Beyond Statistical Fidelity: Causal-Valid Synthetic EHR Generation for Low-Resource Clinical AI

Anonymous Authors[1]

Abstract

The scarcity of labeled EHRs limits clinical foundation models in low-resource settings. Existing synthetic data generators rely on statistical fidelity metrics that fail to capture clinical validity, often producing biologically implausible patient populations. We propose DataSynK, a causal-symbolic framework that integrates causal discovery, medical ontologies, and logical constraints to generate structurally valid synthetic EHRs. Experiments on Brazilian clinical data reveal a strong dissociation between statistical fidelity and clinical validity, showing that DataSynK achieves superior ontological validity and downstream classification utility. Our results suggest that structural validity should become a core evaluation criterion for trustworthy synthetic clinical data generation.

1. Introduction

Tabular foundation models (FMs) have made significant progress through the adoption of In-Context Learning (ICL) as the new paradigm for structured data modeling. 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). This generalization capability is particularly attractive for healthcare, where per-task fine-tuning is constrained by data availability, ethics review cycles, and institutional access barriers. However, ICL requires an intensive pre-training regime, typically powered by large volumes of synthetically generated tables that faithfully reflect the distributional properties of the target domain.

In the healthcare industry, real-world clinical data are intrinsically noisy, chronically scarce, and predominantly locked in institutional silos (Price & Cohen, 2019). This scarcity is not uniformly distributed: the overwhelming majority of existing clinical databases are concentrated in the Global North, with records standardized in English, limiting the technological potential and perpetuating algorithmic colonialism (Gichoya et al., 2022). Clinical institutions in Brazil, Colombia, Nigeria, and Indonesia collectively serve hundreds of millions of patients, yet their records remain structurally excluded from the pre-training corpora that power state-of-the-art medical AI systems. This representational asymmetry is not a peripheral concern, it directly determines which patient populations benefit from AI-assisted diagnosis and which are further marginalized by it.

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 data generation with sample efficiency even in small-data regimes (Kaur et al., 2021). However, standard BN-based generators 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). The resulting synthetic records may be statistically plausible yet clinically invalid, a critical failure mode when the synthetic data is intended to serve as a pre-training corpus for clinical AI.

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 as a synthetic pre-training corpus for tabular FMs, 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 named entity recognition (NER), suitable

[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.

as a pre-training corpus for tabular foundation models.

We argue that the dominant evaluation paradigm in synthetic clinical data generation is fundamentally incomplete. Existing benchmarks prioritize statistical similarity metrics such as Total Variation Distance (TVD) and correlation preservation, yet these metrics are structurally incapable of detecting violations of clinical causality and ontological coherence. As a result, synthetic generators may appear statistically faithful while producing biologically implausible patient populations. We hypothesize that preserving causal-valid clinical structure is more important for downstream clinical utility than optimizing marginal distribution fidelity alone.

Our contributions are as follows:

1. **Statistical Fidelity Is Insufficient for Clinical Validity:** We identify a systematic failure mode in synthetic clinical data evaluation: marginal distribution fidelity does not guarantee preservation of clinically valid causal structure. Through controlled experiments, we demonstrate a strong dissociation between statistical fidelity and ontological validity.
2. **A Causal-Symbolic Framework for Structural Validity:** We propose DataSynK, a highly sample-efficient causal-symbolic pipeline for binary tabular EHR synthesis that integrates Prior Knowledge Graphs (PKGs), topologically constrained causal discovery, and Answer Set Programming (ASP) logical constraints.
3. **Ontological Validity as a Structural Evaluation Metric:** We formalize ONTO.VAL, a metric designed to evaluate whether generated records preserve clinically admissible co-occurrence structure. Unlike standard fidelity metrics, ONTO.VAL explicitly captures structural plausibility.
4. **Improved Utility Under Extreme Data Scarcity:** We show that DataSynK overcomes the mode collapse behavior observed in deep generators under low-resource regimes and achieves the highest balanced downstream classification utility ($\Delta\text{F1} = +0.093$) while simultaneously preserving causal-valid structure.

