempts focused on modeling these networks by estimating the pairwise dependencies between the variables (e.g., relevance networks[8]), there is increasing evidence that dependence between a pair of variables need not necessarily be direct, warranting the inclusion of conditional dependencies. Techniques such as *causal Bayesian networks* (CBNs)[9] have proven particularly helpful in this regard, with broad application in medicine and healthcare. These include deciphering associations from sequencing,[10] molecular,[11] epidemiological[12] and electronic health records[13] with a focus on oncology,[14] neurology,[15] rare[16] and infective diseases,[17] among others. CBNs have also been used success-

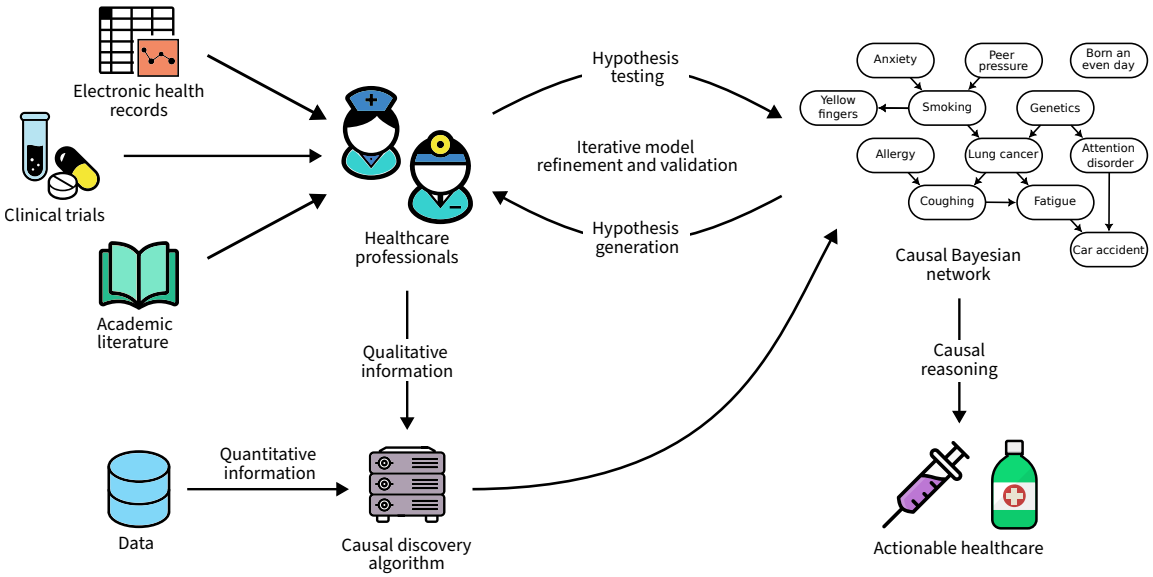

Figure 1: The causal discovery and reasoning pipeline.

fully to model relationships from diverse data sets, including multi-center,[18] temporal,[19] state-space,[20] partially observed,[21] and multimodal data[22] under the broad theme of *causal discovery* (CD).[23]

This perspective examines applications of *federated causal discovery* (FCD) in medicine and healthcare. Specifically, it introduces the "*What*", "*Why*", and "*How*" of FCD, offering a comprehensive and critical overview of current methodologies along with trends, opportunities and challenges. The necessary background and foundations on causal representation and discovery are presented in Section 2. A taxonomy of federated learning algorithms is provided in Section 3. Formal definitions and popular algorithms for FCD are discussed in Section 4. Finally, Section 5 presents current opportunities (5.1) and challenges (5.2) along with FCD case studies with a focus on medicine.

2 Background on Causality

Identifying cause-and-effect relationships is a fundamental step towards clinical reasoning and developing tailored strategies that align with broader precision health initiatives.[24] This section provides the rationale and foundations of causal modelling, along with learning techniques, which are also portrayed in Figure 1.

2.1 Causal Machine Learning

Pearl formalized the distinction between probabilistic and causal reasoning through the *ladder of causation*.[9] The ladder consists of three rungs of increasing complexity and relevance.

1. *Association*: purely about statistical relationships, thus conclusions are based on the patterns and associations in the given observational data. Example: association between changes in biomarker profiles and disease outcomes (e.g., 5-year survival rate in cancer).

2. *Intervention*: studying whether and how in-

terventions impact the system. Example: changes in clinical outcomes in response to a treatment regimen (e.g., cancer recurrence).

3. *Counterfactuals*: Analyzing hypothetical scenarios to determine what would have happened under different conditions than those we actually observed. Example: "Would the survival rate change significantly with a different treatment modality?"

Leveraging observational data and expert knowledge to answer research questions at rungs 2 and 3 falls under the broad theme of *causal inference* or *causal Machine Learning* (CML).[9] CML aims to discover causal mechanisms to predict which *effects* will be observed under interventions, thus allowing us to answer "What if I make a given decision?" questions (rung 2) and "Why or what if I had acted differently?" questions (rung 3). Thus, it transcends classical machine learning, which relies on probabilistic associations and can answer only "What is?" questions (rung 1).[25]

The following section introduces a pragmatic and data-driven approach to causation.

2.2 Causal Bayesian Networks

Causal Bayesian Networks (CBNs) provide a pragmatic and rigorous representation of causality, thereby enabling efficient knowledge representation and reasoning at all three rungs. While inferring the cause and effect solely based on *observational* data can be challenging because of *symmetric* relationships, *interventional* data such as those from clinical interventions (e.g., treatment selection)[26] or sample selection (e.g., randomization and inclusion criteria)[27] can provide insight into potential *asymmetry*, hence the direction of the cause-and-effect relationship.[28] A common assumption is that of *causal sufficiency*[9] which ensures that all the common causes (called *confounders*) are explicitly measured. CBNs also provide a natural representation of confounders between the treatment and the outcome, and can be used to test those assumptions. In RCTs, for instance, the treatment is randomized against a series of factors that might influence both the treatment assignment and the outcome, such as patients' covariates. As a result, we know that causal effects can only flow from the randomized variables, not the other way around.

A CBN is represented by a pair $(\mathcal{G}, \theta)$, where $\mathcal{G}$ is called *causal graph* (CG)[9] and θ is a set of parameters associated with $\mathcal{G}$. Each node i in $\mathcal{G}$ corresponds to a random variable X_i, and an edge $X_i \to X_j$ entails that X_i is a *direct cause* of X_j, with changes in X_i directly influencing changes in X_j. The nodes pointing towards X_j are called the *parent set* of X_j and denoted by Π_j. The parameters θ are associated with the joint probability distribution $P(\mathbf{X})$ over all nodes $\mathbf{X} = \{X_i, \ldots, X_n\}$. Each X_i is stochastically independent of its non-descendants given its parents.[29] Hence, each node in $\mathcal{G}$ is associated a *local* conditional distribution $P(X_i \mid \Pi_i)$ with parameters θ_i, $\bigcup_i \theta_i = \theta$. The CG $\mathcal{G}$ is often assumed to be a directed and acyclic graph (DAG). This is not a practical restriction: duplicating nodes over different time points and locations makes it trivial to represent both as they unfold in time and space.[30]

A CBN may be equivalently represented as a *structural causal model* (SCM) mapping the functional relationships between variables. Each $X_i \in \mathbf{X}$'s stochasticity is then expressed by a separate exogenous variable $U_i \in \mathbf{U}$.[9] The joint distribution $P(\mathbf{U})$ induces, recursively, the joint $P(\mathbf{X})$ over the measured variables; thus, the set of CBN parameters can be easily derived from $P(\mathbf{U})$ and the functional dependencies.

Definition 2.1 (Structural causal model)**.** An SCM $\mathcal{M}$ consists of a 4-tuple $(\mathbf{U}, \mathbf{X}, \mathcal{F}, P(\mathbf{U}))$ where:

- $\mathbf{U}$ represents a set of exogenous variables determined by factors outside the model.
- $\mathbf{X}$ represents a set $\{X_1, \ldots, X_n\}$ of endoge-

nous variables, determined by other variables in the model, that is, those in $\mathbf{U} \cup \mathbf{X}$.

- $\mathcal{F}$ represents a set of functions $\{f_1, \ldots, f_n\}$ such that each f_i is a mapping from the domain of $U_i \cup \Pi_i$ to X_i, where $U_i \subseteq \mathbf{U}$, $\Pi_i \subseteq \mathbf{X} \setminus \{X_i\}$.

