Fair Synthetic Data Generation from Limited Samples: A MIMIC-IV Case Study

Nitish Nagesh
University of California, Irvine
Irvine, USA
nnagesh1@uci.edu

Salar Shakibhamedan
TU Wien
Vienna, Austria
salar.shakibhamedan@tuwien.ac.at

Mahdi Bagheri
University of California, Irvine
Irvine, USA
mahdib1@uci.edu

Ziyu Wang
University of California, Irvine
Irvine, USA
ziyuw31@uci.edu

Nima TaheriNejad
Heidelberg University
Heidelberg, Germany
nima@uni-heidelberg.de

Axel Jantsch
TU Wien
Vienna, Austria
axel.jantsch@tuwien.ac.at

Amir M. Rahmani
University of California, Irvine
Irvine, USA
amirr1@uci.edu

***Abstract*—Synthetic healthcare data generation offers a promising solution to research limitations in clinical settings caused by privacy and regulatory constraints. However, current synthetic data generation approaches require specialized knowledge about training generative models and require high computational resources. In this paper, we propose FairTabGen, an LLM-based tabular data generation framework that produces high-quality synthetic healthcare data using only a small subset of the original dataset. Our method combines in-context learning, prompt curation and embedding structural constraints for data synthesis. We evaluate performance on MIMIC-IV dataset. Our method uses 99% less data and achieves a 40% reduction in fairness through unawareness while better capturing the real-data distribution. However, we observe data distribution of racial groups is skewed affecting demographic parity. We thereafter apply bias mitigation algorithms to improve overall fairness, balancing fidelity and fairness and paving for way for synthetic data generation with limited samples.**

***Index Terms*—synthetic data, tabular health data, counterfactual fairness, bias mitigation**

I. Introduction

Real-world health datasets are prone to systemic biases through protected attributes such as gender, age and race [1]. These biases arise from a multitude of factors including limited health access for certain demographic groups and use of biased clinical proxies [2]. While current evaluation relies on evaluating predictive performance of machine learning models, high accuracy alone is not sufficient since it can negative impacts on vulnerable groups. Therefore, there is a need to assess data fairness to ensure equitable health outcomes [3]. Advances in generative models such as generative adversarial networks (GANs) [4] have enabled high-fidelity data generation across multiple domains. Further, large language models (LLMs) are being used as data generators in tabular settings [5] enabling data augmentation at a larger scale. Synthetic data serves as a mechanism to mitigate bias by imposing fairness constraints in the data generation process in addition to algorithmic approaches for in-, pre- and post-processing techniques [6, 7]. Current efforts for synthetic data generation [8, 9] rely on extensive data curation and training generative models which are time consuming and require high computational resources. The challenge therefore is to generate synthetic data that captures underlying data characteristics without further amplifying biases in resource constrained settings. We seek to answer how to generate synthetic data that retains predictive utility while remaining fair across sensitive attributes using limited data samples. In this work, we propose **FairTabGen**, a prompt-based framework that leverages Large Language Models (LLMs) to synthesize fair, high-utility tabular health data under resource constraints. Our contributions are as follows:

- We present a method to generate tabular synthetic data using curated prompts that incorporate utility and fairness constraints without extensive fine tuning.
- We evaluate this approach across multiple metrics including counterfactual fairness, predictive utility and bias mitigation metrics.
- We demonstrate the effectiveness of our method on the MIMIC-IV dataset, showing comparable fairness using 99% less source data while maintaining competitive predictive utility.

Nitish Nagesh and Salar Shakibhamedan contributed equally to this work.

II. Methodology

Objective. The goal is to generate synthetic data that is matches the distribution to the original data while satisfying fairness constraints for protected attributes and target outcome.

