DataSynK: Causal-Symbolic EHR Synthesis for Tabular Foundation Models in Low-Resource Settings

Anonymous Authors[1]

Abstract

The chronic scarcity of labeled electronic health records (EHRs) limits the development of tabular foundation models, especially in Global South settings. While deep generative models collapse under extreme data scarcity, traditional structured generators fail to guarantee clinical plausibility. To address this, we propose DataSynK, a novel pipeline integrating causal discovery, prior medical ontology, and symbolic logic constraints to synthesize binary tabular EHRs. Empirical evaluations on real-world clinical data demonstrate that DataSynK prevents mode collapse in low-resource regimes and uniquely achieves full ontological validity. Furthermore, it significantly improves downstream predictive utility for imbalanced classes compared to purely statistical baselines, establishing a robust framework for knowledge-guided synthetic data generation.

1. Introduction

Tabular foundation models (FMs) have made significant progress through the adoption of the new paradigm for structured data modeling, In-Context Learning (ICL). Through ICL, predictions can be made in a single forward pass, without the need to update or optimize the model for new tasks (Qu et al., 2025). However, this new generalization mechanism requires an intensive pre-training regime, usually powered by large volumes of synthetically generated tables. However, in the healthcare industry, real-world clinical data are intrinsically noisy, chronically scarce, and predominantly locked in institutional silos (Price & Cohen, 2019). Worsening this scenario, the overwhelming majority of existing clinical databases are concentrated in the Global North, with clinical records standardized in English, limiting the technological potential and perpetuating algorithmic colonialism (Gichoya et al., 2022).

Deep generators (such as TabDDPM (Kotelnikov et al., 2023) and GReaT (Borisov et al., 2022)) represent the state-of-the-art in tabular data generation; however, due to their overparameterized architectures, in extreme data scarcity regimes (e.g., $n < 50$), these models suffer from mode collapse, failing to generalize in latent space (Afonja et al., 2023). In contrast, structured generators based on Bayesian networks (BNs) offer theoretical reliability, enabling parameterization of uncertainties and the generation of data with sample efficiency even in small-data regimes (Kaur et al., 2021). On the other hand, these models often lack clinical validity, making their safe use for training downstream models infeasible.

The current literature reveals a critical methodological gap: the absence of a unified framework capable of integrating (i) the sample efficiency and prior-injection stability of BNs, (ii) principled causal structure learning with identifiability guarantees, and (iii) hard symbolic constraint enforcement to ensure clinical plausibility.

To address this gap, we propose DataSynK, a framework for generating structured binary tabular EHR data to serve as a synthetic pre-training corpus for tabular FM, with particular emphasis on the extremely low-resource regime across the Global South. Unlike text-generative approaches, DataSynK operates entirely in the structured tabular domain: its output is a binary feature matrix encoding clinical entities extracted via ontology-guided NER, suitable as a pre-training corpus for tabular foundation models.

Our contributions are as follows:

1. **Novel Causal-Symbolic Architecture:** We propose DataSynK, a highly sample-efficient pipeline for binary tabular EHR synthesis that uniquely integrates Prior Knowledge Graphs (PKG), topologically constrained causal discovery, and Answer Set Programming (ASP) logical filters.
2. **Empirical Superiority in Utility and Validity:** We demonstrate that DataSynK overcomes the mode collapse inherent to deep generators in low-resource regimes. Empirically, it is the only evaluated method to

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

achieve strict positive ontological validity (Onto.Val) while simultaneously delivering the highest balanced-classification utility ($\Delta F1 = +0.093$) for downstream models.

3. **Robust Prior for Foundation Models:** We establish DataSynK as a principled synthetic data source capable of pre-training tabular foundation models under extreme clinical data scarcity. This provides a scalable pathway to mitigate representational disparities in Global South healthcare settings.

