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

Eduarda Chagas [1 2] **Roberta Viola** [1 2] **Juarez Monteiro** [1] **Saulo Saturnino** [1 2] **Adriano Veloso** [1 2]

Abstract

The scarcity of labeled electronic health records (EHRs) limits the development of clinical foundation models in low-resource settings. Synthetic data generation offers a potential way to mitigate this scarcity, but standard evaluation protocols often emphasize statistical fidelity while providing limited insight into whether generated records preserve clinically admissible structure. We propose DataSynK, a causal-symbolic framework for binary tabular EHR synthesis that combines prior knowledge graphs, constrained causal discovery, Bayesian network sampling, and symbolic clinical constraints. Experiments on a Brazilian public-hospital cohort provide evidence that marginal statistical fidelity and ontology-guided clinical validity capture complementary aspects of synthetic data quality. In this setting, DataSynK achieves the highest ontological validity and downstream classification utility among the evaluated methods, while remaining competitive on standard fidelity metrics. These results suggest that structural validity metrics such as ONTO.VAL can complement distributional fidelity metrics when evaluating synthetic clinical data for low-resource healthcare applications. Our results characterize a fidelity–validity dissociation in synthetic clinical data: marginal statistical fidelity and ontology-guided structural validity capture complementary, and at times divergent, aspects of data quality. Code is available at: https://github.com/eduarda-chagas/DataSynk.

1. Introduction

Tabular foundation models (FMs) have recently advanced structured-data prediction through in-context learning, enabling task adaptation without per-task fine-tuning. This paradigm is attractive for healthcare, where annotated data are scarce, access is institutionally constrained, and model development is shaped by ethics and privacy requirements. However, the quality of tabular FMs depends critically on the synthetic tables used during pre-training or augmentation.

This challenge is particularly acute in low-resource clinical settings. Real-world EHRs from public hospitals in the Global South are often scarce, noisy, and underrepresented in the corpora used to develop medical AI systems. Synthetic data generation offers a potential path toward more inclusive clinical AI, but only if generated records are not merely statistically similar to real data, but also clinically plausible.

Existing tabular generators, including GAN-based, diffusion-based, and language-model-based approaches, can achieve strong distributional fidelity on benchmark datasets. Yet, in extreme low-sample clinical regimes, high-capacity neural generators may overfit, collapse to narrow modes, or fail to recover rare but clinically meaningful dependencies. Structured generators such as Bayesian networks offer a complementary small-data inductive bias, but standard structure learning may encode statistical associations rather than physiologically meaningful mechanisms. This motivates DataSynK, a causal-symbolic framework for binary EHR synthesis that combines prior knowledge graphs, constrained DAG learning, Bayesian network sampling, and symbolic clinical constraints.

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

[1]Kunumi Institute, Brazil [2]Department of Computer Science, UFMG, Brazil. Correspondence to: Eduarda Chagas <eduarda.chagas@kunumi.com>.

LatinX in AI Workshop @ ICML 2026, Seoul, South Korea. 2026.

ontology-guided named entity recognition (NER), suitable 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 provides complementary information for downstream clinical utility beyond marginal distribution fidelity alone.

Our contributions are as follows:

1. **Statistical Fidelity Is Insufficient for Clinical Validity:** We empirically characterize a fidelity–validity dissociation in a low-resource Brazilian EHR setting, showing that marginal fidelity and ontology-guided clinical validity do not coincide.

2. **A Causal-Symbolic Framework for Structural Validity:** We propose DataSynK, a clinically tiered, prior-constrained variant of continuous DAG optimization, combining established causal discovery mechanisms with ontology-derived masks and symbolic constraints for low-resource EHR synthesis.

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 Low-Resource Conditions:** We show that, in the evaluated low-resource EHR setting, DataSynK achieves the highest balanced downstream classification utility ($\Delta F1 = +0.093$) while preserving positive ontology-guided structural validity.

