

Fit for What? Probing a Primary-Care Foundation Model Through Its Synthetic OMOP CDM Trajectories

Stergios Ntanavaras¹, Egill A. Fridgeirsson¹, Peter R. Rijnbeek¹, Tom M. Seinen¹

¹Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands

Background

Electronic health records (EHRs) capture patient care over time as sequences of visits, diagnoses and prescriptions¹. This has driven interest in clinical foundation models: transformers pretrained on patient trajectories with a task-agnostic objective, typically next-event prediction^{2,3}. However, such models are not directly interpretable, but comparing their generated trajectories with real ones is a practical way to probe what they have learned¹. The same comparison tests the synthetic data themselves, which could enable data sharing without exposing real patients⁴ and reproducible benchmarking⁵. One comparison thus answers two questions: where the model's learned structure diverges from real EHRs, and which analyses its synthetic data can support.

Yet evaluations of generated EHRs have relied largely on aggregate or generic similarity, leaving open whether the preserved structure is what a given analysis needs⁶. Task-agnostic pretraining privileges no single task, and characterization, prediction and causal inference⁷ each depend on different properties of the data. A model can match marginal frequencies yet miss the order of events, whether related concepts still appear together, or the treatment-confounder-outcome dependencies a causal estimate requires, which no aggregate metric detects⁶. Fitness for use can therefore be established only per task class⁶, which has not been tested for synthetic OMOP CDM data from primary care. This work trains a Dutch primary-care foundation model on IPCI data and evaluates its synthetic trajectories across six tiers of increasing fidelity and utility complexity. A privacy check is also performed to determine whether the maintained structure indicates learning rather than memorization.

Methods

Dataset: We used the Integrated Primary Care Information (IPCI) database⁸, a Dutch primary care registry containing approximately 3 million patients mapped to the OMOP CDM. For this feasibility study, we selected a representative sample of 200,000 patients with a minimum observation period of 365 days. In addition to that, clinical events were restricted to condition occurrences, drug exposures, and visits, covering full longitudinal histories from patient registration to the end of follow-up.

Sequence representation: Inspired by CEHR-GPT⁹, each patient was encoded as a single token sequence using a custom vocabulary of 16,353 tokens. The vocabulary consists of structural tokens that define observation zones (*OBS_START*, *OBS_END*) and event blocks (*VS/VE* for visit-based blocks, *ES/EE* for blocks with no visit); programmatic tokens that encode time gaps between events (*D1-D1080* days, *LT* for gaps exceeding three years), supply days (*DS1-DS365*), calendar year (*YEAR_1900-YEAR_2030*), and patient age (*AGE_0-AGE_119*); and OMOP concept IDs representing clinical entities. The sequences were structured into three time zones — pre-observation history, the formal observation period, and post-observation events — preserving the full longitudinal trajectory of each patient.

Model architecture and training: A 34.6M-parameter decoder-only transformer (GPT-2-style, implemented with NanoGPT¹⁰; 8 layers, 8 heads, 512-dimensional embeddings) was trained from scratch

by next-token prediction. It used 179,457 patients for training, with 19,939 held out for validation (90/10 split), for up to 20 epochs with early stopping.

Generation and Evaluation: 199,396 synthetic sequences were generated autoregressively from a single *[START]* token, requiring the model to learn all population characteristics including demographic distributions entirely from training data without conditioning signals. We then probed the structure learned by the model through the synthetic trajectories it generated, evaluating them across six increasingly demanding tiers of fidelity and utility, including marginal (Tier 1), temporal (Tier 2), associational (Tier 3), predictive (Tier 4), and causal (Tier 5-6) properties. These tiers evaluate both the quality of the synthetic data and the extent to which the model captured clinically relevant properties of real EHRs.

Results

To assess whether the model reproduces the structure it learned, we compare synthetic trajectories with real trajectories under the same truncation and observation constraints. This is important because 14.41% of real sequences were truncated to the 2,049-token context, with truncated real sequences averaging 825.7 tokens per patient, compared with 782 tokens for synthetic sequences. Complete real sequences average 1,071 tokens, so comparing them directly with synthetic sequences could confound model fidelity with differences in context length. To assess whether synthetic data supports similar scientific conclusions, we compare identical analyses performed on synthetic data and full (not-truncated) real data.

