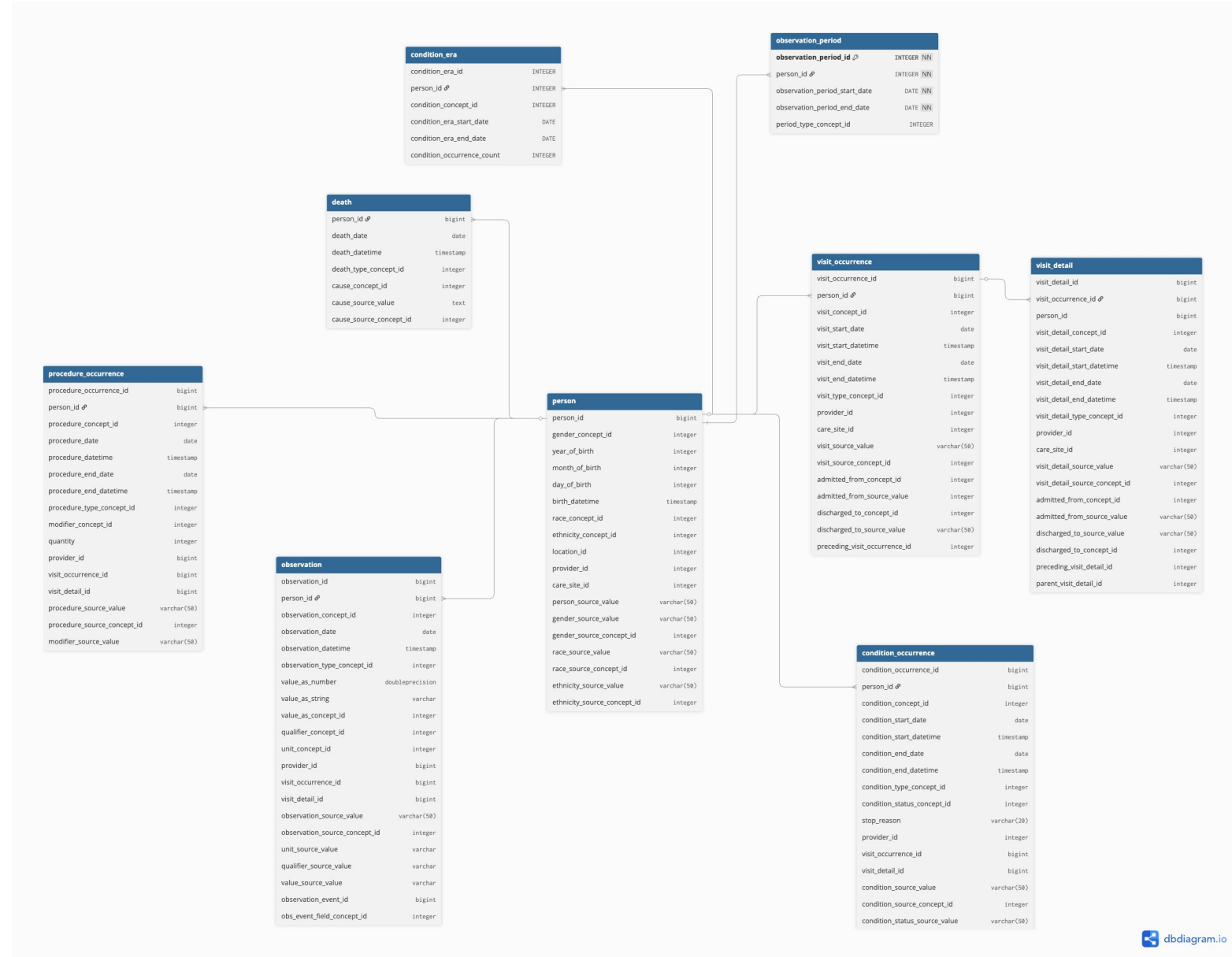


From Observational Data to Synthetic Cohorts

Alberto Labarga
MirrorHealth SL, Pamplona, Spain

The generation of high-quality synthetic health data has become a key enabler for collaborative research, method development, and education within the Observational Health Data Sciences and Informatics (OHDSI) community. Synthetic datasets can mitigate privacy risks, facilitate data sharing across institutions, and support reproducible research workflows without exposing sensitive patient-level information.