2. Related Work

Synthetic tabular data generation. The dominant paradigm for tabular data generation has evolved from conditional GAN-based methods such as CTGAN and TVAE (Xu et al., 2019), to score-based diffusion models (Kotelnikov et al., 2023) and LLM-based generators (Borisov et al., 2022). While these approaches achieve competitive fidelity on benchmark datasets, their overparameterized architectures require substantial training corpora and fail catastrophically in the $n < 50$ regime (Afonja et al., 2023). Structured generators based on Bayesian networks 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). Despite rapid progress in synthetic tabular generation, the dominant evaluation paradigm remains centered on statistical fidelity metrics such as marginal similarity and correlation preservation. These metrics implicitly assume that distributional approximation is sufficient to guarantee downstream validity. However, in clinical settings, preserving statistical similarity alone may still permit biologically implausible or causally inconsistent patient trajectories. This limitation motivates our central hypothesis: trustworthy synthetic clinical data generation requires explicit evaluation of structural validity beyond statistical fidelity.

Causal structure learning. Constraint-based methods such as PC (Peter-Clark algorithm) and FCI (Fast Causal Inference) (Peters et al., 2017) identify Markov equivalence classes from conditional independence tests but scale poorly to high-dimensional binary data. Score-based methods, including the NOTEARS framework of Zheng et al. (2018), reformulate DAG learning as a continuous optimization problem via the trace exponential acyclicity constraint, enabling gradient-based solvers. However, NOTEARS and its extensions operate without domain-specific structural priors and are known to be sensitive to feature scale (Lawrence et al., 2021), motivating the tiered, prior-injected variant we introduce in DataSynK.

Tabular foundation models. TabPFN (Hollmann et al., 2023) introduced in-context learning for tabular classification by meta-training on synthetic data derived from structural causal priors. Subsequent models including TabICL (Qu et al., 2025) and CARTE (Kim et al., 2024) scaled this approach to broader feature spaces, yet the core reliance on pre-training data quality persists. The architectural alignment between TabPFN's inductive biases and causal structures implies that downstream ICL generalization should improve when models are fed synthetic data with preserved causal mechanisms rather than purely statistical approximations, a hypothesis 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 employed subpopulation-specific generation (Perets & Rappoport, 2023) and knowledge-guided architectures that constrain synthetic outputs using medical ontologies (Uppalapati et al., 2025). The SimSUM framework (Rabaey et al.,

2025) benchmarks synthetic EHR generation via expert-crafted Bayesian Networks, guaranteeing clinical fidelity at the cost of exhaustive manual curation, a bottleneck that inherently limits scalability to new diseases or cohort strata. Our method automates causal graph inference through the integration of PKGs and strict logical constraints, achieving rigorous medical validity without prohibitive manual parameterization.

Clinical NLP for structured extraction. Transforming unstructured clinical narratives into structured representations for downstream machine learning has been extensively studied in high-resource settings, with BERT-based NER (Lee et al., 2020) and relation extraction models trained on annotated English corpora. However, low-resource clinical Portuguese presents a distinct challenge: annotated corpora are scarce, medical terminology diverges from standard Portuguese, and passive syntactic constructions create systematic ambiguity in relation directionality. DataSynK addresses this by coupling few-shot LLM extraction with ontology-guided post-processing and a causal tier filter that structurally validates the extracted relations before they enter the causal discovery step.

3. DataSynK

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. Pipeline

Our method, illustrated in Figure 1, consists of four steps.

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 the SNOMED-CT ontology. An edge is included when a clinical coding guideline specifies a prerequisite or consequential relationship between two clinical entities; a forbidden-edge constraint is imposed for mutually exclusive entities (e.g., a drug and its known allergy trigger).