2. Related Work

Synthetic tabular data generation. The dominant paradigm for tabular data generation has evolved from conditional GAN-based methods, CTGAN and TVAE (Xu et al., 2019), to score-based diffusion models (Kotelnikov et al., 2023) and large-language-model generators (Borisov et al., 2022). While these approaches achieve competitive fidelity on benchmark datasets, their overparameterized architectures require substantial training corpora. Structured generators based on Bayesian networks (BNs) offer superior sample efficiency in small-data regimes (Kaur et al., 2021) but rely on structure-learning algorithms that optimize statistical fit without causal identifiability guarantees, producing graphs that encode conditional associations rather than true physiological mechanisms (Pearl, 2009; Peters et al., 2017) .

Tabular foundation models. TabPFN (Hollmann et al., 2023) introduced the paradigm of in-context learning for tabular classification by meta-training on synthetic data derived from structural causal priors. Although subsequent models like TabICL (Qu et al., 2025) and CARTE (Kim et al., 2024) have scaled this approach to broader feature spaces, the core reliance on data quality persists. The architectural alignment between TabPFN's inductive biases and causal structures implies that downstream ICL generalization should drastically improve when models are fed synthetic data with preserved causal mechanisms, rather than purely statistical approximations. This hypothesis is strongly supported by analogous findings in the time-series domain, where (Xie et al., 2025) showed that causal-kernel synthetic pre-training allows foundation models to achieve performance on par with real-data-pretrained baselines.

Knowledge-Guided EHR Generation. To enhance predictive utility for underrepresented cohorts, previous work has successfully employed subpopulation-specific generation (Perets & Rappoport, 2023) and knowledge-guided architectures that constrain synthetic outputs using medical ontologies (Uppalapati et al., 2025). In the realm of structurally rigorous data synthesis, the SimSUM framework (Rabaey et al., 2025) was recently proposed as a benchmark for generating synthetic EHRs via expert-crafted Bayesian Networks. While this guarantees clinical fidelity, its strict reliance on exhaustive human curation inherently limits scalability. Addressing this critical bottleneck, our proposed method automates causal graph inference through the integration of prior knowledge graphs (PKG) and strict logical constraints (ASP), achieving rigorous medical validity without the prohibitive cost of manual parameterization.

3. Our Contributions

3.1. Problem Setup

Let $\mathcal{D}_k = \{\mathbf{x}_i^{(k)}\}_{i=1}^{n_k}$ denote the dataset for clinical subgroup $k \in \mathcal{K}$, where $\mathbf{x}_i^{(k)} \in \{0,1\}^d$ encodes d binary clinical features extracted from pt-BR EHRs. The index k identifies a disease-age stratum with the *critical low-resource regime* defined as $n_k < 50$.

Objective. Learn a per-subgroup generator $G_k : \mathcal{E} \to \{0,1\}^d$ whose samples $\tilde{\mathbf{x}} \sim G_k(\varepsilon)$ simultaneously satisfy:

(i) **Statistical fidelity**: Marginal and pairwise distributions approximate $P(\mathbf{x}^{(k)})$;

(ii) **Ontological validity**: Generated feature values map to valid SNOMED-CT® codes in the reference ontology;[1]

(iii) **Causal validity**: Samples preserve the conditional independence structure of reference DAG G^* and satisfy hard clinical constraints $\mathcal{R}$.

3.2. DataSynK

Our method for generating binary tabular EHR data, as illustrated in Figure 1, consists of a four-step process described as follows:

Step 1: **Prior Knowledge Graph Construction**. Given clinical features $\{f_1, \ldots, f_d\}$, we construct $\mathcal{G}_{\text{PKG}} = (\mathcal{V}, \mathcal{E}_{\text{prior}})$ where each directed edge $(v_i, v_j) \in \mathcal{E}_{\text{prior}}$ encodes a causal hypothesis derived from SNOMED-CT ontology. An edge is included when a clinical coding guideline specifies a prerequisite or consequential relationship. A forbidden-edge constraint is imposed for mutually exclusive entities.

