

Developing a functional pipeline for generating synthetic OMOP CDM data

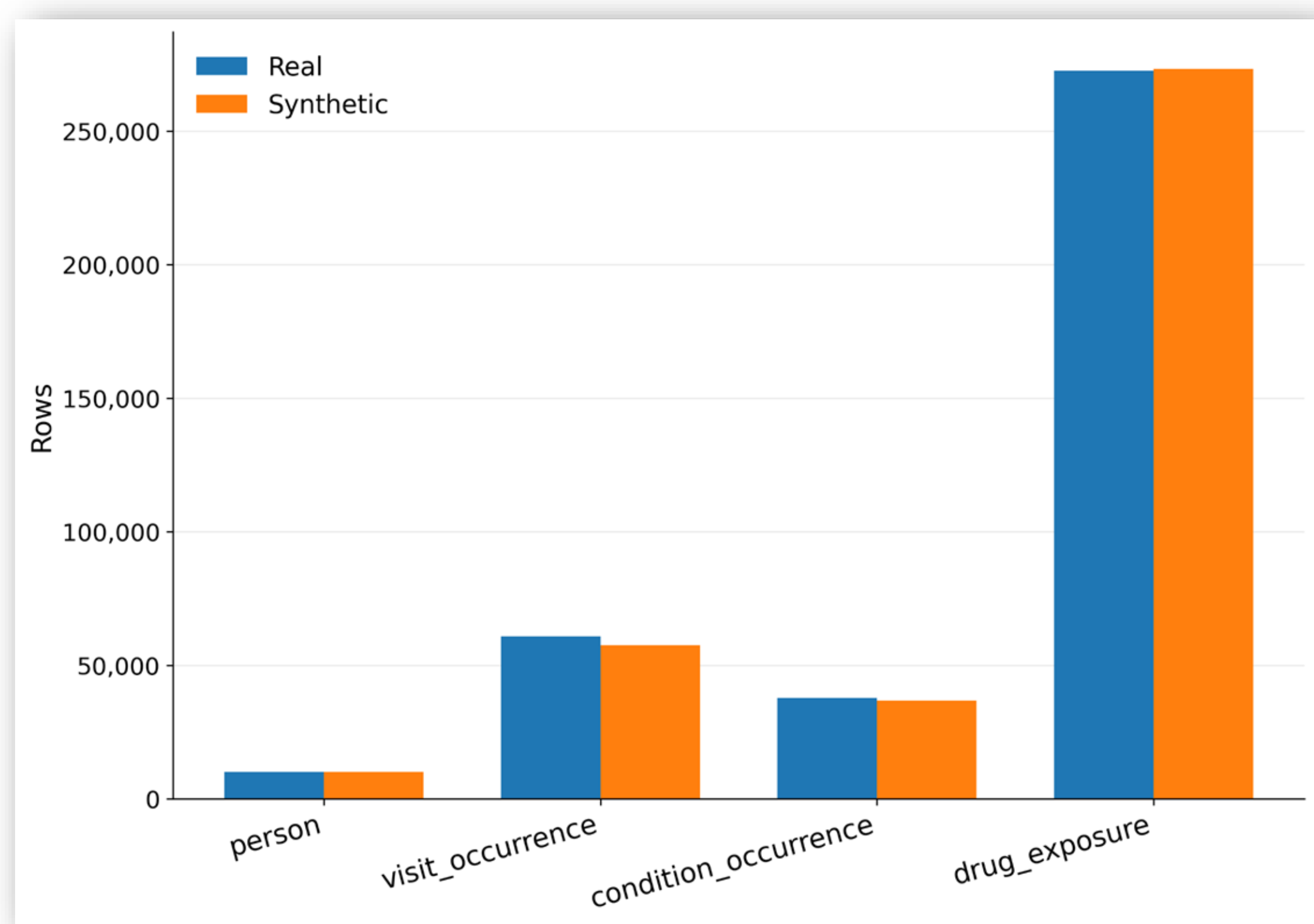
While structural validity and referential integrity were readily achieved, preserving **high-level clinical semantics logic** remains a **key challenge** for these models