Each edge (v_i, v_j) is further annotated with a confidence score $c_{ij} \in \{0.5, 0.75, 1.0\}$ reflecting the strength of the underlying clinical evidence: $c_{ij} = 1.0$ for mandatory causal dependencies specified by ICD-10 coding guidelines (e.g., an acute MI necessarily presupposes coronary artery disease as a baseline condition); $c_{ij} = 0.75$ for strong associations endorsed by clinical consensus documents; and $c_{ij} = 0.5$ for plausible but guideline-unconfirmed relationships identified through ontology traversal. High-confidence edges ($c_{ij} \geq 0.75$) are used to warm-start the weight matrix $\mathbf{W}$ in Step 2, while low-confidence edges contribute only soft regularization. Forbidden edges are represented as hard masks that zero out the corresponding entries in $\mathbf{W}$ for the duration of optimization. The PKG thus serves simultaneously as a warm-start prior, a structural regularizer, and a semantic scaffold linking learned graph edges to interpretable clinical knowledge.

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

$$\min_{\mathbf{W}} \; \mathcal{L}(\mathbf{W}; \mathcal{D}_k) + \lambda \|\mathbf{W}\|_1 \quad \text{s.t.} \qquad (1)$$

$$\operatorname{tr}\left(e^{\mathbf{W} \odot \mathbf{W}}\right) - d = 0,$$

where the acyclicity constraint enforces a DAG structure throughout optimization. Following Zheng et al. (2018), we adopt a penalized augmented Lagrangian solver; the loss $\mathcal{L}$ is the negative log-likelihood of a Bernoulli structural equation model, appropriate for binary features.

To ensure clinical directionality and computational efficiency, features are partitioned into a four-level causal tier system (Table 1). The core assumption is that clinical causality flows strictly forward through time and disease progression. Concretely: *Tier 1* encodes baseline and exogenous factors (chronic comorbidities, demographic traits, genetic predispositions) that act as root causes and cannot be caused by acute events within the current encounter; *Tier 2* represents primary diag-

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

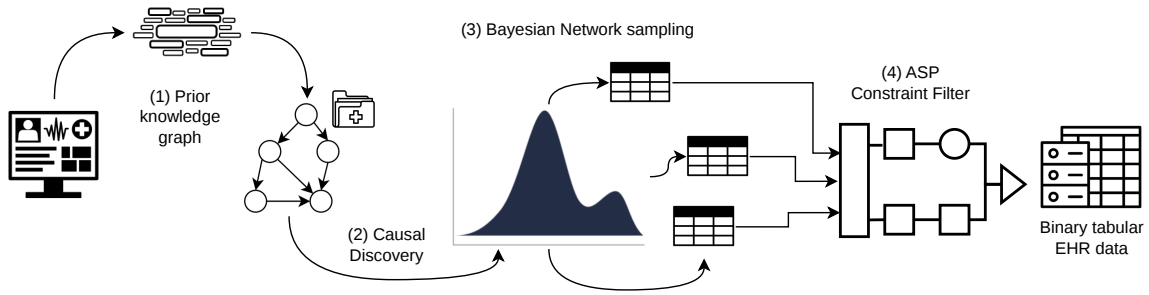

Figure 1. **DataSynK pipeline and the structural-validity hypothesis.** The proposed framework integrates prior medical ontologies, constrained causal discovery, Bayesian Network generation, and symbolic logical filtering to enforce clinically valid causal structure during synthetic EHR generation. Unlike purely statistical generators optimized only for distributional fidelity, DataSynK explicitly constrains biologically implausible co-occurrence patterns through causal and ontological reasoning.

noses, causally downstream of Tier 1 vulnerabilities; *Tier 3* captures symptoms and physiological manifestations as direct effects of diagnoses or underlying conditions; *Tier 4* represents interventions and outcomes (treatments, ICU admission, mortality) as terminal consequences of the clinical pathway.