For $i = 1, \ldots, n$, X_i is determined by the structural assignment:

$$X_i \coloneqq f_i(\Pi_i, U_i)\,.$$

This alternative representation is crucial in defining *counterfactuals*, but does not provide any advantage over the standard CBN definition in CD.[9]

Overall, our ability to perform causal inference is founded on knowing (parts of) the underlying cause-and-effect relationships. These may derive from assumptions, expert opinion, literature research or data and are encoded into the CG. The CG formalizes testable implications about conditional (in)dependences between variables: it is not only a descriptive map or a tool for causal inference. As we will discuss in the following, it provides the language and machinery for articulating assumptions, generating and testing hypotheses, controlling bias during inferences, creating synthetic data for exploratory purposes, performing *in silico* experiments, conducting sensitivity analyses to unmeasured confounding, and even planning scenarios for clinical trials.

2.3 Causal Discovery

Causal discovery (CD) consists in learning a CBN $(\mathcal{G}, \theta)$ from the given data while accommodating expert knowledge and established clinical evidence as priors.[31]. Two broad areas under CD include *structure learning* and *parameter learning*, which enable causal *inference* as a downstream task. While structure learning involves determining the DAG $\mathcal{G}$ that best represents associations in the given multivariate biomedical observational data, parameter learning estimates the marginal conditional probability distributions given the structure $\mathcal{G}$. Inference corresponds to posing queries to the resulting CBN $(\mathcal{G}, \theta)$. Under rare circumstances, it might be possible to construct $\mathcal{G}$ by leveraging expert knowledge and an extensive literature search.[32] However, such models have inherent limitations. For instance, *(a)* the retrieved knowledge may be generic and not specific to system under investigation (e.g. molecular signaling mechanism of a tumor subtype); *(b)* the retrieved knowledge may be from diverse hetergeneous sources challenging its integration (e.g. cell lines, tissues, species);*(c)* the retrieved knowledge may not be representative of the population or demographics of interest (e.g. variations in social determinants); *(d)* the system under investigationg may be novel with minimal precedence or prior knowledge (e.g. COVID-19). These challenges implicitly demand modeling the CBN $\mathcal{G}$ and its parameters θ in a data-driven $(\mathcal{D})$ and evidence-based manner.

To this end, CD is posed as a Bayesian optimization problem comprising structure and parameter learning:

$$\underbrace{P(\mathcal{G}, \theta \mid \mathcal{D})}_{\substack{\text{CBN} \\ \text{learning}}} = \underbrace{P(\mathcal{G} \mid \mathcal{D})}_{\substack{\text{Structure} \\ \text{learning}}} \cdot \underbrace{P(\theta \mid \mathcal{G}, \mathcal{D})}_{\substack{\text{Parameter} \\ \text{learning}}}.$$

As mentioned in Section 2.2, the DAG $\mathcal{G}$ may not be uniquely identifiable when relying solely on observational data, resulting in probabilistically indistinguishable structures, also referred to as the *Markov equivalence class* of $\mathcal{G}$.[29] Examples include network *motifs* or prevalent structures such as *chains* $X_i \to X_j \to X_k$, $X_i \leftarrow X_j \leftarrow X_k$, and *fork* $X_i \leftarrow X_j \to X_k$. The corresponding Markov equivalence class is given by the undirected graph $X_i - X_j - X_k$ since chains and forks are probabilistically indistinguishable. Therefore, a CBN is represented as *completed partially DAG* (CPDAG), typically comprising directed as well as undirected edges. An edge in a CPDAG is directed if and only if it is directed across all DAGs in the Markov equivalence class. However, interventional data can help identify the direction of undirected edges in the CPDAG.[28] Several factors, such as the struc-

ture learning algorithm, sample size, dimensionality, and distributional assumptions of the multivariate data, can also impact identifying the direction of the edges.[31]

Structure learning is a hard computational problem because the number of possible DAG structures increases super-exponentially with dimensionality.[33] Algorithms to tackle it broadly comprise *constraint-based*, *score-based*, and *hybrid approaches*. Score-based approaches that identify the DAG that best represents the given data using a search criterion in conjunction with a scoring function, such as LiNGAM and NOTEARS, have been used successfully to determine the direction of the edges. LiNGAM assumes that the distribution of exogenous variables is such that different edge directions imply different likelihood values, allowing for disambiguation. Differentiable algorithms like NOTEARS leverage scale differences between the exogenous variables and add penalty terms to the likelihood to ensure $\mathcal{G}$ is acyclic. In contrast, constraint-based approaches recover the optimal DAG using tests for conditional independence and are implicitly limited in determining edge directions by the symmetry of the relationships. Hybrid approaches adopt a combination of constraint-based and score-based algorithms in identifying the optimal DAG. The performance of these approaches and their computational complexity vary considerably.[34] As noted earlier, once $\mathcal{G}$ is determined, the associated parameters can be learned from data and domain knowledge using *maximum likelihood* (MLE) and *Bayesian posterior* estimators.[29]

While expert knowledge may not be enough to construct $\mathcal{G}$ by itself, it may assist in constraining the search space of candidate DAGs. These constraints in turn can be either *soft* or *hard*.[35] Soft constraints help guide the structure learning algorithm; *hard* constraints explicitly eliminate specific structures, minimizing the search space of the structure learning process.

3 Federated Learning

Federated Learning (FL) trains a model locally via global updates without explicit sharing of *protected health information* across healthcare organizations, thereby complying with data protection regulations.[36] Recent studies show how FL improves breast density classification models (accuracy up by 6%, generalizability up by 46%),[37] COVID-19 outcome prediction at both 24h and 72h (up 16% and 38%)[38] and rare tumour segmentation (up by 23–33% and 15%)[39] compared to single-organization analyses. Early-stage applications that build predictive models from electronic health records[40] have also confirmed that there is no practical performance degradation compared to a centralized analysis. Xu *et al.*[41] provides an excellent introduction to FL in healthcare.

FL architectures typically follow a *client-server* design where models are trained locally in a decentralised manner from data at individual clients and updated globally.[42] A taxonomy of FL along its key aspects (*data partitioning*, *architecture*, *algorithms*) is shown in Figure 2. *Data partitioning* describes how data are split across clients (see Figure 4). The *architecture* defines the roles of clients and servers and how they exchange information. Finally, the FL *algorithm* defines how models are learned and what privacy-preserving techniques are adopted.

4 Federated Causal Discovery

Federated causal discovery (FCD) enables CD from multiple data sources without explicit data sharing. Since FCD is a subset of FL, FCD algorithms follow the FL taxonomy in Figure 2. In particular:

- FCD is typically implemented on horizontally partitioned data, as it is challenging to compute sufficient statistics involving variables that are never observed jointly across clients. Typically, it cannot identify the true

Trait	Type	Description
Partitioning	Horizontal	Full overlap in variables, no overlap in data points.
	Vertical	No overlap in variables, full overlap in data points.
	Hybrid	Partial overlap in variables, partial overlap in data points.
Learning	Offline	Model is learnt once from the available data.
	Online	Model is updated whenever new data are available.
Task	Single-task	Learning one global model.
	Multi-task	Learning different models for different clients.
Aggregation	Synchronous	Each client contributes to each aggregation.
	Asynchronous	Some clients may not contribute to each aggregation.
Sharing	Model	The entire local models are shared with the server.
	Knowledge	Model's intermediate information is shared.
	Synthetic data	Local synthetic data are shared instead of raw data.
Scale	Cross-silo	Few clients, high computational power, and availability.
	Cross-device	Many clients, limited computational power, and availability.
Topology	Star schema	One server is in charge of aggregation.
	Distributed	Multiple aggregators are allowed.

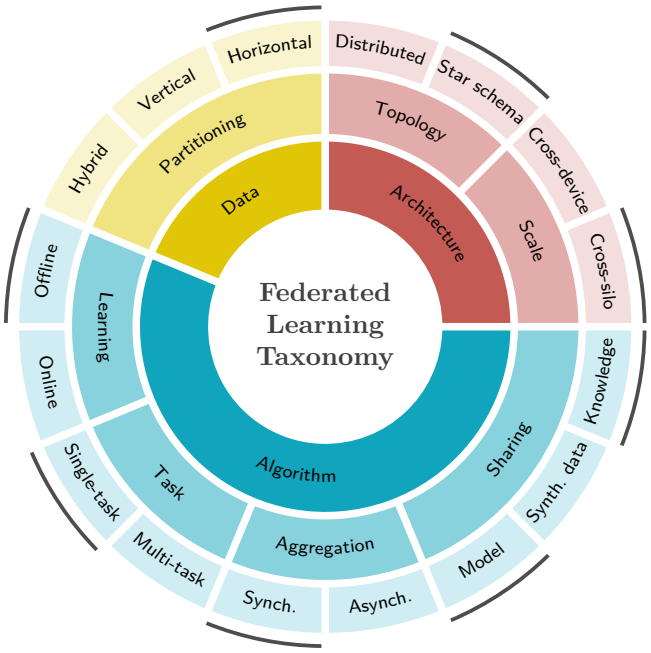

Figure 2: FL taxonomy based on different traits. Yellow represents data partitioning schema, blue represents the algorithmic structure and task, and red represents the client-server architecture. Gray small tabs identify the framework of FCD within FL; exceptions are marked in Figure 5.