Step 2: **Tiered Causal Discovery**. To extract the causal skeleton G^* from the binary features, we employ a topologically constrained variant of the continuous DAG optimization of Zheng et al. (2018):

$$\min_{\mathbf{W}} \ \mathcal{L}(\mathbf{W}; \mathcal{D}_k) + \lambda \|\mathbf{W}\|_1 \quad \text{subject to}$$

[1] SNOMED CT® is a registered trademark of the International Health Terminology Standards Development Organization (IHTSDO). Used under license.

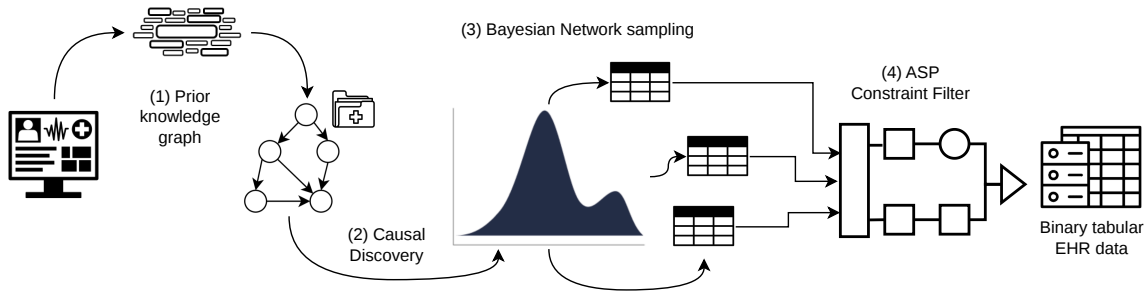

Figure 1. **Proposed pipeline architecture for synthetic binary tabular EHR data generation.** The process begins by integrating clinical consensus into a **(1) prior knowledge graph (PKG)**. A robust causal model, parameterized by initial EHR data and informed by the PKG, is learned via **(2) Causal Discovery**. Unconstrained synthetic patient cohorts are generated through **(3) Bayesian Network sampling**. These records are then passed through an **(4) ASP Constraint Filter**, which prunes biologically implausible cases by enforce rigid logical rules based on clinical guidelines, ensuring the validity of the final synthetic dataset.

$$\mathrm{tr}\big(e^{\mathbf{W}\odot\mathbf{W}}\big) - d = 0.$$

To ensure clinical directionality and computational efficiency, features are partitioned into causal tiers **(detailed in Appendix B)**. The optimization is solved pairwise across valid tier combinations, explicitly masking biologically impossible reverse-causal edges. Here, the continuous objective serves solely as a proxy to estimate structural edge weights $\mathbf{W}$ for structural recovery, rather than generative parameters. High-confidence prior edges initialize $\mathbf{W}$, and forbidden edges are masked throughout optimization. A sparsity threshold ε is applied post-optimization to extract the final boolean DAG. Features are normalized prior to optimization to mitigate the known scale-sensitivity of the NOTEARS objective (Lawrence et al., 2021).

Step 3: **Network Parameterization and Sampling**. The learned DAG G^* defines the factorization $P(\mathbf{x}) = \prod_{j=1}^{d} P(x_j \mid \mathbf{x}_{\mathrm{Pa}(j)})$. Conditional probability tables are estimated by maximum likelihood with Laplace smoothing $\alpha = 1/n_k$, critical for sparse parent configurations in the $n < 50$ regime. The samples are drawn by ancestral sampling over the topological ordering of G^*.

Step 4: **Constraint Filtering.** Raw BN samples are subjected to a hard-rejection filter implemented in Answer Set Programming. A set $\mathcal{R}$ of clinical integrity constraints, encoding parent-child code consistency and cardinality requirements, is enforced **(concrete examples of our dynamic ASP rule injection are provided in Appendix A)**: a sample $\tilde{x}$ is accepted if and only if $\tilde{x} \models r$ for all $r \in \mathcal{R}$.

4. Experimental Results

DataSynK is evaluated on a real-world clinical dataset derived from de-identified EHRs collected at a Brazilian public hospital within the SUS network[2]. The dataset comprises 643 records across three clinical conditions (MI, CVA, sepsis) and three age strata, yielding nine distinct subgroups with pronounced size imbalance ($n_k \in \{7, \ldots, 227\}$). All records are de-identified.