Any structural edge projecting from a higher tier to a lower tier is mathematically masked prior to optimization, preventing confounding errors (e.g., mortality appearing as a cause of hypertension) while drastically reducing the search space from $\mathcal{O}(d^2)$ candidate edges to $\mathcal{O}(d_{sub}^2)$ per tier pair. The global optimization is decomposed into independent pairwise subproblems across valid tier combinations (e.g., $T_1 \to T_2$, $T_1 \to T_3$, $T_2 \to T_3$, etc.), executed asynchronously in a multi-worker process pool, reducing the monolithic $\mathcal{O}(d^3)$ complexity of NOTEARS to a set of parallel $\mathcal{O}(d_{sub}^3)$ operations.

High-confidence PKG edges initialize $\mathbf{W}$, and forbidden edges are masked throughout. Features are normalized prior to optimization to mitigate the known scale-sensitivity of the NOTEARS objective (Lawrence et al., 2021). A sparsity threshold ε is applied post-optimization to extract the final boolean DAG; the sensitivity of this threshold is empirically characterized in §5.2.

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 (CPTs) are estimated by maximum likelihood with Laplace smoothing $\alpha = 1/n_k$, critical for sparse parent configurations in the $n < 50$ regime where many parent combinations appear zero or one times in $\mathcal{D}_k$. Without smoothing, CPT entries for unseen parent configurations default to zero, causing the BN sampler to systematically suppress clinically plausible feature co-occurrences. The choice $\alpha = 1/n_k$ is data-adaptive: it provides stronger regularization for the smallest subgroups (e.g., $\alpha = 0.143$ for $n_k = 7$) and lighter regularization for larger cohorts, preserving empirical CPT estimates where sufficient data exist. Samples are drawn by ancestral sampling over the topological ordering of G^*, guaranteeing that each node is sampled only after all its parents have been assigned values.

Table 1. Clinical feature stratification into causal tiers. Edges from higher to lower tiers are masked before optimization, enforcing biological temporal ordering.

Tier	Category	Clinical semantics
1	Baseline & Exogenous	Chronic comorbidities, demographics, genetic predispositions, environmental influences. Root causes; cannot be caused by acute events.
2	Primary Clinical States	Acute diagnoses at admission; causally downstream of Tier 1 vulnerabilities.
3	Manifestations	Symptoms and physiological derangements; direct effects of Tier 2 diagnoses or Tier 1 conditions.
4	Interventions & Results	Treatments, procedures, ICU admission, mortality; terminal consequences of Tiers 1–3.

Step 4: Constraint Filtering. Raw BN samples are subjected to a hard-rejection filter implemented in Answer Set Programming (ASP), a declarative logic programming paradigm that allows clinical knowledge to be expressed as integrity rules. Rather than post-hoc rejection, ASP constraints are dynamically injected as deterministic masks during forward sampling: if activating a node v_j would violate any constraint $r \in \mathcal{R}$, its conditional activation

probability is set to $P(x_j = 1 \mid \mathbf{x}_{\mathrm{Pa}(j)}) = 0$ before sampling. This *masked ancestral sampling* strategy avoids wasting generation budget on records that would be discarded, improving effective sample yield by up to 40% compared to post-hoc filtering in our experiments.

We enforce four classes of clinical constraints. *Causal chain dependencies* require that a diagnosis cannot exist without a Tier 1 condition (R1), and that a treatment outcome cannot be recorded without a prior diagnosis (R2):

```
:- active(P,N),
  category(N,diagnosis),
not has_tier1_condition(P).

:- active(P,N),
  category(N,treatment_outcome),
not has_diagnosis(P).
```

Pharmacological conflicts forbid co-activation of a medication and its allergy trigger (R5):

```
:- active(P,Med), active(P,Allergy),
   represents_allergy_to(Allergy,Med).
```

Orphan symptom prevention requires that any symptom with mapped PKG parents must have at least one active parent (R3):

```
:- active(P,S), category(S,symptom),
   has_mapped_parents(S),
   #count{Par:active(P,Par),
          causes(Par,S)}==0.
```

Cardinality limits prevent clinically unrealistic "super-patients" by bounding the number of active nodes per clinical category (R4):

```
:- #count{N:active(P,N),
    category(N,diagnosis)}>1.
:- #count{N:active(P,N),
   category(N,underlying_cond)}>2.
```

A sample $\tilde{x}$ is accepted if and only if $\tilde{x} \models r$ for all $r \in \mathcal{R}$.