Metric	Real (full)	Real (trunc-2048)	Synthetic
Patients	199,396	199,396	199,396
Female	51.2%	51.2%	50.6%
Deceased	4.3%	2.1%	1.8%
Age (median)	37.0	35.0	34.0
Obs period median (yr)	5.0	5.0	5.0
Total events	56,262,209	41,992,395	40,447,417
Events/patient (med)	145.0	145.0	147.0
Distinct conditions	1,430	1,423	1,430
Distinct drugs	13,204	12,691	13,204
Visits/patient (med)	45.0	43.0	42.0

Table 1. Marginal fidelity key metrics for real data (full and truncated) and synthetic data, 199,396 patients each.

Tier 1 (marginal fidelity): Truncation changes the real data more than the model changes the synthetic data (Table 1). Cutting real sequences at 2,049 tokens removes about 25% of events and halves the deceased proportion, because the cut sequences lose their endings. Against the truncated real data, the synthetic data is close on every metric: patient count, sex split, observation period and median events per patient match, and total events, deceased proportion and median age differ only slightly. The synthetic data also covers the full vocabulary, with the same number of distinct conditions and drugs as the full real data, whereas truncation drops some from the real side. The model therefore kept the marginal structure it was trained on, and the synthetic data can answer marginal distributional questions well. Against full real data it understates total events, deaths and median age, but that gap mostly reflects the truncation.

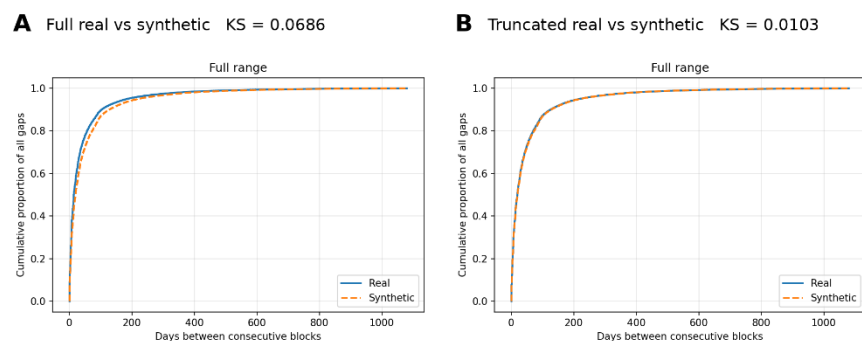


Figure 1. Tier 2 inter-block gap distribution (days between consecutive visits), real versus synthetic: (A) full real data, (B) real data truncated.

Tier 2 (temporal fidelity): Here we test whether the synthetic data reproduces the time between a patient's consecutive visits: each visit is one block of events, and the inter-block gap is the number of days between one block and the next. We compare the real and synthetic gap distributions with the Kolmogorov-Smirnov (KS) statistic, the maximum vertical distance between the two empirical cumulative distribution functions, where 0 means identical distributions and 1 means completely separated ones¹¹. Against full real data, the synthetic gaps run slightly longer, so the synthetic curve rises more slowly at short gaps (KS = 0.0686, Figure 1A). Against truncated real data, the two curves almost coincide across the whole range (KS = 0.0103, Figure 1B). Relative to the observation period, 59.4% of synthetic events fall before it, 40.0% during it and 0.55% after it, against 57.9%, 41.6% and 0.52% in truncated real data, but 50.7%, 48.6% and 0.75% in full real data. The model therefore reproduced the timing and placement of visits, and most of the differences against full real data come from the longer real sequences the model never saw.

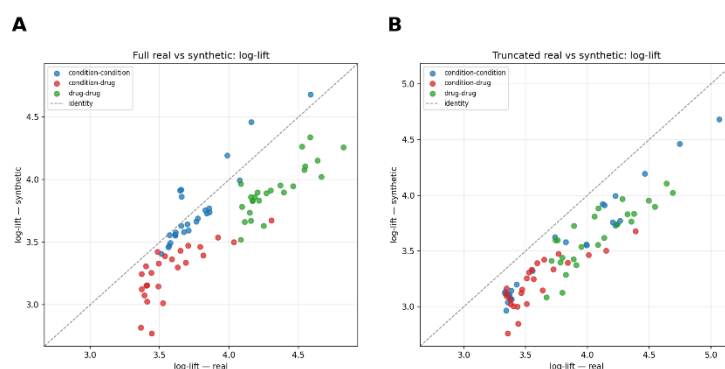


Figure 2. Tier 3 associational fidelity: log-lift of the highest-lift pairs in real versus synthetic data, by pair type, for (A) full real data and (B) real data truncated. The dashed line marks perfect agreement.

