Conditional Adversarial Random Forest for Synthetic Electronic Health Record Generation

Abstract

Synthetic Electronic Health Records (EHRs) enable privacy-preserving healthcare data sharing for machine learning research. However, existing methods struggle with: maintaining temporal consistency across patient visits while preserving demographic-clinical correlations. Current approaches either sacrifice temporal fidelity or require extensive postprocessing.

We propose Conditional Adversarial Random Forest (CARF), extending Adversarial Random Forest [1] with a two-model strategy. The first model generates patient-level demographics that remain static across visits. The second conditional model produces visit-level clinical variables, incorporating visit rank and time progression to create complete patient trajectories. This eliminates manual postprocessing while preserving temporal patterns inherently.

Key Contributions: (1) Two-step conditional generation model that preserved temporal patterns without postprocessing; (2) Comprehensive privacy evaluation revealing vulnerabilities in rare subpopulations; (3) Demonstrated reduction in demographic distance metrics while maintaining clinical validity.

Technical Approach: CARF learns joint distributions of demographics, visit patterns, and clinical variables through the forest structure. During generation, demographic features condition sampling through leaf-node filtering, ensuring temporal consistency. Age advances naturally with days-since-last-visit as visit rank increases, maintaining realistic patient progression without explicit sequence modeling.

Results: Evaluating on MIMIC-III [2] (46,000+ patients), CARF significantly outperforms baseline ARF approaches. For demographic consistency, weighted Euclidean distances improved dramatically—gender average distance dropped from 0.209 to 0.035, ethnicity from 0.327 to 0.151, and religion from 0.364 to 0.196. Medical code correlations improved with MSE reducing from 0.0125 to 0.0102. At patient level, CARF better preserved unique/total code distributions with standard deviations closer to real data. Temporal analysis showed superior Wasserstein distances for age across visits, particularly for early visits where data is more abundant.

Privacy Analysis: We developed comprehensive privacy evaluation using Distance to Closest Records (DCR) and two attack scenarios. Random selection attacks achieved 9.8% high-similarity matches at 5% DCR threshold, affecting 56 unique records. Consensus-based attacks proved more effective with 40.7% high-similarity matches, affecting 131 records. Rare demographic categories (separated marital status, self-pay insurance) showed heightened vulnerability in small equivalence classes.

Limitations and Future Work: Challenges persist with rare procedures (present in < 10 admissions) and age-inappropriate diagnoses. Future work includes developing age-specific models for neonates and incorporating expert-coded gender constraints.

This work advances synthetic healthcare data generation by unifying generation and evaluation, enabling broader access to medical datasets while quantifying and mitigating privacy risks for vulnerable populations.

References: [1] Watson et al. Adversarial random forests for density estimation. AISTATS 2023. [2] Johnson et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016.