Fairness-aware synthetic data generation. Our methodology implements a modular framework for high-fidelity and fairness-aware tabular data synthesis, illustrated in Figure 1. It comprises four stages – first being data curation and preprocessing to extract high-quality seed samples, followed by synthetic data generation through prompt curation, in-context learning and constraint operationalization which then feeds into the multi-dimensional utility and fairness evaluation module and eventually bias mitigation, in which we compare four bias mitigation strategies and retain the one whose fairness

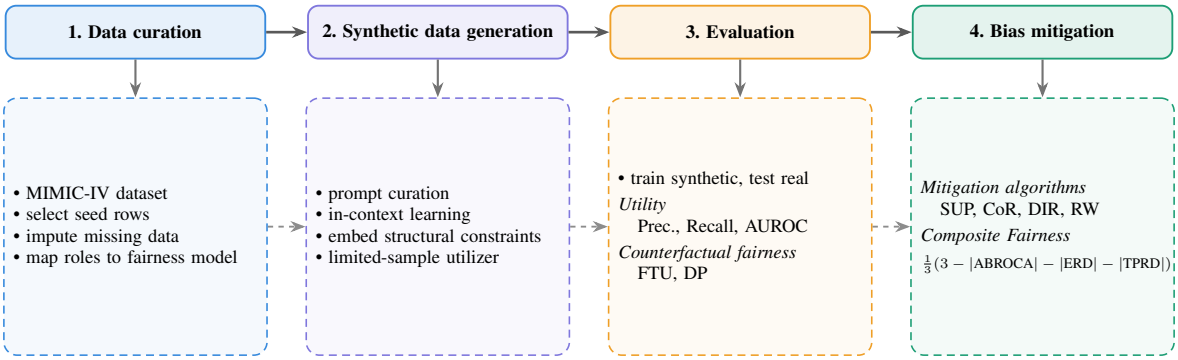

Fig. 1: FairTabGen pipeline on the MIMIC-IV case study. Curated seed data is tagged with structural-causal roles (X sensitive, Z confounders, W mediators, Y outcome) and supplied in-context to a frozen LLM under structural constraints; synthetic data is assessed under train-synthetic–test-real for utility (precision, recall, AUROC) and counterfactual fairness (FTU, DP); and residual bias is reduced by four pre-processing algorithms (SUP, CoR, DIR, RW), with the composite fairness score used to select the strategy whose deviation from the real baseline is smallest.

profile deviates least from the real-data baseline. To address systemic disparities, we apply four bias mitigation strategies prior to downstream model training [10] including Suppression (SUP), Correlation Remover (CoR), Disparate Impact Remover (DIR) and Reweighing (RW). SUP achieves fairness through unawareness by excluding sensitive attributes from the feature set, CoR applies a linear transformation to remove sensitive attributes and their correlations with non-sensitive features, DIR edits feature values so distributions are similar across groups while preserving rank-order, and RW assigns instance weights by group membership and label to neutralize dependencies.

Prompt curation We leverage an existing prompt structure [9] to integrate curate data samples and supply it to a large language model for data synthesis. The components of the prompt and accompanying contents are elucidated in Table I.

Data Synthesis and Training. We adopt the Train Synthetic, Test Real (TSTR) paradigm [10]. By training predictive models on generated data and validating them against a held-out test set from the original sample, we can assess robustness of data quality while satisfying fairness constraints.

Utility Metrics. We use area under the receiver operating curve (AUROC) to measure utility, which is a combination of precision and recall [11]. Precision quantifies how realistic the synthetic data is by calculating the fraction of synthetic points that reside within the manifold of the real data. Recall measures the diversity of the generated data by calculating the fraction of real data points that fall within the manifold of the synthetic distribution. A higher value of AUROC indicates that the synthetic data and real have have similar distributions.

Counterfactual Fairness. We compute counterfactual fairness [12] through Fairness Through Unawareness (FTU) and Demographic Parity (DP). FTU asserts that a model is fair if its prediction remains constant when the protected attribute is interchanged. Ideally, this value approaches zero. DP measures the independence of the prediction and the protected attribute, calculated as the absolute difference in positive prediction rates between groups.

TABLE I: Structure of the synthetic-data generation prompt. Blue tokens are dataset-specific placeholders instantiated per dataset.

Component	Content
System role	You are a tabular synthetic data generation model. You are a synthetic data generator. Your goal is to produce data which mirrors the given examples in causal fairness within a structural causal model (SCM) framework and feature and label distributions but also produce as diverse samples as possible. I will give you real examples first.
Context	Assign roles: Leverage your knowledge about healthcare and causal fairness to generate realistic but diverse samples. Generated data should consider |{Sensitive Features}| as the sensitive attribute (X), |{Confounders}| as the confounders (Z), |{Dataset Label}| as the target variable/Outcome (Y), and the rest of the features as the mediator attribute (W). Generated data must be structured to allow evaluation of fairness through causal pathways, capturing both direct and indirect effects of the sensitive attribute on the target variable, as well as possible confounding influences.
Output	The output should be a markdown code snippet formatted in the following schema: |{Header}|
Example	|{In-context Samples}|
Constraint	DO NOT COPY THE EXAMPLES but generate realistic but new AND diverse samples which have the correct label conditioned on the features.