Our evaluation focuses on a head-to-head comparison against established generators using the most populous subgroups, *Adult-MI* ($n = 227$) and *Adult-CVA* ($n = 211$), which provide sufficient support for robust statistical and utility benchmarking. For these cohorts, we employ an 80/20 stratified split protocol.

Baselines: SDV (Patki et al., 2016), medGAN (Choi et al., 2017), CTGAN and TVAE (Xu et al., 2019), PrivBayes (Zhang et al., 2017), and TabDDPM (Kotelnikov et al., 2023). Synthetic sets generated at 1:1 ratio with training size. All metrics reported represent mean between the respective cohorts and over three generation seeds.

4.1. Benchmark evaluation

Downstream utility was assessed via a Train-on-Synthetic-Test-on-Real (TSTR) protocol (Hyland et al., 2018) using TabPFN (Hollmann et al., 2023), with synthetic labels inferred via k-NN algorithm (Cover & Hart, 1967). Overall performance was quantified by statistical fidelity (TVD, Δcorr), structural preservation (Onto.Val), and balanced

[2]This study was conducted in compliance with Brazilian National Health Council Resolution CNS 466/12 and the General Data Protection Law (LGPD). The project was approved by the institutional Research Ethics Committee under opinion number [blinded for review] and CAAE [blinded for review], with a waiver of informed consent due to the use of a retrospective and de-identified dataset.

Table 1. Benchmark evaluation averaged over two primary clinical subgroups (*Adult-MI*, n=227; *Adult-CVA*, n=211). $\Delta\text{AUC} = \text{AUC}_{\text{TSTR}} - \text{AUC}_{\text{TRTR}}$ and $\Delta\text{F1} = \text{F1}_{\text{TSTR}} - \text{F1}_{\text{TRTR}}$, where per-subgroup upper bounds are $\text{AUC}_{\text{TRTR}} \in \{0.785, 0.571\}$ and $\text{F1}_{\text{TRTR}} \in \{0.452, 0.450\}$. **Bold** indicates best result per column among non-collapsed methods. † Collapsed generators (TVD > 0.30).

Method	TVD ↓	|ΔAUC| ↓	ΔF1 ↑	Δcorr ↓	Onto.Val ↑
TRTR (real → real)	0.000	0.000	0.000	0.000	1.000
SDV	0.015	0.040	+0.040	0.061	0.052
PrivBayes†	0.336	0.094	−0.032	0.078	0.000
medGAN†	0.462	0.160	−0.006	0.082	0.000
CTGAN	**0.009**	0.017	+0.013	0.059	0.062
TVAE	0.023	**0.014**	+0.069	**0.056**	0.038
TabDDPM†	0.435	0.094	−0.037	0.139	0.007
DataSynK (ours)	0.018	0.091	**+0.093**	0.066	**0.089**

predictive utility ($\Delta\text{F1} = \text{F1}_{TSTR} - \text{F1}_{TRTR}$). For mortality prediction under label imbalance, ΔF1 is the clinically preferred metric as it penalizes failures on the minority class equally; $|\Delta\text{AUC}|$ is reported for completeness.

From the results presented in Table 1, we note that PrivBayes, TabDDPM, and medGAN produce marginal distributions that deviate substantially from the real data. DataSynK achieves the highest ΔF1, surpassing TVAE, SDV, and CTGAN. This result indicates that causally structured synthetic data trains more class-balanced classifiers, a property of direct clinical relevance for mortality prediction under label imbalance. On Onto.Val, DataSynK is the only method that substantially exceeds zero across both subgroups, confirming that statistical generation without ontological constraints does not preserve the clinically validated co-occurrence patterns of the Prior Knowledge Graph. Taken together, DataSynK is the only evaluated method to achieve positive ontological validity and the highest balanced-classification utility simultaneously, properties that are individually attainable but jointly absent in all competitive baselines.