Generating Clinically Coherent Synthetic Electronic Health Records in OMOP: Bridging the Semantic Gap

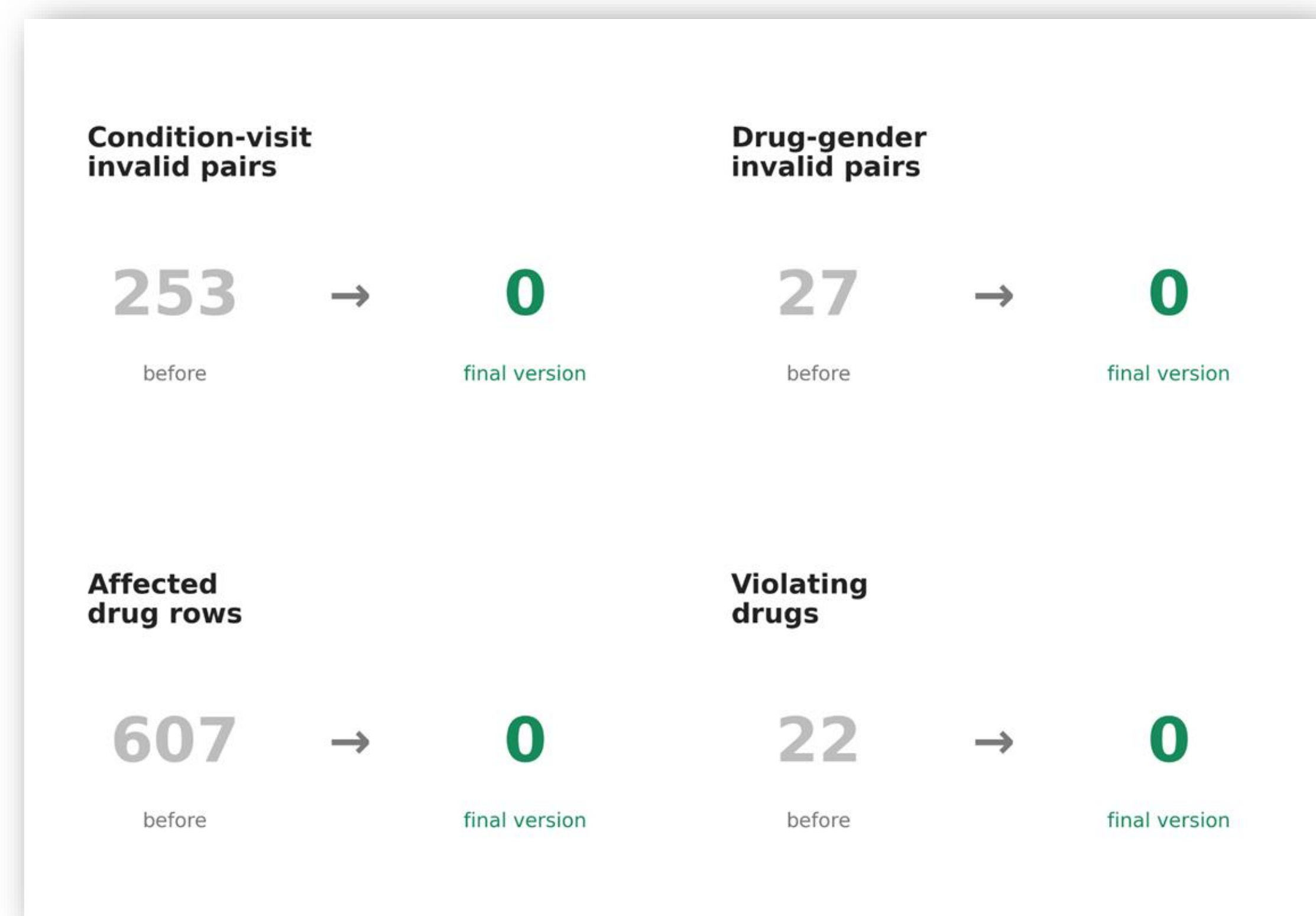
Background: Biomedical datasets are often constrained by stringent privacy requirements and limited accessibility. These factors hinder the widespread use of Electronic Health Records (EHRs) for observational research and methodological development. While Synthetic Data Generation (SDG) has emerged as a promising solution to share granular patient data while preserving privacy, producing synthetic clinical data of sufficient quality to serve as a reliable proxy for real patient records remains challenging. We aimed to build a usable synthetic OMOP dataset that preserves core clinical structure and coherence across four linked tables.