CG because multiple DAGs may equally explain the data at hand.[43]

- FCD algorithms follow an offline learning procedure, as online CD is typically challenging even in single-source settings. This is not a significant limitation, as clinical data are more often collected retrospectively than prospectively.
- There are instances of multi-task FCD in the literature (see Figure 5). They typically assume a shared CG (Section 4.1) to be realistic. How to account for distribution shifts among clients is an open research question (Section 5.2).
- Synchronous aggregation is commonly carried out, although asynchronous learning may be practicable in principle.
- Most privacy-preserving techniques in FCD involve sharing scores or statistics, which are sometimes encrypted to protect the data. The extent to which original data can actually be traced back from score or statistics remains an unexplored area (Section 5.2).
- CD requires significant computational resources and time for big data, which favours the cross-silo architecture. There are no examples of (fully) distributed FCD, but it could be a solution to develop personalized models for multi-task FL.

FCD often makes additional assumptions, which are discussed in the following section.

4.1 Assumptions

In the most general setting, a source-dependent SCM is assumed to be responsible for generating each local data set $\mathcal{D}^1, \ldots, \mathcal{D}^k$. Specifically, let $\mathcal{M}^1, \ldots, \mathcal{M}^K$ be a collection of SCMs underlying the K data sources where $\mathcal{M}^k$ is given by:

$$\mathcal{M}^k = (\mathbf{U}^k, \mathbf{X}^k, \mathcal{F}^k, P^k(\mathbf{U}^k)),$$
$$X_i^k \coloneqq f_i^k(\Pi_i^k, U_i^k).$$

Without loss of generality,[1] we set $\mathbf{U}^k = \mathbf{U}, \forall k$. Because FCD research and applications focus overwhelmingly on the horizontally-partitioned data, we set $\mathbf{X}^k = \mathbf{X}, \forall k$.

Figure 3 lists a set of common assumptions regarding both the model and the data. They are relevant beyond FL, but we discuss them in the context of FCD.

Assumption 1 (Shared causal graph)**.** Cause-and-effect relationships are invariant across clients, that is $\Pi_i^k = \Pi_i, \forall i, k$. Hence, all CGs coincide: $\mathcal{G}^k = \mathcal{G}, \forall k$.

Despite the inherent noisiness and heterogeneity of the $\mathcal{D}^k$, relationships are often robust and preserved after controlling for potential confounders. Realistically, cause-and-effect relationships are likely invariant in many real-world settings, while parameters may vary.[44] For instance, disease etiology and drug mechanisms of action are similar regardless of where a physician collects the measurements, but drug effectiveness may vary in different subpopulations. Exceptions exist, for instance, in FMRI data.[45]

Under Assumption 1, FCD consists of obtaining a single $\mathcal{G}$ from $\mathcal{D}^1, \ldots, \mathcal{D}^k$. Otherwise, we talk about *multi-task* FCD as the task of collaboratively finding $\mathcal{G}^k, \forall k$.

Assumption 2 (Observational data)**.** Data were collected during routine work without deliberate manipulations of the variables. In other words, the variables are not *intervened* on.[9]

For a client k and variable X_i, a *perfect* intervention sets a structural assignment for X_i to $X_i \coloneqq x_i$ for a certain x_i. On the other hand, an *imperfect* intervention modifies either $P^k(\mathbf{U})$ or $f_i^k(\cdot)$ in $\mathcal{M}^k$ without fixing them to a single, deterministic value.

Assumption 3 (Pseudo-causal sufficiency)**.** Any variable causing two or more variables in $\mathbf{X}^k$ is included in $\mathbf{X}^k$, $\forall k$.

[1] In fact, we may set $\mathbf{U} = \bigcup_k \mathbf{U}^k$, $P(\mathbf{U})$ its joint distribution, and $P^k(\mathbf{U}) = \sum_{i \neq k} P(\mathbf{U}^k, \mathbf{U}^i)$, $\forall k$.

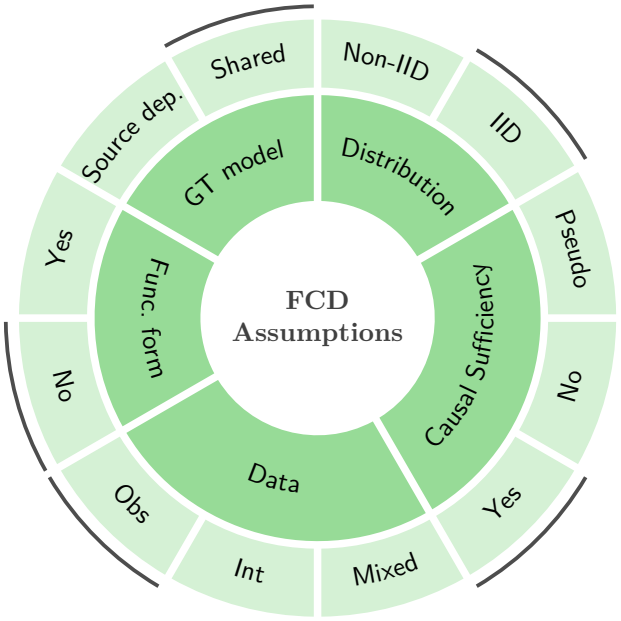

Figure 3: Overview of possible assumptions in FCD. Gray small tabs identify the most adopted ones; exceptions are marked in Figure 5. "GT model" stands for "ground-truth model", namely the CG; "Func. form" means "functional form" and identifies whether the approach needs to assume an SCM form.

Assumption 3 coincides with the standard *causal sufficiency* assumption holding locally at each data source.[31] However, multi-source causal sufficiency requires stronger guarantees, as one variable may cause two other variables measured at different sites.

Assumption 4 (Causal sufficiency)**.** Any variable causing two or more variables in $\bigcup_k \mathbf{X}^k$ is included in $\mathbf{X}^k$, $\forall k$.

The hybrid and vertical data partitioning settings (Figure 2) conflict with Assumption 3 and Assumption 4, because not all variables are necessarily measured in all sources. Note that causal sufficiency implies pseudo-causal sufficiency, but not vice versa.

Assumption 5 (IID data)**.** Data are *independent and identically distributed* (IID) within and across the clients, hence $P^k(\mathbf{U}) = P(\mathbf{U}), \forall k$.

Non-IID data commonly arise from *distribution drifts*, or *shifts*, between different populations: $P^i(\mathbf{X}) \neq P^j(\mathbf{X})$, for some $i \neq j$. If Assumptions 2 and 3 hold, we can account for them using a set of *context variables* as illustrated in Figure 4.[46] Consider the CBN $(\mathcal{G}^*, \theta^*)$, where (i) $\mathcal{G}^*$ is the enhanced version of $\mathcal{G}$ obtained by adding a set of variables $\mathbf{C}$ and an edge from $C \in \mathbf{C}$ to $X \in \mathbf{X}$ whenever $P^i(X) \neq P^j(X)$, $i \neq j$; (ii) θ^* parametrizes the joint distribution $P^*(\mathbf{X}, \mathbf{C})$. Each assignment $\mathbf{C} = \mathbf{c}$ defines specific environmental, individual, or measurement conditions.[46] The local data set $\mathcal{D}^k$ is collected under the environment $\mathbf{C} = \mathbf{c}^k$, and the underlying CBN is $(\mathcal{G}, \theta^k)$. Here, θ^k parametrizes the joint distribution $P^k(\mathbf{X}) = P^*(\mathbf{X}, \mathbf{C} \mid \mathbf{C} = \mathbf{c}^k)$.