4.2. Causal validity: What standard metrics miss?

Table 1 reveals a systematic dissociation between marginal fidelity and structural validity. For instance, while CTGAN achieves the lowest TVD, its Onto.Val remains significantly below DataSynK. DataSynK is the only evaluated method to achieve positive structural preservation across the primary clinical cohorts, providing direct evidence that our causal-symbolic approach captures a structural dimension invisible to standard fidelity metrics.

Four baselines report Onto.Val $= 0.000$, confirming that statistical generation without ontological constraints does not preserve the co-occurrence structure of the reference medical ontology.

The sensitivity of Onto.Val to the DAG sparsity threshold (ε) is empirically validated in Appendix D. Strikingly, increasing ε from 0.10 to 0.20 marginally improves TVD but completely collapses Onto.Val to zero and degrades downstream utility ($|\Delta\text{AUC}|$). This provides direct empirical evidence that standard fidelity metrics like TVD are blind to severe structural and clinical degradation, validating the necessity of an ontology-aware approach.

4.3. Privacy analysis

We evaluated the empirical privacy of the generated datasets using Distance to Closest Record (DCR), Nearest-Neighbour Distance Ratio (NNDR), and a random forest-based Membership Inference Attack (MIA-AUC). Results demonstrate that DataSynK does not memorize training records, achieving distance metrics (DCR and NNDR) strictly comparable to established non-collapsed baselines like CTGAN and TVAE. While DataSynK exhibits a slight increase in MIA susceptibility compared to purely statistical generators, this reflects a standard fidelity-privacy trade-off inherent to its superior structural preservation and clinical validity. A comprehensive numerical breakdown and the analysis of mode-collapsed generators are provided in Appendix E.

5. Conclusion

We introduced DataSynK, a causal-symbolic pipeline for high-fidelity tabular EHR synthesis in low-resource settings. By integrating prior ontologies, topologically constrained causal discovery, and Answer Set Programming, DataSynK uniquely achieves strict ontological validity while delivering superior balanced-classification utility. Ultimately, it establishes a principled framework for generating biologically plausible pre-training corpora for tabular foundation models under extreme clinical data scarcity.

6. Impact Statement

The scarcity of structured clinical data outside the Global North drives algorithmic colonialism in healthcare. DataSynK mitigates these representational gaps by synthesizing biologically valid EHRs without relying on massive, English-centric datasets. Although validated in Brazilian Portuguese, our ontology-driven, language-agnostic architecture offers a scalable, privacy-conscious framework to generate inclusive pre-training data for clinical AI in diverse low-resource settings worldwide.

References

Afonja, T., Chen, D., and Fritz, M. Margctgan: A "marginally" better ctgan for the low sample regime. In *DAGM German Conference on Pattern Recognition*, pp.

524–537. Springer, 2023.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference (MLHC)*, pp. 286–305. PMLR, 2017.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L.-C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.-C., et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations 2023*, 2023.

Hyland, S., Esteban, C., and Rätsch, G. Real-valued (medical) time series generation with recurrent conditional gans. 2018.

Kaur, D., Sobiesk, M., Patil, S., Liu, J., Bhagat, P., Gupta, A., and Markuzon, N. Application of bayesian networks to generate synthetic health data. *Journal of the American Medical Informatics Association*, 28(4):801–811, 2021.

Kim, M. J., Grinsztajn, L., and Varoquaux, G. CARTE: Pretraining and transfer for tabular learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23843–23866. PMLR, 2024.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pp. 17564–17579. PMLR, 2023.

Lawrence, A. R., Kaiser, M., Sampaio, R., and Sipos, M. Data generating process to evaluate causal discovery techniques for time series data. *arXiv preprint arXiv:2104.08043*, 2021.

Patki, N., Wedge, R., and Veeramachaneni, K. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410. IEEE, 2016.

Pearl, J. *Causality*. Cambridge university press, 2009.

Perets, O. and Rappoport, N. Ensemble synthetic ehr generation for increasing subpopulation model's performance. *arXiv preprint arXiv:2305.16363*, 2023.

Peters, J., Janzing, D., and Scholkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Price, W. N. and Cohen, I. G. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.