In this work, we present a novel synthetic data generation pipeline that leverages existing OMOP Common Data Model (CDM)² databases to automatically learn debiased graph representations of clinical events and their temporal and semantic relationships. These learned representations are then translated into executable Synthea¹ modules, enabling scalable and transparent generation of synthetic cohorts that preserve clinically meaningful structure while explicitly addressing bias and fairness concerns.



OMOP-CDM

Extract data into a graph

First, standardized OMOP-CDM tables (e.g., PERSON, CONDITION_OCCURRENCE, DRUG_EXPOSURE, VISIT_OCCURRENCE, etc) are transformed into an event-centric graph representation³. **Nodes** correspond to clinical concepts (conditions, drugs, procedures) and patient states, while **edges** encode observed co-occurrence, temporal ordering, and transition probabilities derived from the source database.

Summary

The proposed pipeline consists of five main stages: (1) representation learning from OMOP-CDM data, (2) debiased graph construction, and (3) Synthea module instantiation, (4) synthetic data generation (5) Synthea to OMOP conversion. The pipeline is designed to be modular, reproducible, and compatible with standard OHDSI tooling, enabling integration into existing observational research workflows. A fully open-source Python implementation orchestrates this process, generating executable Synthea modules and subsequently producing synthetic patient-level data that can be exported directly into OMOP-CDM-compliant datasets.

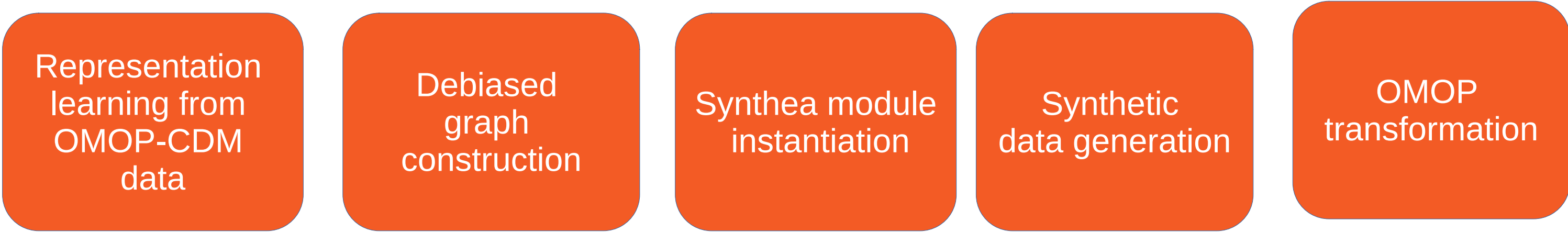
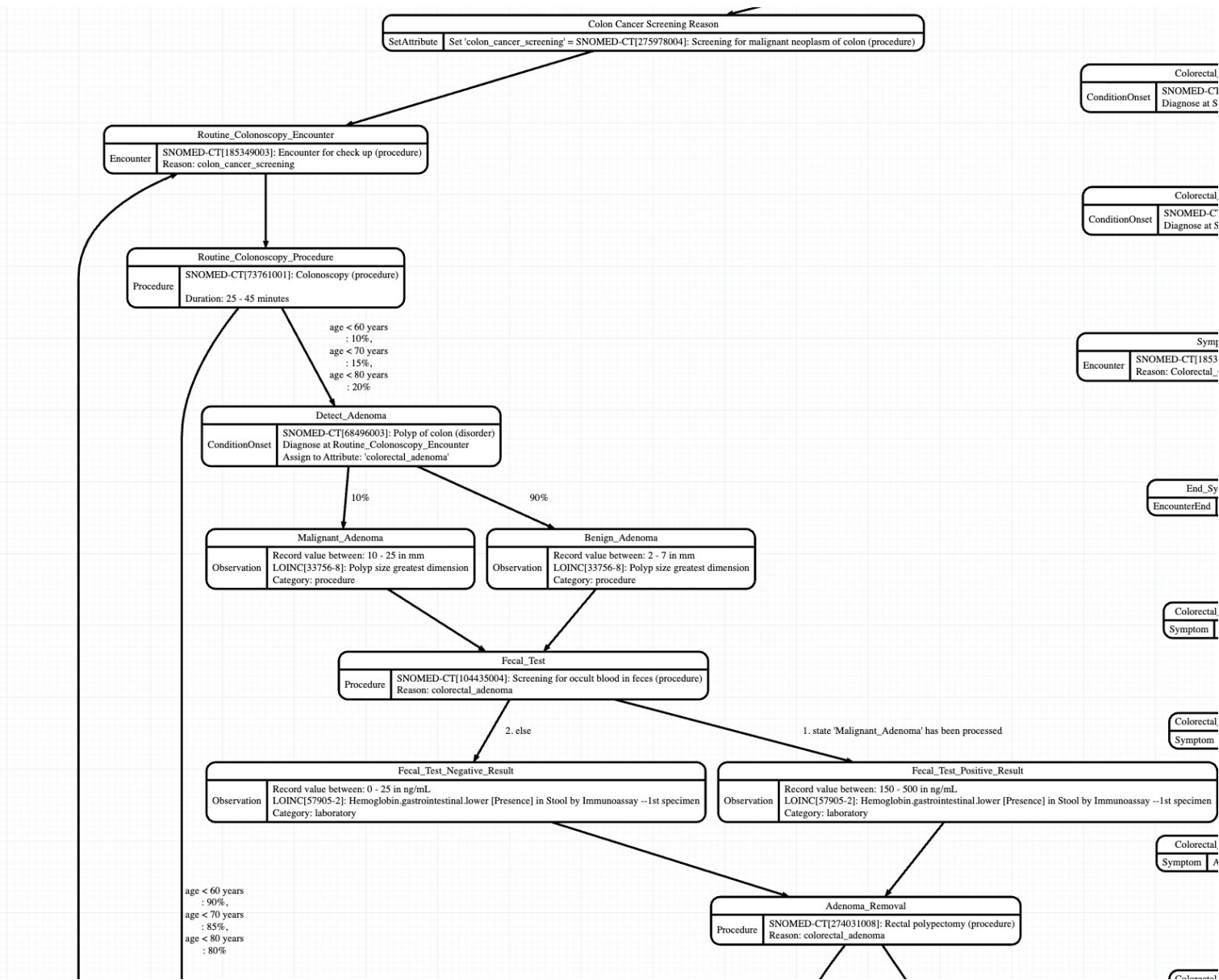