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). These approaches can achieve competitive fidelity on benchmark datasets, but their performance in extreme low-sample regimes remains challenging: high-capacity neural generators may overfit, collapse to narrow modes, or fail to recover rare but clinically meaningful dependencies when only a few dozen records are available. Prior work has documented the instability of GAN-based tabular synthesis under limited sample sizes (Afonja et al., 2023), motivating alternative generators whose inductive biases are better aligned with small-data settings. Structured generators based on Bayesian networks offer an attractive complementary direction because they can explicitly encode conditional dependencies and uncertainty under data scarcity (Kaur et al., 2021). However, standard BN-based generators typically rely on structure-learning algorithms that optimize statistical fit without causal identifiability guarantees, producing graphs that may 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 are necessary but not sufficient: they can detect distributional mismatch, yet they do not directly test whether generated records satisfy domain-specific clinical constraints. In clinical settings, a generator may preserve marginal statistics while still producing biologically implausible or causally inconsistent patient trajectories. This limitation motivates our central hypothesis: trustworthy synthetic clinical data generation requires explicit evaluation of structural validity alongside 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. Beyond generation, ONTO.VAL offers a candidate criterion for selecting or filtering pre-training corpora for tabular foundation models, complementing the data-quality considerations raised by Hollmann et al. (2023) and Xie et al. (2025)

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. Clinical NLP for structured extraction is more mature in English high-resource settings than in low-resource clinical Portuguese, where annotated corpora are scarce and relation directionality may be ambiguous. DataSynK uses few-shot LLM extraction with ontology-guided post-processing and tier-based validation to convert Portuguese clinical notes into binary structured features before causal discovery.

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.

Objective. Learn a subgroup-specific distribution $P_{\theta,k}$ over $\{0,1\}^d$, from which synthetic records $\tilde{x} \sim P_{\theta,k}$ are sampled by masked ancestral sampling, and 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 $\mathcal{R}$.

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

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

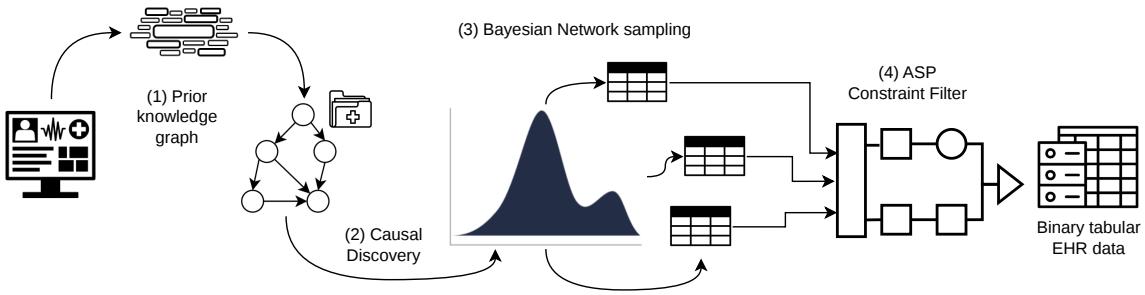

Figure 1. **DataSynK pipeline and the structural-validity hypothesis.** The proposed framework integrates prior medical ontologies, constrained causal discovery, Bayesian Network generation, and symbolic logical filtering to enforce clinically valid causal structure during synthetic EHR generation. Unlike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

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 diagnoses, 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., $T_1 \rightarrow T_2$, $T_1 \rightarrow T_3$, $T_2 \rightarrow T_3$, 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.

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 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 a data-scarce 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.

Step 4: Constraint Filtering. Raw BN samples are sub-

jected 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).

Table 2. Tier-validity by relation type in the CVA 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

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 CVA 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.

[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 3. Dataset composition by condition and age stratum.