Qu, J., HolzmÃžller, D., Varoquaux, G., and Morvan, M. L. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.

Rabaey, P., Heytens, S., and Demeester, T. Simsum–simulated benchmark with structured and unstructured medical records. *Journal of Biomedical Semantics*, 16(1): 20, 2025.

Uppalapati, K., Abdulkareem, S., and Yimenicioglu, B. Raregraph-synth: Knowledge-guided diffusion models for generating privacy-preserving synthetic patient trajectories in ultra-rare diseases. *arXiv preprint arXiv:2510.06267*, 2025.

Xie, S., Feofanov, V., Zhang, J., Palpanas, T., and Redko, I. Cauker: Classification time series foundation models can be pretrained on synthetic data. In *The Fourteenth International Conference on Learning Representations*, 2025.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems*, 42 (4):1–41, 2017.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A. Implementation Details of Symbolic Logic Constraints (ASP)

In the DataSynK pipeline, raw Bayesian Network sampling can theoretically produce biologically implausible configurations due to the stochastic nature of the generation process. To enforce strict clinical guidelines (Step 4), we formalize medical knowledge using Answer Set Programming (ASP), a declarative logic programming paradigm. To optimize sample yield and computational efficiency, these ASP constraints are dynamically injected as deterministic masks during the forward sampling process. If a node evaluates to true for a constraining rule, its adjusted activation probability is forced to $P = 0.0$. Below, we provide concrete examples of the clinical rules utilized to prune invalid patient configurations, represented in standard ASP syntax.

Causal Chain Dependencies. Clinical events must follow a valid chronological and causal sequence. For instance, a diagnosis cannot exist without a prior underlying condition or external influence (R1), and a treatment outcome cannot be recorded without a prior diagnosis (R2). In ASP, these hard constraints are defined as integrity rules (where ':-' denotes "it cannot be the case that"):

```
% R1: A patient cannot have a diagnosis without a Tier 1 condition
:- active(Patient, Node), category(Node, diagnosis),
   not has_tier1_condition(Patient).

% R2: A patient cannot have a treatment outcome without a diagnosis
:- active(Patient, Node), category(Node, treatment_outcome),
   not has_diagnosis(Patient).
```

Clinical and Pharmacological Conflicts. Mutually exclusive events, such as administering a medication to a patient with a known allergy to that specific compound, are strictly forbidden (R5).

```
% R5: Prevent conflicting assignments (e.g., Medication vs. Allergy)
:- active(Patient, Medication), active(Patient, Allergy),
   represents_allergy_to(Allergy, Medication).
```

Orphan Symptom Prevention. To ensure structural validity, if a symptom has known clinical causes mapped in the Prior Knowledge Graph (PKG), it cannot spontaneously activate without at least one of its causal parents being active (R3).

```
% R3: An induced symptom requires at least one active parent cause
:- active(Patient, Symptom), category(Symptom, symptom),
   has_mapped_parents(Symptom),
   #count { Parent : active(Patient, Parent),
            causes(Parent, Symptom) } == 0.
```

Cardinality and Complexity Limits. To prevent generating clinically unrealistic "super-patients" with extreme multimorbidities, we enforce cardinality constraints based on clinical categories. For instance, a synthetic record is constrained to a maximum of one primary diagnosis and two underlying conditions.

```
% Enforce maximum cardinality per clinical category
:- #count { Node : active(Patient, Node),
            category(Node, diagnosis) } > 1.
:- #count { Node : active(Patient, Node),
            category(Node, underlying_condition) } > 2.
```

B. Clinical Feature Stratification and Causal Tiers

To ensure biological plausibility and prevent temporal leakage during the continuous causal discovery process the clinical features in DataSynK are strictly partitioned into a topological hierarchy. This stratification is defined by a 4-level causal tier system.