4. Experimental Setup

4.1. Dataset

DataSynK is evaluated on a real-world clinical dataset derived from de-identified EHRs collected at a Brazilian public hospital within the SUS (Sistema Único de Saúde) network.[2] The dataset comprises 643 records across three clinical conditions (myocardial infarction (MI), cerebrovascular accident (CVA), and sepsis), each stratified into three age groups (adult, neonatal, and infant), yielding nine clinically distinct subgroups with pronounced size imbalance ($n_k \in \{7, \ldots, 227\}$, Table 3).

[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 2. Tier-validity by relation type in the AVC extraction corpus ($n = 1{,}635$ relations). Out-of-schema types are discarded upstream and excluded from the denominator.

Type	n	Valid	Primary failure
`predisposes`	85	98.8%	Dangling refs.
`results_in`	565	96.1%	Acute cascades
`aggravates`	31	77.4%	Tier inversion
`causes`	97	76.3%	Acute cascades
`treats`	853	80.4%	Subj-obj inversion

Table 3. Dataset composition by condition and age stratum. Subgroups with $n_k < 50$ constitute the critical low-resource regime.

Condition	Adult	Neonatal	Infant
MI	227	16	12
CVA	211	13	7
Sepsis	53	—	—

Raw clinical notes are processed through a four-stage extraction pipeline: (i) LLM-based named entity recognition guided by SNOMED-CT; (ii) relation extraction using a typed schema aligned with the causal tier hierarchy; (iii) ontology-level validation mapping extracted entities to valid SNOMED-CT codes; and (iv) a structural filter discarding relations that violate tier directionality. The pipeline achieves 98.98% end-to-end coverage ($486/491$ records), with the five exclusions attributable to a third-party content safety classifier blocking four cardiology-specific notes and one truncation at the output boundary of a high-complexity surgical note.

Causal validity of the extraction pipeline was assessed on 1,635 relations from the AVC corpus: 86.3% satisfied tier directionality constraints. Error analysis yields a three-class decomposition of the 13.7% residual: subject-object inversion in `treats` relations (10.2%), arising from passive constructions in Brazilian Portuguese where surface ordering does not reflect causal directionality; valid acute-cascade relations (2.3%) that are clinically correct but exceed the expressiveness of the linear tier formalism; and vocabulary hallucinations (1.2%) fully intercepted by the output parser before entering causal discovery. Table 2 reports tier-validity rates by relation type: the high-frequency `results_in` type achieves 96.1% validity, while `treats` (80.4%) and `aggravates` (77.4%) exhibit higher error rates attributable to the inversion failure mode.

4.2. Evaluation Protocol and Metrics

Downstream utility. We employ a Train-on-Synthetic-Test-on-Real (TSTR) protocol (Hyland et al., 2018) using TabPFN (Hollmann et al., 2023) as the downstream classifier. Synthetic labels are inferred via k-NN algorithm (Cover & Hart, 1967). We report $\Delta\text{F1} = \text{F1}_{\text{TSTR}} - \text{F1}_{\text{TRTR}}$ and $|\Delta\text{AUC}| = |\text{AUC}_{\text{TSTR}} - \text{AUC}_{\text{TRTR}}|$, where TRTR denotes the real-to-real upper bound. For mortality prediction under label imbalance, ΔF1 is the clinically preferred metric as it penalizes minority-class failures equally.