Composite Fairness. To provide a comprehensive assessment of fairness, we utilize metrics that evaluate both predictive quality and the disparity between groups. ABROCA (Area Between ROC Curves) quantifies the divergence between the ROC curves of different demographic groups, where a value closer to zero indicates higher fairness [13]. The True Positive Rate Difference (TPRD) measures the gap in the ratio of correctly identified positive cases between the protected and unprotected

TABLE II: Data quality and counterfactual fairness evaluation across real and synthetic datasets. ↑ indicates higher is better, ↓ indicates lower is better.

Dataset	Method	Data Quality ↑			Counterfactual Fairness ↓	
		Prec.	Recall	AUROC	FTU	DP
MIMIC-IV	Original	0.701 ± 0.006	0.647 ± 0.006	0.851 ± 0.003	0.042 ± 0.003	0.065 ± 0.013
	DECAF[8]	0.670 ± 001	0.600 ± 002	$\mathbf{0.816 \pm .001}$	0.097 ± 0.001	0.012 ± 0.001
	CLLM[9]	0.684 ± 0.005	0.736 ± 0.002	0.639 ± 0.001	0.081 ± 0.002	0.024 ± 0.001
	FairTabGen	$\mathbf{0.905 \pm 0.003}$	$\mathbf{0.969 \pm 0.008}$	0.644 ± 0.003	$\mathbf{0.025 \pm 0.003}$	$\mathbf{0.001 \pm 0.005}$

groups [14], while the Error Rate Difference (ERD) captures the disparity in overall misclassification rates, providing insight into whether one group bears a disproportionate burden of model error [15]. We normalize these metrics into a single composite fairness score [10], where values closer to one indicate higher fairness. We evaluate the synthesized data on three fronts: counterfactual fairness (FTU, DP), distributional fidelity and downstream utility (precision, recall, AUROC), and residual bias after applying the four bias mitigation strategies, scored by ABROCA, TPRD, and ERD.

Experimental Setup. We implement our framework in Python and PyTorch and orchestrate synthetic data using the OpenAI API using `gpt-4o` model [16]. We set the temperature to 0.90 and top-p sampling to 0.95. We generate 100 samples per batch to account for context window limitations. We perform intial experiments and data curation on an Intel Core i7-10700K CPU with 16 GB. For training classifiers and data generation at scale, we use the NVIDIA RTX 3090 and 4090 GPUs hosted on an Ubuntu-based environment. We reproduce baselines [8, 9] to ensure comparative analysis.

III. Results

Dataset and Preprocessing. We use MIMIC-IV [17] for its breadth of clinical features across a diverse population. From the cohort of 418,640 records we extract 200 seed rows as in-context samples and impute missing values.

Prompt attributes. The roles of clinical variables are defined based on the standard fairness model [18] and is outlined below. Prompts are processed in batches of 100 over five independent runs.

- **Sensitive Attribute (X):** `race` (binarized to White and Black/African American).
- **Confounders (Z):** `age` and `gender`, which influence both and the outcome.
- **Mediators (W):** 23 clinical features including the Charlson Comorbidity Index, elective status, and 20 frequent diagnostic codes (e.g., ICD codes for heart failure, hypertension, and diabetes).
- **Target Outcome (Y):** Length of Stay (`los_seconds`), binarized at a 4-day threshold (345600 seconds).
- The `header` represents list of all the features and 200 `in-context samples` are considered. The prompts are processed in batches of 100 samples to accommodate context window limitations and optimize token throughput. To ensure statistical reliability, all experiments are repeated across five independent runs.

Data Distribution and Representation. The original data is skewed towards the majority racial group while gender is relatively balanced. In our synthesis, the generated data for the majority racial group decreased by 30.27%, while the representation for the minority group increased by 30.75%. In contrast, gender distribution remained highly consistent with less than 1% variation between male and female groups. While racial representation in generated samples is balanced across groups under consideration, the ensuing results should be interpreted within this context.