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

4.2. Evaluation Protocol and Metrics

We evaluate synthetic data quality along four dimensions: marginal fidelity, downstream utility, privacy risk, and structural clinical validity. Marginal fidelity is measured using feature-wise total variation distance (TVD) and pairwise correlation distortion Δ_{corr}. Downstream utility is assessed through a train-on-synthetic, test-on-real (TSTR) protocol with TabPFN, reporting $\Delta F1$ and $|\Delta AUC|$ relative to the real-to-real (TRTR) reference. Privacy risk is evaluated using Distance to Closest Record (DCR), Nearest-Neighbor Distance Ratio (NNDR), and membership inference AUC (MIA-AUC). Formal definitions of all metrics are provided in Appendix A.

ONTO.VAL: Ontology-Guided Structural Validity. We introduce a generator-agnostic metric that measures whether the active clinical concepts in each synthetic record form an admissible configuration under a reference knowledge graph. For a generated record $\tilde{x}$, let $A(\tilde{x}) = \{f_j : \tilde{x}_j = 1\}$ be its active feature set, and let $\mathcal{P}_{G_{\text{PKG}}}(f_i, f_j)$ indicate whether f_i and f_j are connected by at least one clinically admissible path in G_{PKG}. We define the record-level validity indicator as

$$v(\tilde{x}) = \mathbf{1}\left[\begin{array}{l} \forall f_i, f_j \in A(\tilde{x}),\ i \neq j, \\ \mathcal{P}_{G_{\text{PKG}}}(f_i, f_j) = 1 \end{array}\right].$$

Then, ontological validity is

$$\text{ONTO.VAL}(\tilde{X}) = \frac{1}{|\tilde{X}|} \sum_{\tilde{x} \in \tilde{X}} v(\tilde{x}).$$

Although instantiated here with a clinical PKG, ONTO.VAL is a generator-independent evaluation metric applicable to any domain equipped with an admissibility graph over features, including genomics, industrial telemetry, and financial systems.

Baselines. We compare DataSynK against 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 the training data. Unless otherwise stated, benchmark tables report point estimates averaged over the two primary cohorts (Adult-MI, $n = 227$; Adult-CVA, $n = 211$) and three generation seeds.

Table 4. Benchmark point estimates averaged over two primary clinical subgroups (Adult-MI, $n = 227$; Adult-CVA, $n = 211$) and three generation seeds. $\Delta\text{AUC} = \text{AUC}_{\text{TSTR}} - \text{AUC}_{\text{TRTR}}$ 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**

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 reports the full benchmark. In our evaluation, PrivBayes, TabDDPM, and medGAN exhibit high TVD (> 0.30), indicating poor distributional recovery under the evaluated low-resource setting. 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 suggests that marginal fidelity alone may be insufficient for downstream utility in this setting. This suggests that evaluation practices based exclusively on statistical fidelity may overestimate the trustworthiness of synthetic clinical data. 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. DataSynK slightly increases TVD while achieving

Table 5. Single-cohort sensitivity analysis of the DAG sparsity threshold ε on the Adult-MI cohort. Bold: best configuration.

ε	TVD↓	|ΔAUC| ↓	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

the highest ONTO.VAL among the evaluated methods.

Taken together, these results reveal a previously underexplored fidelity–validity dissociation in synthetic clinical data generation: optimizing statistical similarity does not necessarily 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 shows that the DAG sparsity threshold ε induces a sharp trade-off between statistical fidelity and structural validity. Although increasing ε from 0.10 to 0.20 slightly improves TVD and Δ_{corr}, it collapses ONTO.VAL to zero and substantially degrades downstream utility, suggesting that clinically meaningful low-weight causal edges are being over-pruned. Conversely, a looser threshold ($\varepsilon = 0.05$) appears to introduce spurious structural edges, also harming ontological validity. The best performance at $\varepsilon = 0.10$ therefore reflects a narrow but clinically meaningful sparsity regime that balances structural recovery and noise suppression. In the current implementation, ε is selected by validation on the training cohort rather than tuned on test labels; nevertheless, this sensitivity remains a limitation and motivates future adaptive thresholding strategies based on stability selection or expert-calibrated edge confidence.

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.

Table 6. Privacy metric point estimates 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.

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

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.