The core assumption of this topology is that clinical causality flows strictly forward through time and disease progression (from lower to higher tiers). Consequently, any structural edge projecting from a higher tier to a lower tier is mathematically masked prior to the NOTEARS optimization. The clinical semantics of each tier are defined as follows:

- **Tier 1: Baseline and Exogenous Factors (*underlying_condition*, *external_influence*).** This foundational tier encompasses pre-existing patient states and external factors that precede the current clinical episode. It includes chronic comorbidities, demographic traits, genetic predispositions, and environmental influences. By definition, these variables act as root causes and cannot be caused by acute events within the current medical encounter.
- **Tier 2: Primary Clinical States (*diagnosis*).** This tier represents the core clinical diagnoses or acute diseases identified during the patient's admission. D iagnoses are causally downstream of baseline vulnerabilities (Tier 1) and act as the primary drivers for subsequent physiological manifestations.
- **Tier 3: Manifestations (*symptom*).** This tier captures the observable clinical findings, symptoms, and physiological derangements presented by the patient. Symptoms are modeled strictly as the direct effects of the primary diagnoses (Tier 2) or underlying conditions (Tier 1).
- **Tier 4: Interventions and Results (*treatment_outcome*).** The final tier represents the culmination of the clinical pathway. It includes medical interventions, pharmacological treatments, and ultimate patient outcomes (e.g., ICU admission, survival, or mortality). These events are the downstream consequences of the patient's baseline state, diagnosis, and symptomatic presentation.

By enforcing this deterministic 4-tier topological ordering, DataSynK prevents classic statistical confounding errors, such as a symptom or death incorrectly appearing as the cause of a chronic underlying condition, while drastically reducing the computational search space for the structural optimization algorithm.

C. Computational Efficiency: Tiered Subgraph Parallelization

A well-known limitation of continuous DAG optimization frameworks, such as NOTEARS, is their cubic time complexity $\mathcal{O}(d^3)$ with respect to the number of variables d. For comprehensive EHR datasets containing dozens or hundreds of clinical features, running a monolithic optimization becomes computationally prohibitive.

To overcome this bottleneck, DataSynK leverages the clinical tier structure (described in Section 3.2) to decompose the global DAG discovery into a set of highly efficient, parallelizable sub-problems. Because clinical causality is strictly unidirectional across tier, the full $d \times d$ adjacency matrix does not need to be optimized simultaneously.

Our parallelization strategy is implemented as follows:

1. **Pairwise Subgraph Decomposition:** The global feature set is partitioned into independent subgraphs representing valid causal directions between adjacent and non-adjacent tiers (e.g., Tier 1 $\rightarrow$ Tier 2, Tier 1 $\rightarrow$ Tier 3, Tier 2 $\rightarrow$ Tier 3, etc.).
2. **Dimensionality Reduction:** For each valid pair, a local NOTEARS instance is formulated containing only the nodes belonging to those two specific tiers. This reduces the effective dimensionality of each optimization routine from d to d_{sub}, where $d_{sub} \ll d$. Consequently, the local complexity drops to $\mathcal{O}(d_{sub}^3)$.
3. **Asynchronous Parallel Execution:** Because the structural constraints strictly isolate the causal flow between these tier pairs, the subgraphs are entirely independent. We execute these local optimizations simultaneously using a multi-worker process pool, fully utilizing modern multi-core architectures.
4. **Global Recombination:** After all parallel workers converge, the local continuous weights are aggregated to reconstruct the global adjacency matrix $\mathbf{W}$. In cases where an edge is evaluated in multiple overlapping sub-contexts, the maximum learned causal strength is preserved. Finally, the global structural priors and the global sparsity threshold ε are applied.

This tiered subgraph parallelization transforms a monolithic $\mathcal{O}(d^3)$ bottleneck into a scalable set of parallel $\mathcal{O}(d_{sub}^3)$ operations. This architectural choice not only accelerates the Causal Discovery step by orders of magnitude but also mathematically prevents the algorithm from wasting computational cycles searching for biologically impossible reverse-causal edges.

D. Ablation Study: DAG Sparsity Threshold

To isolate the effect of the continuous DAG optimization on the structural validity of the generated data, we conducted an ablation study on the sparsity threshold parameter (ε) applied to the learned adjacency matrix $\mathbf{W}$. Table 2 reports the generation metrics on the primary Adult-MI cohort for $\varepsilon \in \{0.05, 0.10, 0.20\}$.