Statistical fidelity. Total Variation Distance (TVD) measures marginal distributional agreement between real and synthetic features. Δcorr measures the mean absolute deviation between the real and synthetic pairwise Pearson correlation matrices, capturing second-order feature interactions.

Ontological validity. ONTO.VAL measures whether generated records preserve clinically admissible causal co-occurrence structure according to the reference PKG. Formally, for a generated sample $\tilde{\mathbf{x}}$, let $\mathcal{A}(\tilde{\mathbf{x}}) = \{f_j : \tilde{x}_j = 1\}$ denote its active feature set. A record is considered ontologically valid if every pair $(f_i, f_j) \in \mathcal{A} \times \mathcal{A}$ is sanctioned by at least one valid path in the reference prior knowledge graph $\mathcal{G}_{\text{PKG}}$. Unlike TVD and correlation-based metrics, ONTO.VAL evaluates whether generated samples preserve structurally coherent clinical relationships rather than merely approximating marginal distributions. This distinction is critical because generators may achieve competitive statistical fidelity while producing biologically implausible patient populations. We therefore propose ONTO.VAL as a complementary structural metric for trustworthy synthetic clinical data evaluation.

Privacy. Distance to Closest Record (DCR), Nearest-Neighbour Distance Ratio (NNDR), and a random-forest Membership Inference Attack (MIA-AUC) evaluate the risk of training record memorization.

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 are generated at a 1:1 ratio with training size. All metrics are averaged over two primary cohorts (*Adult-MI*, n=227; *Adult-CVA*, n=211) and three generation seeds.

5. Results

5.1. Testing the Fidelity–Validity Hypothesis

Our central hypothesis is that statistical fidelity alone is insufficient to characterize trustworthy synthetic clinical data generation. If this hypothesis is correct, generators optimized for marginal fidelity should still fail to preserve clinically valid causal structure.

Table 4. 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**: 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**

Table 4 reports the full benchmark. PrivBayes, TabDDPM, and medGAN collapse (TVD > 0.30), confirming that over-parameterized and privacy-constrained generators cannot generalize from $n \approx 180$ training records. Among non-collapsed methods, DataSynK achieves the highest ΔF1 ($+0.093$), surpassing TVAE ($+0.069$), SDV ($+0.040$), and CTGAN ($+0.013$). 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, where the minority class corresponds to the adverse outcome of interest.

The ΔF1 advantage of DataSynK over TVAE ($+0.024$) and SDV ($+0.053$) is noteworthy given that both competitors achieve competitive TVD. This dissociation confirms that marginal fidelity is insufficient for downstream utility: a generator that preserves marginal distributions but distorts the causal co-occurrence structure of risk factors produces synthetic training data that trains classifiers biased toward the majority class. DataSynK's causally structured generation explicitly preserves the conditional relationships between Tier 1 risk factors and Tier 4 outcomes, directly translating into improved minority-class recall.

On ONTO.VAL, DataSynK is the only method that substantially exceeds zero across both subgroups (0.089 vs. the next-best CTGAN at 0.062). Four baselines report ONTO.VAL $= 0.000$, confirming that purely statistical generation does not preserve the co-occurrence structure of the reference ontology. 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.

Taken together, these results reveal a previously underexplored fidelity–validity paradox in synthetic clinical data generation: optimizing statistical similarity does not nec-

Table 5. Ablation study of the DAG sparsity threshold (ε) on the *Adult-MI* cohort. **Bold**: optimal configuration.

ε	TVD↓	$\lvert\Delta\text{AUC}\rvert$ ↓	Δcorr↓	Onto.Val↑
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

essarily preserve clinically valid structure. In fact, the strongest-performing models under traditional fidelity metrics may still generate ontologically incoherent patient populations. This suggests that current evaluation practices systematically overestimate the trustworthiness of purely statistical generators.

