International Journal of Population Data Science

Journal Website: www.ijpds.org

Privacy-Preserving Record Linkage in Population Data Systems: Leveraging Synthetic HDSS Data and Deterministic Identifier Masking for Machine Learning Models

Tathagata Bhattacharjee[1], Emma Slaymaker[1], Chodziwadziwa Kabudula[2], and Jim Todd[3]

[1]London School of Hygiene & Tropical Medicine, London, United Kingdom

[2]University of the Witwatersrand, Johannesburg, South Africa

[3]London School of Hygiene & Tropical Medicine, London, United Kingdom; Catholic University Of Health And Allied Sciences, Mwanza, Tanzania, United Republic of

Background

Linking population-based datasets requires identifiable attributes that raise privacy and governance concerns. In low- and middle-income countries (LMIC), Health and Demographic Surveillance Systems (HDSS) are key sources of longitudinal population data. Synthetic data offers a privacy-preserving route for developing and evaluating record linkage (RL) methodologies without exposing sensitive microdata. This study introduces a framework that leverages Conditional Tabular Generative Adversarial Network (CTGAN)-generated synthetic HDSS data and deterministic name masking to support ethical RL and the development of supervised and ensemble machine-learning linkage models.

Methods

A CTGAN-based pipeline was applied to clean and harmonise Kisesa HDSS data to generate a statistically comparable synthetic dataset, while sex-specific deterministic name masking ensured full replacement of identifiers. Synthetic data quality was evaluated using distributional similarity, relationship preservation, and privacy-risk assessment. Synthetic adult clinic datasets were derived from the synthetic HDSS by introducing graded, field-specific error patterns via controlled probabilistic perturbations, creating conditions to evaluate RL models. Supervised models (logistic regression, random forest, gradient boosting, support vector machines) and an ensemble meta-classifier were trained and compared on these synthetic linkage tasks.

Results

The synthetic HDSS dataset preserved key demographic structures and multivariate dependencies, while deterministic masking eliminated direct identifiers. Across increasingly noisy error scenarios, supervised and ensemble models maintained strong match–nonmatch discrimination, with ensemble methods exhibiting the highest robustness as data quality deteriorated.

Conclusion

CTGAN-derived synthetic HDSS data combined with deterministic name masking provides a secure testbed for machine-learning-based RL, enabling ethical, reproducible, and governance-aligned population data linkage research in LMICs.

https://doi.org/10.23889/ijpds.v11i5.3673