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 dissociation. 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 dissociation 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 re-

sults (§5.2) further demonstrate that optimizing solely for TVD leads to structural validity instability, suggesting that future synthetic data benchmarks for clinical AI should mandate ONTO.VAL or equivalent ontological metrics alongside statistical fidelity.

Importantly, our results should not be interpreted as suggesting that structural validity replaces standard statistical fidelity metrics. Rather, the evidence points to their complementarity. TVD and correlation preservation remain necessary for detecting marginal and pairwise distributional mismatch, whereas $\mathrm{ONTO.VAL}$ captures a different failure mode: clinically inadmissible co-occurrence structure. The main empirical message is therefore not that ontological validity dominates all statistical objectives, but that clinical synthetic data evaluation is incomplete when it relies exclusively on distributional similarity. A trustworthy evaluation protocol should account for both statistical approximation and domain-specific structural admissibility.

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. Because DataSynK operates on structured clinical codes after extraction, its generative component is not tied to Portuguese text; however, transfer to new languages, institutions, and disease domains requires expert validation of the PKG and extraction pipeline.

Limitations. Several limitations warrant acknowledgment. First, our findings are based on a single Brazilian public-hospital cohort and should therefore be interpreted as evidence within this setting rather than as a claim of universal generalization across healthcare systems, where coding practices, disease prevalence, documentation conventions, and administrative workflows may differ substantially. Second, $\mathrm{ONTO.VAL}$ is defined relative to the reference prior knowledge graph G_{PKG}, making it intentionally conservative but sensitive to PKG incompleteness: missing ontology edges may incorrectly penalize valid but unencoded clinical co-occurrences. Third, the current causal tier model enforces a forward temporal abstraction of clinical progression, which improves interpretability and reduces the causal discovery search space but cannot fully represent acute bidirectional cascades, such as septic shock and multi-organ failure mutually reinforcing each other. Finally, DataSynK is currently specialized for structured binary EHR features, and extensions to continuous physiological variables, high-frequency longitudinal signals, and multimodal clinical data will require additional parameterization and validation. Together, these limitations motivate multi-institutional validation, PKG perturbation analysis, and more expressive temporally indexed clinical causal models.

7. Conclusion

We introduced DataSynK, a causal-symbolic framework for synthetic EHR generation under clinical data scarcity. Using a Brazilian public-hospital cohort, our experiments provide evidence that marginal statistical fidelity alone may overlook clinically relevant structural violations: synthetic records can approximate real feature marginals while failing to preserve ontology-guided clinical co-occurrence structure. DataSynK addresses this gap by combining prior knowledge graphs, constrained causal discovery, Bayesian network sampling, and symbolic constraint enforcement to generate structurally admissible binary EHR data. Our results suggest that structural validity metrics such as $\mathrm{ONTO.VAL}$ can complement standard fidelity, utility, and privacy metrics, particularly in low-resource clinical settings where synthetic data may support downstream model development. Future work should validate these findings across multiple institutions, quantify robustness to incomplete prior knowledge graphs, and extend the framework to richer temporal, continuous, and multimodal clinical data.

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.

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.

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.

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., Alonso, M., Odonnat, A., Zhang, J., and Redko, I. Cauker: classification time series foundation models can be pretrained on synthetic data only. In *1st ICML Workshop on Foundation Models for Structured Data*, 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. Metric Definitions

Let X and $\tilde{X}$ denote the real and synthetic datasets, respectively, each represented over d binary clinical features. Let $p_X(x_j = 1)$ and $p_{\tilde{X}}(\tilde{x}_j = 1)$ denote the empirical activation probabilities of feature j in the real and synthetic datasets.

Marginal fidelity. We measure marginal distributional agreement using the average feature-wise total variation distance:

$$\mathrm{TVD}(X, \tilde{X}) = \frac{1}{d} \sum_{j=1}^{d} \left| p_X(x_j = 1) - p_{\tilde{X}}(\tilde{x}_j = 1) \right| .$$

Correlation preservation. To assess second-order dependency preservation, we compute the mean absolute deviation between the real and synthetic pairwise Pearson correlation matrices:

$$\Delta_{\mathrm{corr}}(X, \tilde{X}) = \frac{2}{d(d-1)} \sum_{1 \le i < j \le d} \left| \rho_X(i,j) - \rho_{\tilde{X}}(i,j) \right| ,$$

where $\rho_X(i,j)$ and $\rho_{\tilde{X}}(i,j)$ denote the empirical Pearson correlations between features i and j in the real and synthetic datasets.