Hypothetically, $\mathcal{G}^*$ can be learned whenever the context $\mathbf{C}$ is recorded at each client. However, those variables are usually unknown, so at least one indicator variable C may be used as a

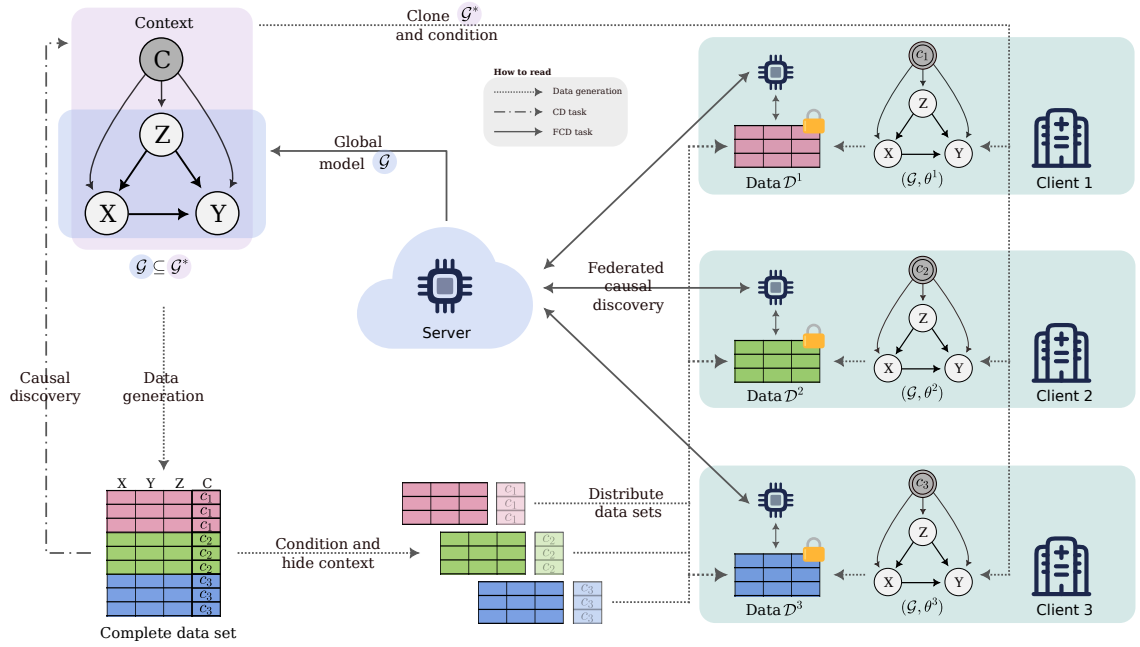

Figure 4: Overview of data generation in the federated setting, along with the CD and FCD tasks. Note that **C** may represent multiple variables.

proxy, with $C = k$ indicating the k-th client.[2] Note that this approach requires the server to know which client each local update originates from, which is not always the case in highly private settings. Figure 2 outlines how FCD relates to the FL taxonomy.

4.2 Algorithms

Causal discovery using multiple data sets is well-explored in the literature.[46] Here, we focus on those algorithms that can be strictly classified as FL and, as such, follow the characterization provided in Figure 2.

Basic approaches to FCD involve performing CD locally and aggregating the learned graphs through edge voting, union, or intersection.[47] These approaches prove effective when local sample sizes in individual clients are large enough and Assumption 5 assumption holds. However, smaller sample sizes at the client level (Figure 4) may not be representative of the underlying population, leading to statistically significant differences in local distributions. Moreover, Assumption 5 may not be satisfied in many scenarios.

Figure 5 presents the FCD algorithms linked to the assumptions and settings discussed in Section 4.1. Their classification is based on how the server explores the DAG space. In the following, we summarize and provide details about some relevant instances.

Constraint-based The FedC2SL algorithm[66] extends PC and the Fast Causal Inference (FCI) algorithms[23] to the federated setting. FCI works in contexts where some variables may be unobserved, while PC assumes causal sufficiency (Assumption 3). The server iteratively constrains the DAG space using *federated conditional independence tests* that distribute the computation of sufficient statistics and aggregate them securely. The FedCDH algorithm[65] does the same, but

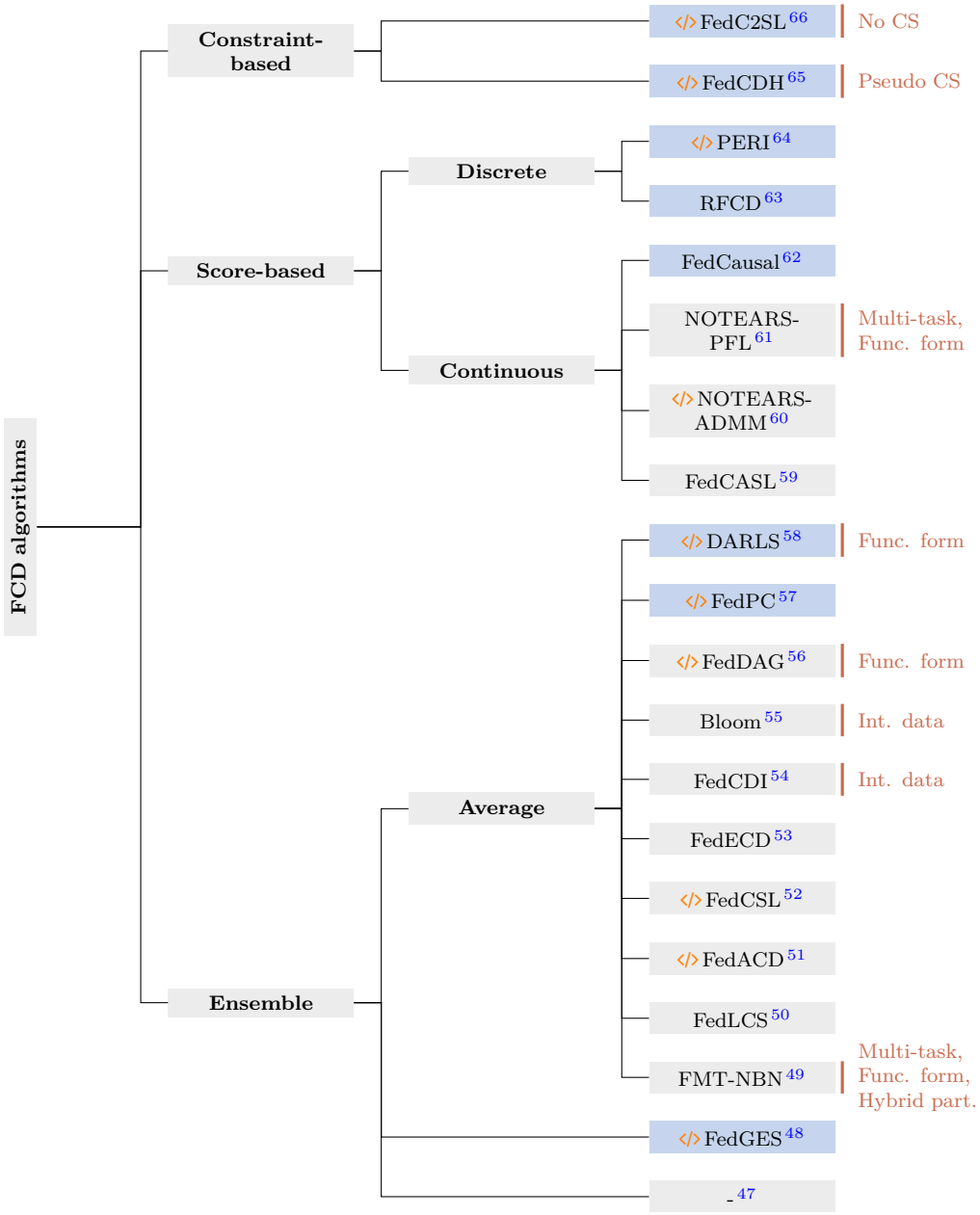

Figure 5: Overview of FCD algorithms classified on the server strategy of DAG space exploration. In blue, those ensuring the acyclicity of the global CG. The </> symbol marks the available implementations. In red, those approaches that do not follow strictly the FCD characterization as in Figures 2 and 3. "CS" stands for "causal sufficiency"; "Func." for functional; "Int." for "interventional"; "Part." for "partitioning".

clients release sufficient statistics to the server only once.

Score-based The PERI algorithm[64] builds on the Greedy Equivalent Search (GES) algorithm.[67] It minimizes the worst-case *regret* of a function that measures how far a DAG is from the DAG that best fits the data across all clients. PERI ensures privacy and convergence guarantees. The FedCausal algorithm[62] extends the gradient-based NOTEARS-ADMM algorithm[60] by providing a global optimization function.

Ensemble Model ensemble techniques consist of finding a *consensus* DAG from a given collection.[68] The FedGES algorithm[48] uses model ensemble techniques to combine locally learned DAGs at the server level. The procedure is iterative: the global DAG is fused with the client's DAG, becoming the starting point for the subsequent local optimization. Note that naive aggregation of edges, such as union or averaging, does not ensure acyclicity, while edges lose their causal and statistical meanings. The FedLCS algorithm[50] is a local FL extension of the PC algorithm that finds only the edges incident on a single variable, rather than the entire CG.

5 Discussion

As noted earlier, increasing digitization in medicine has led to an exponential growth in multivariate and multimodal observational data. Deciphering novel patterns and associations from these data sets has the potential to discover novel associations while validating what is known. However, non-causal machine learning approaches focus exclusively on predictive performance and are not designed for clinical reasoning.[24]

FCD has the potential to decipher causal patterns from these observational data sets across, without explicit data sharing, overcoming sample size constraints. CBNs obtained through FCD excel at *causal reasoning* (Section 2.1) and thus produce valid conclusions about treatment effect, exposure effect, risk factor modification, and ultimately better patient outcomes. They reduce the risk of mistaking correlation patterns for causation and provide a principled approach to investigating scenarios when RCTs are unfeasible, thereby bridging the gap between observational and experimental evidence.