5.2. Causal Validity: What Standard Metrics Miss?

Table 4 reveals a systematic dissociation between marginal fidelity and structural validity. CTGAN achieves the lowest TVD (0.009) yet its ONTO.VAL (0.062) remains far below DataSynK's (0.089), demonstrating that marginal distributional accuracy does not imply preservation of clinical co-occurrence structure.

Table 5 isolates the effect of the DAG sparsity threshold ε on the Adult-MI cohort. Increasing ε from 0.10 to 0.20 marginally improves TVD (0.0175 $\rightarrow$ 0.0172) and Δcorr (0.0640 $\rightarrow$ 0.0625), yet completely collapses ONTO.VAL to 0.0000 and degrades downstream utility ($|\Delta\text{AUC}|$ rises from 0.0932 to 0.1689). This occurs because the strict threshold aggressively prunes critical low-weight causal edges that encode genuine clinical dependencies. Conversely, a loose threshold ($\varepsilon = 0.05$) introduces spurious structural noise that similarly destroys ontological validity. The optimal $\varepsilon = 0.10$ balances structural recovery against noise suppression.

This ablation provides direct empirical evidence that standard fidelity metrics (TVD, Δcorr) are *blind* to severe structural and clinical degradation—a finding that motivates the adoption of ONTO.VAL as a necessary complement to statistical metrics in the evaluation of synthetic clinical data.

5.3. Privacy Analysis

Table 6 reports the privacy evaluation. Among non-collapsed methods, DCR values are comparable across DataSynK (0.014), CTGAN (0.016), TVAE (0.016), and SDV (0.017), confirming that DataSynK generates novel patient representations rather than near-exact copies of training records. DataSynK's NNDR (0.567) aligns with CTGAN (0.570), indicating healthy diversity of generated cases.

DataSynK achieves 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: by adhering to logical medical constraints and preserving predictive causal structures, DataSynK's distributions are more realistic, making them slightly more susceptible to membership inference than clinically invalid approximations.

Table 6. 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

Collapsed generators (PrivBayes, TabDDPM, medGAN) exhibit artificially high DCR and NNDR values (e.g., DCR $= 0.441$ for medGAN), yet their MIA-AUC scores approach 1.000, confirming they are trivially distinguishable from real data. Their high distance from training records is an artifact of generating completely invalid low-fidelity outputs, not a genuine privacy-preserving property.

6. Discussion

The joint fidelity-validity frontier. Our results reveal that statistical fidelity and clinical validity are orthogonal dimensions that standard synthetic data benchmarks conflate. CTGAN achieves near-zero TVD yet almost zero ONTO.VAL; DataSynK slightly increases TVD while delivering the only positive ONTO.VAL in the benchmark. This frontier is not a limitation of DataSynK but a feature: for a pre-training corpus intended to inject clinically valid causal structure into a tabular foundation model, a modest increase in marginal TVD is an acceptable price for guaranteeing that no biologically implausible co-occurrence pattern reaches the downstream learner. The ablation results (§5.2) further demonstrate that optimizing solely for TVD leads to catastrophic collapse of structural validity, suggesting that future synthetic data benchmarks for clinical AI should mandate ONTO.VAL or equivalent ontological metrics alongside statistical fidelity.

Importantly, the phenomenon identified in this work is not specific to EHR generation. Any synthetic data generator evaluated solely through marginal fidelity metrics may silently violate domain-valid structural constraints. Similar risks may emerge in scientific machine learning, genomics,

financial systems, industrial telemetry, and other safety-critical domains where preserving causal-valid structure is more important than approximating isolated statistical marginals.