Tier 3 (associational fidelity): Here we test whether pairs of concepts that occur together more often than expected by chance in real patients do so to the same degree in synthetic data. We selected the top 25 pairs by lift for each pair type (condition-condition, condition-drug and drug-drug), where lift is the observed co-occurrence of a pair divided by the co-occurrence expected if the two concepts were independent, and we plot it on a log scale. Most points fall below the identity line, so the model preserved the associations but weakened them (Figure 2A). Condition-condition pairs sit closest to the

line, whereas condition-drug and drug-drug pairs fall further below it. The shortfall is larger than truncated real data (Figure 2B), because truncation raises the real lift by removing separate occurrences of each concept more than joint ones, so the synthetic data keeps the strong associations but understates their strength.

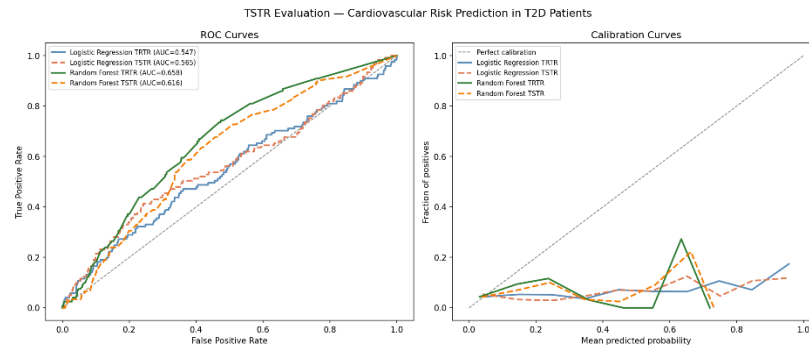


Figure 3. Tier 4 predictive utility: ROC (left) and calibration (right) curves for cardiovascular risk prediction in type 2 diabetes patients, for models trained on synthetic data and tested on real data (TSTR) and trained and tested on real data (TRTR), using logistic regression and random forest. AUC, area under the ROC curve.

Tier 4 (predictive utility): Here we test whether a prediction model trained on synthetic data works on real patients, using cardiovascular risk prediction in patients with type 2 diabetes as the task. We train logistic regression and random forest models either on synthetic data (train on synthetic, test on full real, TSTR) or on full real data (train on full real, test on full real, TRTR)¹², and compare discrimination (ROC curves) and calibration on the same real test patients (Figure 3). The TSTR and TRTR curves lie close together for both models, with the TSTR logistic regression slightly above and the TSTR random forest slightly below its TRTR counterpart, so training on synthetic data gives about the same discrimination as training on real data. Discrimination is modest even for real-trained models, though, and the logistic regression is close to chance, so this test has limited power to separate real from synthetic data. All models are poorly calibrated, with predicted risks well above the observed fraction of events (Figure 3, right). The synthetic data is therefore usable for developing a risk model of this kind, but it supports only limited conclusions about predictive utility, and any model trained on it would need recalibration on real data.

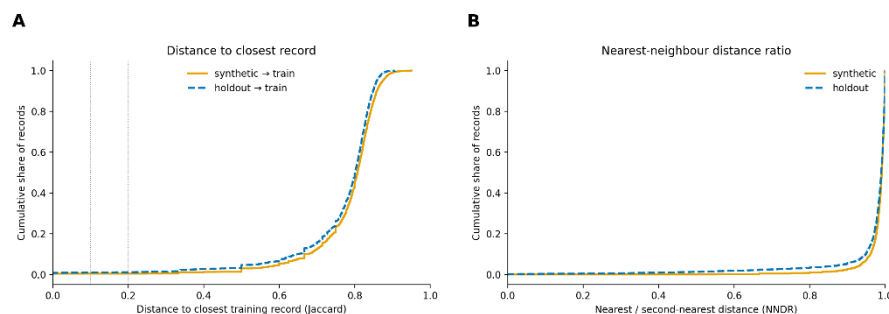


Figure 4. Privacy check: cumulative distribution of the distance to the closest training record (Jaccard distance) for synthetic patients (synthetic to train) and for real patients held out from training (holdout to train).