Still, they require large, well-balanced samples to produce transferable models that generalise across clinical and population cohorts and between different clinical cohorts. Crucially, the close relationship between several key concepts in causal inference and trial design (for instance, backdoor adjustment vs randomization, collider bias and stratification, path analysis and mediation) makes CBNs a powerful tool for boosting systematic reviews and RCTs.

In the remainder of this section, we will focus on some of the opportunities and challenges of FCD in this context, as well as less-explored research areas.

5.1 Opportunities

FCD can perform causal inference for treatment effects, pooling information from external control arms with real-world control patients from different institutions. Building on the effectiveness of FL[37,39] and CD[24], it can leverage RWE to improve drug discovery and patient outcomes by improving trial effectiveness, fairness, data integration and privacy.[41]

5.1.1 Clinical Translation

The need for cross-institutional cooperation to foster medical research had been apparent for years. Experts' collaboration and data sharing enhance RWE, thereby strengthening the generalizability of analyses, reducing local biases, and promoting fairness.[69] Popular architectures enabling multi-institutional storage and analytics of medical data include both centralized and federated architectures, as described in Section 3.

Centralized architectures have been adopted to accelerate precision medicine research efforts,

such as the *All of Us* initiative funded by the National Institutes of Health (NIH).[70] All of Us aims to provide equitable access to diverse data sets from participating institutions in a secure, centralized cloud-based environment, with centralized governance and a paid model to support analytics. Similarly, the European ELIXIR infrastructure supports the coordination and development of life sciences research, including a network of cross-domain experts, and promotes best practices for data analysis.[71] The *Health Information Exchange* (HIE) also aims to improve care coordination while supporting surveillance.[72] In addition, the *Global Alliance for Genomics and Health* (GA4GH)[73] aims to establish standards for sharing genomic and health data while ensuring privacy, security, and adherence to ethical values.

Unlike centralized architectures, federated architectures have supported federated querying of de-identified medical data with the potential to accelerate clinical trials. An example is the *Shared Health Research Information Network* (SHRINE), funded by the *National Center for Advancing Translational Science* (NCATS).[74] SHRINE has been adopted by *Patient Centered Outcomes Research* (PCORI), enabling federated querying of healthcare data from PCORnet members. While these established projects targeted cohort discovery, new developments are focusing on model learning through FL instead.

FL has gained traction in health research, with studies exploring different data types and applications.[69] To this end, several initiatives promoted the creation of distributed infrastructures to host ready-to-use, private data. Among those, the *IDEA4RC* project focuses on creating a decentralized ecosystem of rare cancers data in Europe, compliant with the standardization of common data models;[75] *FeederNet* is a South Korean initiative supporting the development of a biohealth data ecosystem for federated analyses.[76] In principle, these platforms could be leveraged for large-scale FCD.

However, limited applied research has been published on FCD to date. Notably, Zanga *et al.*[77] evaluated their FCD algorithm on decentralized data sets related to endometrial cancer. Their method is suited to incomplete data, especially when the missing mechanism differs across centres. Chen *et al.*[78] modeled gene expression by leveraging a time-dependent version of the CG. Zhang *et al.*[79] proposed a method to aggregate outputs from different CD algorithms and applied it to the treatment of acute kidney injury. Other applied works[80,81] focused on *federated causal inference* by assuming the CGs are known instead of learning them. Furthermore, no existing distributed medical data set currently provides a ground-truth DAG for FCD benchmarking and experimentation.

Cross-institutional collaborations may aspire to learn more robust CBNs through FCD than individual institutions; the reference population may be clearly specified and taken into consideration (see Section 5.2.2).[82] Involving their respective experts in the CD process increases the trust in the learned CBN and fosters its adoption in clinical practice. However, it may present some challenges in harmonizing heterogeneous and possibly conflicting domain knowledge.[83]

5.1.2 Efficiencies in RCTs

RCTs are limited in their ability to translate to real-world scenarios, such as cancer care, due to their inclusion and exclusion criteria.[84] Non-causal models have shown mixed performance in providing information comparable to that from RCTs from *real-world evidence* (RWE) available in electronic health records.[85] CBNs, on the other hand, can effectively harness RWE to better understand the real-world patient experience and outcomes thanks to their ability to learn digital twins of such settings.[13] CBNs can identify relevant hypotheses to test and subpopulations with different treatment responses; provide prior estimates of effect sizes to identify relevant biomarkers for drug development[86] and perform power analysis for sample size determination; reduce the reuse of trial control arms, which limits transferability.[87] In addition, they could be used

more often to answer secondary analyses of existing RCT data.[88]

Furthermore, CBNs can inform the RCT design and assess its feasibility from observational RWE by emulating it in the "target trial" framework. Making sound design decisions about the length of follow-up, sample size, potential confounders, relevant subpopulations, minimization of lost-to-follow-up, and expected cost of treatment delivery is essential for trials to provide useful information at their conclusion.[89] At the same time, RCTs are increasingly expensive, reaching a median budget of $650,000, of which 27.4% is spent in the planning phase and 12.7% in the finalization phase.[90] Even so, about 70% of RCTs exceeded the budget by over 50%. Furthermore, 57% of RCTs had one or more substantial amendments, each costing $141,000 to $535,000.[91] Of these, 45% originated from protocol design flaws, inconsistencies in the protocol narrative, and infeasible eligibility criteria, which were "avoidable" and would have been flagged by a "target trial" CBN built from readily available RWE. Any additional planning efficiencies and reduced overruns resulting from more targeted designs are essential to advancing evidence-based healthcare in a cost-effective manner.

5.1.3 Fairness

Fairness is a well-documented issue in clinical trials. In addition to biases in historical control data,[92] they may also be biased due to limitations in the trial design,[93] self-selection among minorities,[94] physicians' implicit biases,[95] and other factors. FCD offers a comprehensive solution to this class of issues because of its unique combination of FL and causal inference.

FL can potentially provide better coverage of a target population by allowing multiple institutions to pool information without sharing patient-level data, which is one of the major concerns that pushes minorities to self-select themselves out of trial enrollment. Indirectly, it would also reduce the impact of physicians' implicit biases by giving them access to scrutinized, larger control arms.[96]

In addition to mitigating known sources of biases, better population coverage provides CD with the data it needs to construct a CBN that captures the characteristics of the trial and its patients. Such a CBN serves as the foundation for achieving *counterfactual fairness*[97], which builds on counterfactuals to examine differential outcomes as a function of legally protected attributes. This is the most rigorous framework for fairness assessment and remediation in the literature; notably, it goes beyond descriptive statistics and allows for disambiguating between different biases mediated by different causal pathways.[98]

5.1.4 Synthetic Data Generation

Synthetic data generation is becoming an increasingly attractive research area due to its significant potential in addressing many of the issues and limitations that arise when learning a model from scarce data.

In particular, synthetic data generation is receiving increasing attention in medicine, biology and healthcare[99]. Causal networks, being generative models, are an increasingly explored option for synthetic data generation in biology and healthcare[100,101]. Therefore, this renewed interest in causal networks offers the opportunity to study how a causal network, learned from multiple data sources, and thus suffering from many types of bias and other limitations, can or must be used to synthetically generate reliable data.

On the other hand, the FCD problem can also be tackled and studied in those cases where synthetic data is available to increase the sample size of different data sources. Indeed, the synthetic data generation landscape seems to pay a little to no attention to the relevant issue of disambiguating *random zeros* from *structural zeros*. Indeed, structural zeros occur when zero counts in a dataset may arise because a subject or group is fundamentally unable to have a non-zero value due to a restriction of the system being studied.

We think that studying, designing and developing FCD algorithms are worth attention because they offer a great opportunity for academia, companies and practitioners.

5.2 Challenges

5.2.1 Data Heterogeneity

Most clinical data are stored in heterogeneous data silos with different logical and physical structures, each one tailored to meet specific technical needs. Combining different databases may lead to inappropriate and biased results, even within the same institution.[102]

A software solution to data harmonization is given by *Beacon v2*[103], which provides a secure and flexible protocol for querying heterogeneous databases. However, Beacon does not support FL. Collaborative networks such as *Observational Health Data Sciences and Informatics* (OHDSI)[76] provide *common data models* (CDMs) to address data heterogeneity in medicine, thereby improving the generalizability of findings. *Common* refers to the data logical structure and shared vocabularies, fostering homogeneous semantics. A CDM enhances data owners' management, improves data users' interoperability while complying with privacy and security standards, and facilitates the development of standardized analytical tools. The *Observational Medical Outcomes Partnerships* (OMOP) CDM[104], developed within OHDSI, provides a longitudinal view of each patient. Its *oncology extension* further increases the information granularity to best support cancer research.