Figure 1. Synthetic data generation pipeline

Knowledge graph

Knowledge graph consolidation

Second, a debiasing step is applied to the learned graph to mitigate known sources of observational bias, including demographic imbalance, care-setting-specific practice patterns, and spurious correlations driven by data sparsity or coding artifacts. This is achieved through a combination of constraint-based normalization, re-weighting strategies, and graph pruning informed by domain-agnostic fairness criteria and configurable demographic stratifications.



Synthea module

Expose graph as Synthea module

Third, the resulting debiased graph is automatically translated into Synthea disease modules, preserving state transitions, conditional logic, and temporal dynamics.

Results

We evaluated the pipeline using several real-world OMOP-CDM databases covering diverse patient populations and care settings. The generated synthetic datasets were assessed along three dimensions: structural fidelity, statistical similarity, and bias attenuation.⁴

Structural evaluation demonstrated that the synthetic data preserved key OMOP-CDM characteristics, including visit structures, concept hierarchies, and longitudinal patient trajectories. Statistical comparisons showed that marginal and joint distributions of major clinical events closely approximated those of the source data, while avoiding direct record-level replication.

Importantly, bias-focused analyses indicated a measurable reduction in demographic and utilization imbalances when compared with naïve data-driven generation approaches. The resulting Synthea modules were human-readable, version-controlled, and reusable, supporting transparency and downstream validation by domain experts.

Conclusion

This work introduces a principled, data-driven approach for generating synthetic health data that bridges real-world observational datasets and rule-based simulation frameworks. By learning debiased graph representations from OMOP-CDM databases and translating them into executable Synthea modules, the proposed pipeline combines the realism of empirical data with the interpretability and controllability of simulation models.

References

[1] OHDSI. The Book of OHDSI. <https://ohdsi.github.io/TheBookOfOhdsi/>
[2] Walonoski J, et al. Synthea™: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association.
[3] Kang, Mengjia, et al. "Evolution of a graph model for the OMOP common data model." Applied Clinical Informatics 15.05 (2024)
[4] Kiran, A., P. Rubini, and S. Saravana Kumar. Comprehensive review of privacy, utility and fairness offered by synthetic data. IEEE Access (2025)