Downstream utility. Downstream utility is evaluated using a train-on-synthetic, test-on-real protocol. Let TRTR denote the real-to-real reference setting, where the downstream classifier is trained and tested on real data, and let TSTR denote the synthetic-to-real setting, where the classifier is trained on synthetic data and tested on held-out real data. We report:

$$\Delta F1 = F1_{\mathrm{TSTR}} - F1_{\mathrm{TRTR}}, \qquad |\Delta AUC| = |AUC_{\mathrm{TSTR}} - AUC_{\mathrm{TRTR}}| .$$

Positive $\Delta F1$ indicates that training on synthetic data improves F1 relative to the real-data reference, while lower $|\Delta AUC|$ indicates closer agreement with the real-data reference.

Ontological validity. For a generated record $\tilde{x}$, let

$$A(\tilde{x}) = \{f_j : \tilde{x}_j = 1\}$$

be its active feature set. Let $\mathcal{P}_{G_{\mathrm{PKG}}}(f_i, f_j)$ indicate whether at least one clinically admissible path connects f_i and f_j in the prior knowledge graph G_{PKG}. Ontological validity is defined as:

$$\mathrm{ONTO.VAL}(\tilde{X}) = \frac{1}{|\tilde{X}|} \sum_{\tilde{x} \in \tilde{X}} \mathbf{1} \left[\forall f_i, f_j \in A(\tilde{x}),\ i \neq j : \mathcal{P}_{G_{\mathrm{PKG}}}(f_i, f_j) = 1 \right] .$$

This metric intentionally implements a conservative validity criterion: a synthetic record is counted as valid only if all active clinical concepts are mutually supported by the reference PKG. As a consequence, $\mathrm{ONTO.VAL}$ is sensitive to incompleteness in G_{PKG}, a limitation discussed in Section 6.

Distance to Closest Record. Distance to Closest Record (DCR) measures the minimum normalized Hamming distance between each synthetic record and the real training set:

$$\mathrm{DCR}(\tilde{X}, X) = \frac{1}{|\tilde{X}|} \sum_{\tilde{x} \in \tilde{X}} \min_{x \in X} \frac{1}{d} \sum_{j=1}^{d} \mathbf{1}[\tilde{x}_j \neq x_j].$$

Higher DCR values indicate that synthetic records are farther from the nearest real training record.

Nearest-Neighbor Distance Ratio. Nearest-Neighbor Distance Ratio (NNDR) compares the distance from each synthetic record to its nearest and second-nearest real neighbors:

$$\mathrm{NNDR}(\tilde{X}, X) = \frac{1}{|\tilde{X}|} \sum_{\tilde{x} \in \tilde{X}} \frac{d_1(\tilde{x}, X)}{d_2(\tilde{x}, X) + \epsilon},$$

where $d_1(\tilde{x}, X)$ and $d_2(\tilde{x}, X)$ denote the normalized Hamming distances to the closest and second-closest real records, respectively, and ϵ is a small numerical constant. Values closer to one suggest that synthetic records are not concentrated around a single memorized training example.

Membership inference risk. Membership inference risk is evaluated by training a random-forest classifier to distinguish real training records from synthetic records. We report the area under the ROC curve:

$$\text{MIA-AUC} = \text{AUC}\left(h_{\text{MIA}}(z), y_{\text{member}}\right),$$

where h_{MIA} is the membership inference classifier and y_{member} indicates whether a record originates from the real training set. Values closer to 0.5 indicate lower distinguishability between real and synthetic records, whereas values closer to 1.0 indicate higher membership inference risk.