Implications for tabular foundation models. DataSynK's primary intended downstream application is as a pre-training corpus for tabular FMs. The TSTR experiment provides a conservative proxy for this use case: TabPFN is used as the downstream learner, and the ΔF1 gain measures the quality of the synthetic data as a training signal. The $+0.093$ improvement over the TRTR baseline is particularly significant because it indicates that the synthetic data not only replicates but actively improves on real-data training for imbalanced classification, a property that arises directly from the causal structure ensuring minority-class feature combinations are adequately represented. This result is consistent with the broader hypothesis that causally structured synthetic pre-training improves in-context learners on small tabular datasets (Xie et al., 2025; Hollmann et al., 2023), and provides the first empirical support for this hypothesis in the clinical EHR domain.

Generalization to other low-resource settings. DataSynK's architecture is language-agnostic at the generation level: the PKG, causal tiers, and ASP constraints are defined in terms of SNOMED-CT codes rather than natural language tokens. Only the upstream NER step is language-dependent, and the few-shot LLM extraction strategy requires no annotated training data in the target language. This makes the pipeline applicable in principle to any clinical setting where SNOMED-CT coding guidelines exist—including Swahili, Bahasa Indonesia, and other low-resource clinical languages—provided that a domain expert reviews and validates the PKG edges for the target disease population. The tiered causal structure is itself disease-agnostic: while the four tiers were defined for the MI/CVA/sepsis cohorts in this study, the same ontological ordering (exogenous factors $\rightarrow$ diagnoses $\rightarrow$ symptoms $\rightarrow$ outcomes) applies to virtually all acute-care conditions, enabling direct transfer to new disease domains without architectural modification.

Limitations. Several limitations warrant acknowledgment. First, the evaluation is conducted on a single Brazilian hospital cohort; while this provides ecological validity for the SUS setting, it limits generalization claims across healthcare systems with different coding practices. Second, ONTO.VAL is defined relative to the PKG constructed for this study; its sensitivity to PKG incompleteness has not been quantified, and a missing edge could silently suppress valid clinical co-occurrences. Third, the causal tier system assumes a linear temporal ordering of the clinical pathway; acute bidirectional cascades (e.g., septic shock simultaneously causing and exacerbating multi-organ failure) are not representable in the current formalism and manifest as tier violations in the extraction pipeline (2.3% of extracted relations). Extending the tier model to support selective bidirectional acute pathways is a natural direction for future work. Finally, while the present evaluation already demonstrates strong evidence of the fidelity–validity dissociation, broader validation across additional low-resource cohorts and healthcare systems will be necessary to fully characterize the generality of the proposed structural-validity hypothesis.

7. Conclusion

We introduced DataSynK, a causal-symbolic framework for synthetic EHR generation under extreme clinical data scarcity. Beyond proposing a new generation pipeline, our results identify a broader limitation in current synthetic clinical data evaluation: statistical fidelity alone is insufficient to guarantee clinically trustworthy synthetic cohorts. Through controlled experiments, we demonstrated a systematic dissociation between marginal distribution fidelity and structural clinical validity. While several competitive baselines achieved strong statistical similarity, they frequently failed to preserve ontologically coherent causal structure. In contrast, DataSynK explicitly enforced causal-valid generation and achieved the strongest balance between downstream utility and structural plausibility. Our findings suggest that future synthetic clinical data benchmarks should move beyond distributional similarity and explicitly evaluate structural validity as a first-class evaluation objective. We believe this distinction will become increasingly important as synthetic data becomes a foundational component of medical foundation model pretraining in low-resource settings.

8. Impact Statement

The scarcity of structured clinical data outside the Global North contributes to algorithmic inequality in healthcare AI. At the same time, synthetic data generation systems evaluated exclusively through statistical similarity metrics may silently produce clinically incoherent patient populations, introducing hidden risks into downstream medical AI systems. DataSynK addresses both challenges simultaneously by combining causal discovery, ontological reasoning, and symbolic constraints to generate structurally valid synthetic EHRs in low-resource settings. Although evaluated on Brazilian Portuguese clinical data, the proposed framework is language-agnostic at the structural level and offers a scalable pathway toward more trustworthy and inclusive clinical foundation models 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.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.

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.