Results

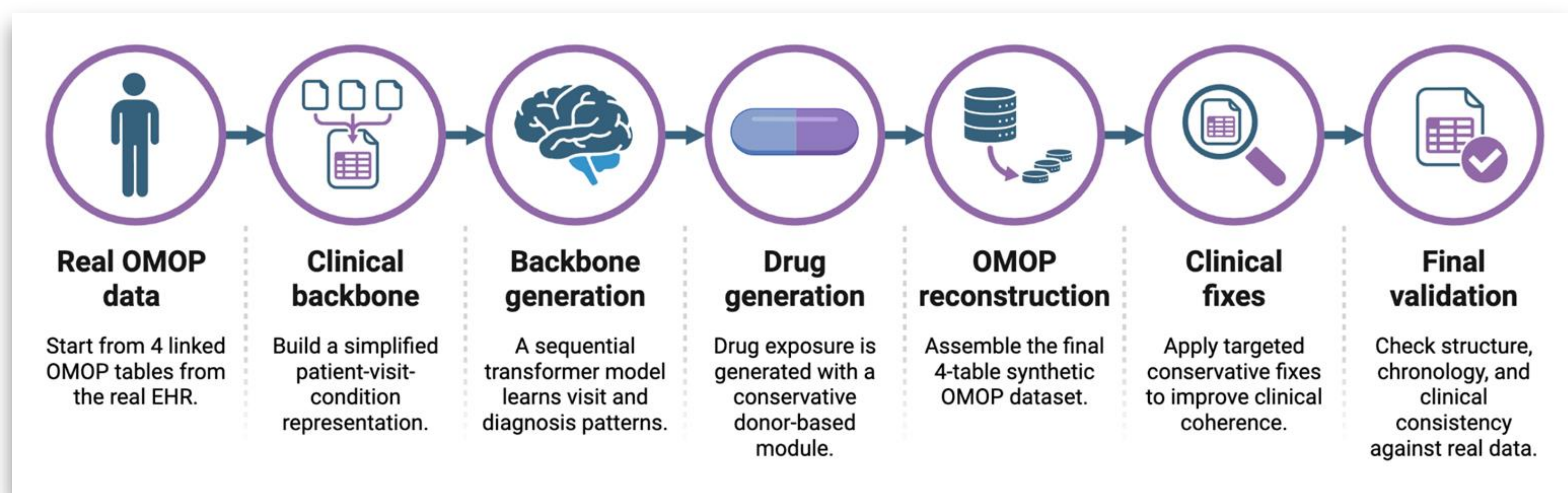
A full 4-table OMOP dataset at realistic scale
Synthetic row counts closely matched the original data across all 4 OMOP tables, supporting a complete and usable relational release rather than isolated synthetic records.



Targeted post-processing removed clinically invalid relationships
Conservative post-processing removed impossible condition-visit and drug-gender combinations, improving clinical coherence without changing the overall dataset structure.



Methods



We used a hybrid pipeline combining transformer-based backbone generation, conservative drug generation, OMOP reconstruction, and targeted clinical fixes.

Conclusions

Although the final 4-table synthetic OMOP dataset is usable and clinically safer than earlier candidates, concept-level calibration remains imperfect, and rare concepts are still under-represented. Future work will extend the pipeline to additional OMOP tables and improve long-tail clinical fidelity.



Jan Carreras, Juan Manuel Ramírez-Anguita, Ricard Garcia, Antonio Arcas, Carla Arxé, Angela Leis, Miguel-Angel Mayer