Utility and Counterfactual Fairness Analysis. We compare FairTabGen against a GAN-based [8] and an LLM-based [9] generator. In Table II, precision and recall measure distributional fidelity, indicating how well synthetic samples occupy, and cover, the real-data manifold [11]), whereas AUROC measures the downstream predictive ranking under the train-synthetic–test-real (TSTR) protocol with an XGBoost classifier. FairTabGen attains the highest fidelity, indicating synthetic samples that are both realistic and diverse, while DECAF yields the highest downstream AUROC among generators. Thus DECAF better preserves the global decision boundary, whereas FairTabGen more faithfully reproduces the real-data distribution. On counterfactual fairness, FairTabGen achieves the lowest FTU and DP of all methods, supporting its use for training downstream clinical models.

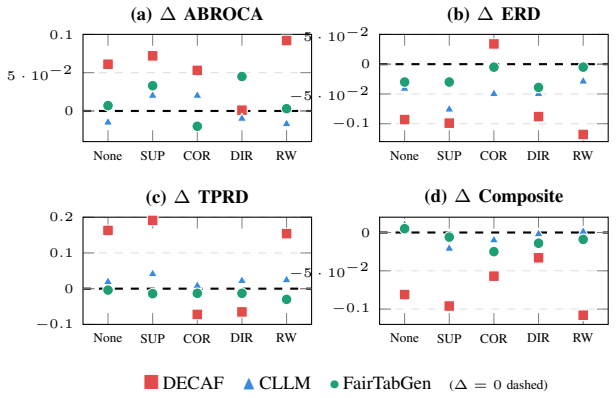

Fig. 2: Signed deviation from the real-world baseline (Δ = Method − Real) across mitigation strategies; the dashed line marks $\Delta = 0$. Each panel is scaled to its own metric.

Bias Mitigation and Composite Fairness. We assess four bias mitigation strategies including supression, correlation removal, disparate impact removal and reweighing on individual

and composite fairness metrics. Figure 2 shows the signed deviation (Δ) of each method from the real-world baseline. For CoR and RW, FairTabGen shows very limited deviation in ERD and TPRD. On composite fairness, FairTabGen exhibits the smallest overall deviation from the real baseline and consistently outperforms the GAN baseline, though LLM-based method is marginally closer for DIR and RW individual strategies.

IV. Discussion

FairTabGen generates fair tabular synthetic data with a frozen LLM under structural constraints, using less than 1% of the source data while achieving fairness gains and maintaining distributional fidelity. Across the four bias-mitigation strategies, its composite fairness deviates least from the real-data baseline overall, supporting the use of such data for training clinical models. Despite these promising results, our work has limitations. The data distribution across racial groups is skewed and does not capture the full range of clinical features. Moreover, reliance on black-box models limits reproducibility in clinical environments. To address these, we aim to extend the framework across multiple baselines and datasets, incorporate open-source models, and generate data at larger scale. By doing so, we contribute an extensible, lightweight framework for synthetic data generation.

References

[1] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4(1):123–144, 2021.

[2] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[3] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

[4] Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082, 2023.

[5] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

[6] Mahed Abroshan, Andrew Elliott, and Mohammad Mahdi Khalili. Imposing fairness constraints in synthetic data generation. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2269–2277. PMLR, 02–04 May 2024.

[7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[8] Boris Van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela Van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

[9] Nabeel Seedat, Nicolas Huynh, Boris Van Breugel, and Mihaela Van Der Schaar. Curated llm: synergy of llms and data curation for tabular augmentation in low-data regimes. In *Proceedings of the 41st International Conference on Machine Learning*, pages 44060–44092, 2024.

[10] Qinyi Liu, Oscar Deho, Farhad Vadiee, Mohammad Khalil, Srecko Joksimovic, and George Siemens. Can synthetic data be fair and private? a comparative study of synthetic data generation and fairness algorithms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 591–600, 2025.

[11] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

[12] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[13] Josh Gardner, Christopher Brooks, and Ryan Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 225–234, 2019.

[14] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[15] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[17] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and S Horng IV. Mimic-iv-ed. *PhysioNet*, 2021.

[18] Drago Plečko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024.