Table 2. Ablation study of the DAG sparsity threshold (ε). Bold indicates the optimal configuration chosen for DataSynK.

ε	**TVD** $\downarrow$	$\lvert\Delta$**AUC**$\rvert \downarrow$	Δ**corr** $\downarrow$	**Onto.Val** $\uparrow$
0.05	0.0180	0.1672	0.0661	0.0000
0.10	0.0175	**0.0932**	0.0640	**0.0769**
0.20	**0.0172**	0.1689	**0.0625**	0.0000

The results perfectly illustrate the "blind spot" of standard statistical metrics. For instance, increasing the threshold to $\varepsilon = 0.20$ marginally improves the statistical distance metrics (TVD drops to 0.0172 and Δcorr to 0.0625). However, this strict threshold aggressively prunes critical causal dependencies, causing the Ontological Validity (Onto.Val) to collapse entirely to 0.0000 and degrading downstream predictive utility ($|\Delta\text{AUC}|$). Conversely, a loose threshold ($\varepsilon = 0.05$) introduces spurious structural noise, similarly destroying ontological validity. Setting $\varepsilon = 0.10$ provides the optimal balance, preserving the true clinical co-occurrence patterns (highest Onto.Val) and maximizing predictive utility without sacrificing marginal fidelity.

E. Comprehensive Privacy Analysis

This section provides an extended analysis of the privacy metrics summarized in Section 4.3. We evaluate the generators across three dimensions of privacy risk: Distance to Closest Record (DCR), Nearest-Neighbour Distance Ratio (NNDR), and Membership Inference Attacks (MIA-AUC).

Distance Metrics (DCR and NNDR). DCR measures the minimum distance between any synthetic record and its closest real counterpart. As shown in Table 3, among non-collapsed methods, DCR values are highly comparable across DataSynK (0.014), CTGAN (0.016), TVAE (0.016), and SDV (0.017). This confirms that DataSynK generates novel patient representations rather than outputting near-exact copies of the training records. NNDR complements this by assessing the dispersion of synthetic samples; DataSynK's NNDR (0.567) aligns tightly with CTGAN (0.570), indicating a healthy diversity of generated cases.

Membership Inference (MIA-AUC). This metric evaluates how easily an attacker can distinguish synthetic records from real ones, where an AUC ≈ 0.5 indicates ideal indistinguishability. DataSynK achieves an MIA-AUC of 0.619, marginally higher than CTGAN (0.553) and SDV (0.583). This slight increase is a direct consequence of the fidelity-privacy trade-off because DataSynK strictly adheres to logical medical constraints (Onto.Val) and preserves predictive causal structures (ΔF1), its generated distributions are highly realistic, making it slightly more susceptible to membership inference than purely statistical (and clinically invalid) approximations.

The Artifact of Collapsed Models. Collapsed generators (PrivBayes, TabDDPM, medGAN) exhibit artificially high DCR values (e.g., 0.441 for medGAN) and NNDR values (≈ 0.98). However, their MIA-AUC scores approach 1.000, confirming they are trivially distinguishable from real data. This demonstrates that their high distance from training records is merely an artifact of generating completely invalid, low-fidelity records (as evidenced by their high TVD), rather than a genuine privacy-preserving property.

Table 3. Privacy metrics for the primary evaluation cohorts. DCR ↑ and NNDR ↑: higher = more distant from training records. MIA-AUC: closer to 0.5 = harder to distinguish from real data. † Collapsed generators (TVD > 0.30).

Method	**DCR** ↑	**NNDR** ↑	**MIA-AUC** ≈ 0.5
SDV	**0.017**	**0.637**	0.583
CTGAN	0.016	0.570	**0.553**
TVAE	0.016	0.518	0.587
DataSynK (ours)	0.014	0.567	0.619
PrivBayes†	0.317	0.971	0.999
TabDDPM†	0.368	0.967	1.000
medGAN†	0.441	0.981	1.000