Privacy check: Here we test whether the utility measured in Tiers 1 to 4 could come from the model copying its training patients. For each synthetic patient we find the closest training patient by Jaccard distance on concept sets (distance to closest record, DCR)¹³, where 0 means the same concepts and 1 means nothing in common, and we compare this with the same distance for real patients the model never trained on (Figure 4A). The two curves almost coincide, and the synthetic curve lies slightly to the right, so synthetic patients are not closer to the training data than unseen real patients are. The nearest-neighbor distance ratio, which divides the distance to the nearest training patient by the distance to the second nearest, points the same way: values near 0 would mean one training patient stands out as a likely source, and the synthetic curve sits at or below the unseen real one (Figure 4B). The structure preserved in Tiers 1 to 4 is therefore consistent with learning rather than copying. This is a leakage check, not a privacy guarantee: it is not differential privacy and not a membership-inference attack, and it does not make the data safe to share.

Conclusion

Probing the primary-care foundation model through its synthetic trajectories shows that it learned the marginal structure of the data and the timing of visits, but only part of the finer clinical structure. Ordering of conditions before drugs and the strength of associations between concepts were weakened rather than lost, and a small share of synthetic records contained temporal contradictions that real records cannot. The synthetic data can therefore already support descriptive questions about frequencies and timing, while association studies on them would slightly underestimate real effects. Prediction models trained on them discriminated about as well as models trained on real data, but the task itself was hard and the predicted risks were poorly calibrated, so such models would need recalibration on real data. The leakage checks gave no sign that this utility came from copying training patients, although they are not a privacy guarantee. The answer to “fit for what?” is thus fit for description, usable with caution for association and prediction, and still unknown for causal inference. Tiers 5 and 6 are future work: we will test whether the confounding structure and the causal effect estimates of real data survive in the synthetic data.

References

1. Achterberg JL, Haas MR, Spruit MR. On the evaluation of synthetic longitudinal electronic health records. *BMC Medical Research Methodology*. 2024;24:181.
2. McDermott MBA, Nestor B, Kim EM, et al. A foundation model for electronic health records. *npj Digital Medicine*. 2023;6:96.
3. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*. 2024. arXiv:2311.11483
4. Mendes JM, Barbar A, Refaie M. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*. 2025;7:1563991. <https://doi.org/10.3389/fdgth.2025.1563991>
5. Yan C, Yan Y, Wan Z, et al. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*. 2022;13:7609. <https://doi.org/10.1038/s41467-022-35295-1>
6. Budu E, Soliman A, Etminani K, Rögnvaldsson T. A framework for evaluating synthetic electronic health records. *Studies in Health Technology and Informatics*. 2023;302:378–379.
7. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data

science tasks. CHANCE. 2019;32(1):42–49.

8. Maria A J de Ridder, Marcel de Wilde, Christina de Ben, Armando R Leyba, Bartholomeus M T Mosseveld, Katia M C Verhamme, Johan van der Lei, Peter R Rijnbeek, Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands, *International Journal of Epidemiology*, Volume 51, Issue 6, December 2022, Pages e314–e323, <https://doi.org/10.1093/ije/dyac026>
9. Pang C, Jiang X, Kalluri KS, et al. CEHR-GPT: generating electronic health records with chronological patient timelines. arXiv. 2023;2402.04400. <https://arxiv.org/abs/2402.04400>
10. Andrej Karpathy. A minimal pytorch re-implementation of the openai gpt (generative pretrained transformer) training, <https://github.com/karpathy/nanogpt>
11. Massey FJ. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951;46(253):68-78.
12. Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional GANs. arXiv:1706.02633. 2017.
13. Hancock, J. M., Zvelebil, M. J., & Hancock, J. M. (2014). "Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient)." In *Dictionary of Bioinformatics and Computational Biology*. <https://doi.org/10.1002/9780471650126.dob0956>