Translation into CDMs via *extract-transform-load* (ETL) processes poses several challenges. Only a subset of a center's raw data is typically mapped, and ineffective cooperation among centers will impact the final data homogeneity. Lack of ETL expertise can also undermine final data quality, leading to spurious cause-and-effect relationships in FCD. Both issues must be considered in FCD, and the ETL process must be well-documented for this purpose. Methodological research should also focus on developing FCD techniques for longitudinal and censored data,[105] as well as modalities such as medical imaging[106] and natural language processing.[41]

5.2.2 Distribution and Semantic Drifts

Variable distributions may differ between sources due to genetic population structure, heterogeneous environmental conditions, including unmeasured economic and social factors, and other covariates, such as age and sex.[107] The observational or interventional nature of various parts of the data is also a difficult assumption to test. Heterogeneous machine calibration, both across centres and over time, differing institutional policies and various inductive and deductive processes employed by physicians, may effectively amount to unobservable interventions. Note that interventions differ from distribution drift across populations, as the latter holds irrespective of whether an intervention has been performed.[108] Also, there may be *semantic drift* when variables have different semantics across data sources.

Modelling incomplete data poses similar challenges. The reason behind the presence of missing values is inherently causal; each data source in FCD may not only have varying degrees of missingness but also different missingness mechanisms. Expert prior knowledge and specialized algorithms are required to handle them.[77] Discarding incomplete observations or assuming an identical missingness mechanisms will bias both ETL and FCD itself.

5.2.3 Aggregation Bias

Learning CBNs from a single data source is a well-explored problem; the same is not true in FCD. Under Assumption 1, one may average all the local distributions, weighting them by the local sample size. However, this would not account for distribution drifts across sources. Even a single low-quality or highly biased data set can disproportionately bias the global CBN in the presence of few clients or class imbalance. Druzdzel and Díez[82] demonstrated how not combining knowledge from different sources or using

only data from the setting in which the CBN will be used is neither necessary nor sufficient to ensure model correctness. Hence, whether to adopt FL depends on the population(s) of interest, the research question, and the validity of assumptions. A sensible approach could be to perform a posteriori global model personalisation for individual clients, especially when the research question pertains to the local population. However, determining the value of each client's contribution to FCD is an open research question.

5.2.4 Privacy and Security

FCD algorithms must implement privacy and security by design to fulfil legal and ethical regulations, considering that institutions may be located in different jurisdictions.[109] Privacy involves safeguarding and controlling personal information, while security means protecting the system's integrity, availability, and confidentiality. The nature of attacks against privacy (e.g., *membership inference*) and security (e.g., *model poisoning*) depends on the threat model and the attacker's goal.[110,111]

Privacy and security should be addressed on multiple fronts.[41] Cybersecurity best practices will prevent the most common data leaks. FL is designed to avoid data sharing, thereby preserving privacy by design; most privacy attacks in the literature are only feasible under unrealistic assumptions.[112] Secure multiparty computation[66] (MPC, for encryption) and differential privacy[113] (DP, for privacy) can reach an acceptable trade-off between privacy, performance, and fairness.[110]

The scarcity of existing literature limits a comprehensive analysis of privacy and security in FCD. Membership inference attacks are not possible by sharing an unparameterized CG, but the presence or absence of specific edges might reveal sensitive information about certain subpopulations.[114] For this reason, Xu *et al.*[115] developed the *PrivPC* algorithm to perform CD under DP constraints. Murakonda *et al.*[116] provided a theoretical bound on the error and power of membership inference attacks on an exposed network. Zhang *et al.*[117] showed how to generate synthetic data from a CBN under DP. Rocchi *et al.*[118] used credal networks to mask the released model without degrading its utility, in contrast to DP-based protection techniques. Further research may shed light on practical privacy leaks in existing methods, clarifying what can and cannot be shared, to better enforce security standards while maintaining the scalability and effectiveness of FCD approaches.

5.2.5 Software Availability

The lack of comprehensive software solutions for FCD hinders its development and evaluation on real-world data. Some FCD algorithms have been implemented independently by the respective authors usinge different programming languages and file formats (see Figure 5). Moreover, most implementations require data to be in a tabular format and lack interfaces to learn from CDMs, forcing researchers to use ETL to preprocess CDMs into a tabular form. For instance, Schulz *et al.*[119] conducted a CD study using a single OMOP data set, first ensuring data compatibility with the employed algorithm. The OHDSI community provides the ATLAS and HARES tools for extracting a cohort from an OMOP CDM and populating a data table. At this point, the OHDSI's ARACHNE system may orchestrate federated epidemiological studies, while the *Vantage6* platform[120] may facilitate more complex FL analyses.

Finally, FCD may require enhanced communication and computational overhead than centralized efforts.[81] CD is known to be resource-intensive for big data, so most experimentation in the FCD literature is limited to simple CBNs. Future research should thoroughly investigate the computational bottlenecks of FCD and engineer reliable, efficient infrastructures.

References

[1] Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal Biomedical AI. *Nature Medicine* **28**, 1773–1784 (2022).

[2] Li, R., Romano, J. D., Chen, Y. & Moore, J. H. Centralized and Federated Models for the Analysis of Clinical Data. *Annual Review of Biomedical Data Science* **7**, 179–199 (2024).

[3] Berlin, J. A. & Golub, R. M. Meta-Analysis as Evidence: Building a Better Pyramid. *JAMA* **312**, 603–606 (2014).

[4] Abernethy, A. *et al.* The Promise of Digital Health: Then, Now, and the Future. *NAM Perspectives* **6**, 1–24 (2022).

[5] Kivela, M. *et al.* Multilayer Networks. *Journal of Complex Networks* **2**, 203–271 (2014).

[6] Zitnik, M. & Leskovec, J. Predicting Multicellular Function Through Multi-Layer Tissue Networks. *Bioinformatics* **33**, i190–i198 (2017).

[7] Chatfield, C. *The Analysis of Time Series: Theory and Practice* (Springer, 1975).

[8] Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R. & Kohane, I. S. Discovering Functional Relationships Between RNA Expression and Chemotherapeutic Susceptibility Using Relevance Networks. *Proceedings of the National Academy of Sciences* **97**, 12182–12186 (2000).

[9] Pearl, J. *Causal Inference in Statistics* (Wiley, 2021).

[10] Carapito, R. *et al.* Identification of Driver Genes for Critical Forms of Covid-19 in a Deeply Phenotyped Young Patient Cohort. *Science Translational Medicine* **14** (2022).

[11] Brown, A. A. *et al.* Genetic Analysis of Blood Molecular Phenotypes Reveals Common Properties in the Regulatory Networks Affecting Complex Traits. *Nature Communications* **14**, 5062 (2023).

[12] Petersen, A. H., Ekstrøm, C. T., Spirtes, P. & Osler, M. Constructing Causal Life-Course Models: Comparative Study of Data-Driven and Theory-Driven Approaches. *American Journal of Epidemiology* **192**, 1917–1927 (2023).

[13] Papanastasiou, G. *et al.* Causal Modeling in Large-Scale Data to Improve Identification of Adults at Risk for Combined and Common Variable Immunodeficiencies. *npj Digital Medicine* **8**, 361 (2025).

[14] Balordi, A. *et al.* On Counterfactual Explanations of Cardiovascular Risk in Adolescent and Young Adult Breast Cancer Survivors. *Journal of Medical Systems* **49**, 140 (2025).

[15] Delucchi, M. *et al.* Bayesian Network Analysis Reveals the Interplay of Intracranial Aneurysm Rupture Risk Factors. *Computers in Biology and Medicine* **147**, 105740 (2022).

[16] Lim, R. G. *et al.* Huntington Disease Oligodendrocyte Maturation Deficits Revealed by Single-Nucleus Rnaseq Are Rescued by Thiamine-Biotin Supplementation. *Nature Communications* **13**, 7791 (2022).

[17] Belyaeva, A. *et al.* Causal Network Models of SARS-CoV-2 Expression and Aging to Identify Candidates for Drug Repurposing. *Nature Communications* **12**, 1024 (2021).

[18] Zanga, A. *et al.* Causal Discovery with Missing Data in a Multicentric Clinical Study. In *Proceedings of the 21st International Conference on Artificial Intelligence in Medicine (AIME23)*, 40–44 (2023).

[19] Deshpande, A., Chu, L., Stewart, R. & Gitter, A. Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomics. *Cell Reports* **38**, 110333 (2022).

[20] Scutari, M., Kerob, D., Krutmann, J. & Salah, S. Causal Networks of Infodemiological Data: Modelling Dermatitis. In *Proceedings of the 23rd International Conference on Artificial Intelligence in Medicine (AIME25)*, 397–407 (2025).

[21] Mohammad-Taheri, S. *et al.* Do-Calculus Enables Estimation of Causal Effects in Partially Observed Biomolecular Pathways. *Bioinformatics* **38**, i350–i358 (2022).

[22] Sturma, N., Squires, C., Drton, M. & Uhler, C. Unpaired Multi-Domain Causal Representation Learning. In *Advances in Neural Information Processing Systems*, 34465–34492 (2023).

[23] Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search* (MIT Press, 2000).

[24] Glocker, B., Musolesi, M., Richens, J. & Uhler, C. Causality in Digital Medicine. *Nature Communications* **12**, 5471 (2021).

[25] Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2009), 2nd edn.

[26] Brown, B. C., Tokolyi, A., Morris, J. A., Lappalainen, T. & Knowles, D. A. Large-Scale Causal Discovery Using Interventional Data Sheds Light on Gene Network Structure in k562 Cells. *Nature Communications* **16** (2025).

[27] Colnet, B. *et al.* Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review. *Statistical Science* **39**, 165–191 (2024).

[28] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal Protein-Signaling Networks Derived From Multiparameter Single-Cell Data. *Science* **308**, 523–529 (2005).

[29] Koller, D. & Friedman, N. *Probabilistic Graphical Models* (MIT Press, 2010).

[30] Dean, T. & Kanazawa, K. A Model for Reasoning About Persistence and Causation. *Computational Intelligence* **5**, 142–150 (1989).

[31] Zanga, A., Ozkirimli, E. & Stella, F. A Survey on Causal Discovery: Theory and Practice. *International Journal of Approximate Reasoning* **151**, 101–129 (2022).

[32] Xiao-xuan, H., Hui, W. & Shuo, W. Using Expert's Knowledge to Build Bayesian Networks. In *2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)*, 220–223 (2007).

[33] Bouckaert, R. R. *Bayesian Belief Networks: From Construction to Inference*. Ph.D. thesis, Utrecht University, The Netherlands (1995).

[34] Scutari, M., Graafland, C. E. & Gutiérrez, J. M. Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms. *International Journal of Approximate Reasoning* **115**, 235–253 (2019).

[35] Constantinou, A. C., Guo, Z. & Kitson, N. K. The Impact of Prior Knowledge on Causal Structure Learning. *Knowledge Information Systems* **65**, 3385–3434 (2023).

[36] McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-Efficient Learning of Deep Networks From Decentralized Data. *Proceedings of Machine Learning Research* **54 (AISTATS)**, 1273–1282 (2017).

[37] Roth, H. R. *et al.* Federated Learning for Breast Density Classification: A Real-World Implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: 2nd MICCAI Workshop, DART 2020, and 1st MICCAI Workshop, DCL 2020*, 181–191 (2020).

[38] Dayan, I. *et al.* Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19. *Nature Medicine* **27**, 1735–1743 (2021).

[39] Pati, S. *et al.* Federated Learning Enables Big Data for Rare Cancer Boundary Detection. *Nature Communications* **13**, 7346 (2022).

[40] Brisimi, T. S. *et al.* Federated Learning of Predictive Models From Federated Electronic Health Records. *International Journal of Medical Informatics* **112**, 59–67 (2018).

[41] Xu, J. *et al.* Federated Learning for Healthcare Informatics. *Journal of Healthcare Informatics Research* **5**, 1–19 (2021).

[42] Zhang, C. *et al.* A Survey on Federated Learning. *Knowledge-Based Systems* **216**, 106775 (2021).

[43] Tillman, R. E. & Eberhardt, F. Learning Causal Structure From Multiple Datasets with Similar Variable Sets. *Behaviormetrika* **41**, 41–64 (2014).

[44] Schölkopf, B. *Causality for Machine Learning*, 765–804 (ACM, 2022).

[45] Ramsey, J. *et al.* Six Problems for Causal Inference From FMRI. *Neuroimage* **49**, 1545–1558 (2010).

[46] Mooij, J. M., Magliacane, S. & Claassen, T. Joint Causal Inference From Multiple Contexts. *Journal of Machine Learning Research* **21**, 1–108 (2020).

[47] Ajayi, O. T. & Cheng, Y. Decentralized Learning of Bayesian Networks From Private Data with Applications to Global Pandemic. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*, 961–962 (2023).

[48] Torrijos, P., Gámez, J. A. & Puerta, J. M. FedGES: A Federated Learning Approach for Bayesian Network Structure Learning. In *Discovery Science*, 83–98 (2025).

[49] Yang, X., Niu, B., Lan, T. & Zhang, C. Federated Multi-Task Bayesian Network Learning in the Presence of Overlapping and Distinct Variables. *IISE Transactions* **57**, 773–787 (2025).

[50] Yu, K., Rong, C., Wang, H., Cao, F. & Liang, J. Federated Local Causal Structure Learning. *Science China Information Sciences* **68** (2025).

[51] Guo, X. *et al.* Sample Quality Heterogeneity-Aware Federated Causal Discovery Through Adaptive Variable Space Selection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 4071–4079 (2024).

[52] Guo, X., Yu, K., Liu, L. & Li, J. FedCSL: A Scalable and Accurate Approach to Federated Causal Structure Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**, 12235–12243 (2024).

[53] Guo, X., Yi, L., Wu, X., Yu, K. & Wang, G. Enhancing Causal Discovery in Federated Settings with Limited Local Samples. In *International Workshop on Federated Foundation Models in Conjunction with NeurIPS 2024* (2024).

[54] Abyaneh, A. *et al. Federated Causal Discovery From Interventions* (2022). URL https://arxiv.org/abs/2211.03846.

[55] Qiu, C. & Yang, K. Interventional Causal Structure Discovery Over Graphical Models with Convergence and Optimality Guarantees. *IEEE Transactions on Network Science and Engineering* **12**, 156–172 (2025).

[56] Gao, E. *et al.* FedDAG: Federated DAG Structure Learning. *Transactions on Machine Learning Research* (2023).

[57] Huang, J., Guo, X., Yu, K., Cao, F. & Liang, J. Towards Privacy-Aware Causal Structure Learning in Federated Setting. *IEEE Transactions on Big Data* **9**, 1525–1535 (2023).

[58] Ye, Q., Amini, A. A. & Zhou, Q. Federated Learning of Generalized Linear Causal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**, 6623–6636 (2024).

[59] Liu, C. *et al.* Federated Causal Structure Learning with a Bi-Level Optimization Model. In *2024 IEEE 14th International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (Cyber)*, 708–713 (2024).

[60] Ng, I. & Zhang, K. Towards Federated Bayesian Network Structure Learning with Continuous Optimization. *Proceedings of Machine Learning Research* **151 (AISTATS)**, 8095–8111 (2022).

[61] Liu, S. *et al.* Federated Bayesian Network Learning From Multi-Site Data. *Journal of Biomedical Informatics* **163**, 104784 (2025).

[62] Yang, D. *et al.* Federated Causality Learning with Explainable Adaptive Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**, 16308–16315 (2024).

[63] Mian, O., Kaltenpoth, D. & Kamp, M. Regret-Based Federated Causal Discovery. *Proceedings of Machine Learning Research* **185 (KDD Workshop on Causal Discovery)**, 61–69 (2022).

[64] Mian, O., Kaltenpoth, D., Kamp, M. & Vreeken, J. Nothing but Regrets—Privacy-Preserving Federated Causal Discovery. *Proceedings of Machine Learning Research* **206 (AISTATS)**, 8263–8278 (2023).

[65] Li, L. *et al.* Federated Causal Discovery From Heterogeneous Data. In *International Conference on Learning Representations* (2024).

[66] Wang, Z., Ma, P. & Wang, S. Towards Practical Federated Causal Structure Learning. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, 351–367 (2023).

[67] Chickering, D. M. Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research* **3**, 507–554 (2003).

[68] Puerta, J. M., Aledo, J. A., Gámez, J. A. & Laborda, J. D. Efficient and Accurate Structural Fusion of Bayesian Networks. *Information Fusion* **66**, 155–169 (2021).

[69] Teo, Z. L. *et al.* Federated Machine Learning in Healthcare: A Systematic Review on Clinical Applications and Technical Architecture. *Cell Reports Medicine* **5**, 101419 (2024).

[70] The All of Us Research Program Investigators. The "All of Us" Research Program. *New England Journal of Medicine* **381**, 668–676 (2019).

[71] Hub, E. Elixir Scientific Programme, 2024-2028 (2023). URL https://elixir-europe.org/sites/default/files/documents/elixir-programme-24-28-full.pdf.

[72] Hersh, W. *et al.* Health Information Exchange. Tech. Rep., Agency for Healthcare Research and Quality (2015).

[73] Rehm, H. L. *et al.* GA4GH: International Policies and Standards for Data Sharing Across Genomic Research and Healthcare. *Cell Genomics* **1**, 100029 (2021).

[74] Weber, G. M. *et al.* The Shared Health Research Information Network (Shrine): A Prototype Federated Query Tool for Clinical Data Repositories. *Journal of the American Medical Informatics Association* **16**, 624–630 (2009).

[75] Commission, E. Intelligent Ecosystem to Improve the Governance, the Sharing and the Re-Use of Health Data for Rare Cancers (2022). URL https://cordis.europa.eu/project/id/101057048/results.

[76] You, S. C., Lee, S., Choi, B. & Park, R. W. Establishment of an International Evidence Sharing Network Through Common Data Model for Cardiovascular Research. *Korean Circulation Journal* **52**, 853 (2022).

[77] Alessio, Z. *et al.* Federated Causal Discovery with Missing Data in a Multicentric Study on Endometrial Cancer. *Journal of Biomedical Informatics* **in print** (2025).

[78] Chen, J., Ma, Y. & Yue, X. Federated Learning of Dynamic Bayesian Network via Continuous Optimization From Time Series Data (2024). URL https://arxiv.org/abs/2412.09814.

[79] Zhang, M. *et al.* Development and Validation of a Multi-Causal Investigation and Discovery Framework for Knowledge Harmonization (MINDMerge): A Case Study with Acute Kidney Injury Risk Factor Discovery Using Electronic Medical Records. *International Journal of Medical Informatics* **191**, 105588 (2024).

[80] Xiong, R. *et al.* Federated Causal Inference in Heterogeneous Observational Data. *Statistics in Medicine* **42**, 4418–4439 (2023).

[81] Meurisse, M. *et al.* Federated Causal Inference Based on Real-World Observational Data Sources: Application to a SARS-CoV-2 Vaccine Effectiveness Assessment. *BMC Medical Research Methodology* **23**, 248 (2023).

[82] Druzdzel, M. J. & Díez, F. J. Combining Knowledge From Different Sources in Causal Probabilistic Models. *Journal of Machine Learning Research* **4**, 295–316 (2003).

[83] Borboudakis, G. & Tsamardinos, I. Incorporating Causal Prior Knowledge as Path-Constraints in Bayesian Networks and Maximal Ancestral Graphs. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 427–434 (2012).

[84] Miksad, R. A. & Abernethy, A. P. Harnessing the Power of Real-World Evidence (Rwe): A Checklist to Ensure Regulatory-Grade Data Quality. *Clinical Pharmacology & Therapeutics* **103**, 202–205 (2017).

[85] Franklin, J. M. *et al.* Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results From the RCT Duplicate Initiative. *Circulation* **143**, 1002–1013 (2021).

[86] Wu, J. Q. *et al.* Automated Causal Inference in Application to Randomized Controlled Clinical Trials. *Nature Machine Intelligence* **4**, 436–444 (2022).

[87] Marzano, L. *et al.* Exploring the Discrepancies Between Clinical Trials and Real-World Data: A Small-Cell Lung Cancer Study. *Clinical and Translational Science* **17**, e13909 (2024).

[88] Farmer, R. E. *et al.* Application of Causal Inference Methods in the Analyses of Randomised Controlled Trials: A Systematic Review. *Trials* **19**, 23 (2018).

[89] Glick, H. A., Doshi, J. A., Sonnad, S. S. & Polsky, D. *Economic Evaluation in Clinical Trials* (Oxford University Press, 2015), 2nd edn.

[90] Griessbach, A. *et al.* Resource Use and Costs of Investigator-Sponsored Randomized Clinical Trials in Switzerland, Germany, and the United Kingdom: A Metaresearch Study. *Journal of Clinical Epidemiology* **176**, 111536 (2024).

[91] Getz, K. A. *et al.* The Impact of Protocol Amendments on Clinical Trial Performance and Cost. *Therapeutic Innovation & Regulatory Science* **50**, 436–441 (2016).

[92] Viele, K. *et al.* Use of Historical Control Data for Assessing Treatment Effects in Clinical Trials. *Pharmaceutical Statistics* **13**, 41–54 (2013).

[93] Kahan, B. C., Rehal, S. & Cro, S. Risk of Selection Bias in Randomised Trials. *Trials* **16**, 405 (2015).

[94] Catz, D. S. *et al.* Attitudes About Genetics in Underserved, Culturally Diverse Populations. *Public Health Genomics* **8**, 161–172 (2005).

[95] Chapman, E. N., Kaatz, A. & Carnes, M. Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities. *Journal of General Internal Medicine* **28**, 1504–1510 (2013).

[96] Ogier du Terrail, J. *et al.* Fedeca: Federated External Control Arms for Causal Inference with Time-to-Event Data in Distributed Settings. *Nature Communications* **16** (2025).

[97] Kusner, M., Loftus, J., Russell, C. & Silva, R. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, vol. 31, 4066–4076 (2017).

[98] Gultchin, L., Guo, S. & Malek, A. Pragmatic Fairness: Developing Policies with Outcome Disparity Control. *Proceedings of Machine Learning Research* **236 (CLeaR)**, 243–264 (2024).

[99] Pezoulas, V. C. *et al.* Synthetic Data Generation Methods in Healthcare: A Review on Open-Source Tools and Methods. *Computational and Structural Biotechnology Journal* **23**, 2892–2910 (2024).

[100] Tracy, D. H., Sherman, J. & Baron, M. Abstract 7373: Generative Bayesian Networks for Augmentation of Molecular Data From Commercial Genetics Panels. *Cancer Research* **84**, 7373–7373 (2024).

[101] Gogoshin, G., Branciamore, S. & Rodin, A. S. Synthetic Data Generation with Probabilistic Bayesian Networks. *Mathematical Biosciences and Engineering* **18**, 8603–8621 (2021).

[102] Madigan, D. *et al.* Evaluating the Impact of Database Heterogeneity on Observational Study Results. *American Journal of Epidemiology* **178**, 645–651 (2013).

[103] Rambla, J. *et al.* Beacon v2 and Beacon Networks: A "Lingua Franca" for Federated Data Discovery in Biomedical Genomics, and Beyond. *Human Mutation* (2022).

[104] Voss, E. A. *et al.* Feasibility and Utility of Applications of the Common Data Model to Multiple, Disparate Observational Health Databases. *Journal of the American Medical Informatics Association* **22**, 553–564 (2015).

[105] Rocchi, N. *et al.* *A Causal Discovery Workflow for Rare Diseases: Experts-in-the-Loop Analysis of Sparse Longitudinal Data* (2025).

[106] Castro, D. C., Walker, I. & Glocker, B. Causality Matters in Medical Imaging. *Nature Communications* **11**, 3673 (2020).

[107] Zhao, Y. *et al.* Federated Learning with Non-Iid Data. *Corr* **abs/1806.00582** (2018).

[108] Li, A., Jaber, A. & Bareinboim, E. Causal Discovery From Observational and Interventional Data Across Multiple Environments. In *Advances in Neural Information Processing Systems*, vol. 36, 16942–16956 (2023).

[109] Truong, N., Sun, K., Wang, S., Guitton, F. & Guo, Y. Privacy Preservation in Federated Learning: An Insightful Survey From the GDPR Perspective. *Computers & Security* **110**, 102402 (2021).

[110] Dwork, C., Smith, A., Steinke, T. & Ullman, J. Exposed! A Survey of Attacks on Private Data. *Annual Review of Statistics and Its Application* **4**, 61–84 (2017).

[111] Blanco-Justicia, A. *et al.* Achieving Security and Privacy in Federated Learning Systems: Survey, Research Challenges and Future Directions. *Engineering Applications of Artificial Intelligence* **106**, 104468 (2021).

[112] Wainakh, A. *et al.* Federated Learning Attacks Revisited: A Critical Discussion of Gaps, Assumptions, and Evaluation Setups. *Sensors* **23**, 31 (2022).

[113] Dwork, C., McSherry, F., Nissim, K. & Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, 265–284 (2006).

[114] Ma, P., Ji, Z., Pang, Q. & Wang, S. NoLeaks: Differentially Private Causal Discovery Under Functional Causal Model. *IEEE Transactions on Information Forensics and Security* **17**, 2324–2338 (2022).

[115] Xu, D., Yuan, S. & Wu, X. Differential Privacy Preserving Causal Graph Discovery. In *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, 60–71 (2017).

[116] Kumar Murakonda, S., Shokri, R. & Theodorakopoulos, G. Quantifying the Privacy Risks of Learning High-Dimensional Graphical Models. *Proceedings of Machine Learning Research* **130 (AISTATS)**, 2287–2295 (2021).

[117] Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D. & Xiao, X. PrivBayes: Private Data Release via Bayesian Networks. *ACM Transactions on Database Systems* **42** (2017).

[118] Rocchi, N., Stella, F. & de Campos, C. Towards Privacy-Aware Bayesian Networks: A Credal Approach. In *ECAI 2025* (2025).

[119] Schulz, N. A. *et al.* Learning Debiased Graph Representations From the Omop Common Data Model for Synthetic Data Generation. *BMC Medical Research Methodology* **24**, 136 (2024).

[120] Smits, D. *et al. An Improved Infrastructure for Privacy-Preserving Analysis of Patient Data*, 144–147 (IOS Press, 2022).