Synthetic Medical Data

Vineeth Vellala, 201719460, MSc Data Science and Analytics

Leeds University, 2023/24

Background

- Electronic health records (EHRs) contain valuable information for research, algorithm development, and educational purposes, but access is limited due to privacy concerns
- Synthetic medical data can mitigate privacy issues, especially in resource-constrained settings
- **Hypothesis:** Normalising flow models can learn MIMIC-IV dataset's probability density to generate useful synthetic data

Materials & Methods

- Utilise MIMIC-IV, a large, freely available database of de-identified electronic health records, as the primary data source
- Develop and compare normalising flows (nflows library) and Generative Adversarial Networks (GANs) for generating realistic synthetic EHR data to integrate into OpenMRS

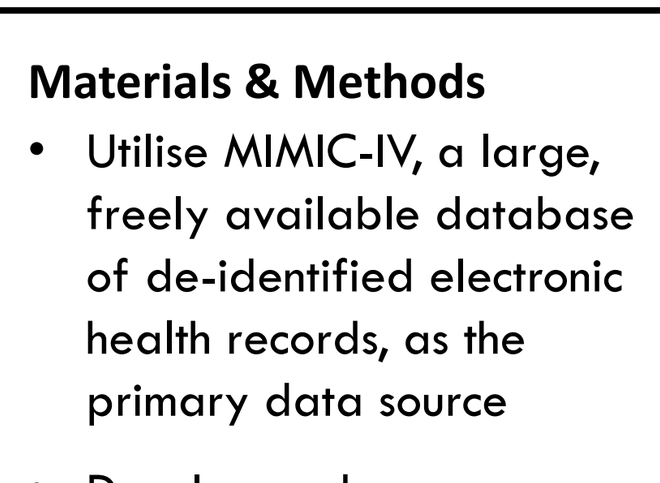


Figure 1: Architecture

Results/stage of the work

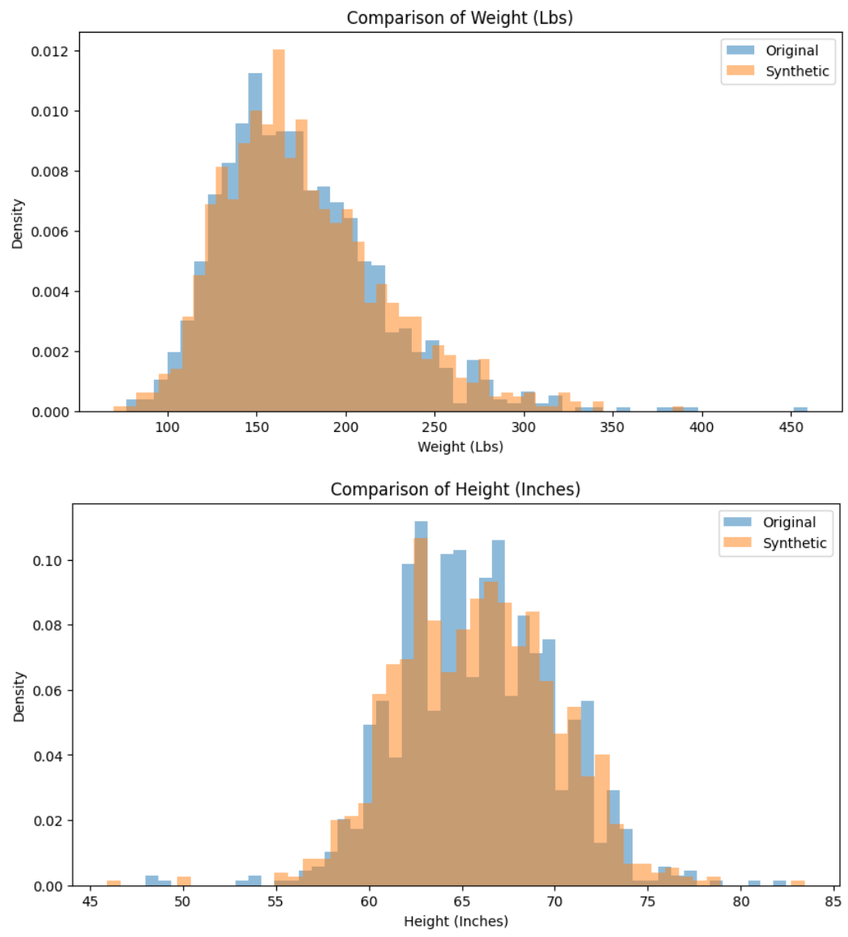
The normalising flow model successfully generated synthetic medical data closely resembling the original MIMIC-IV dataset for patient height and weight.

1. **Statistical measures:** Synthetic data closely matched original data for both height and weight

Feature	Original Mean	Synthetic Mean	Original Std	Synthetic Std
Height (Inches)	65.79	65.87	4.14	4.18
Weight (Lbs)	174.58	179.45	46.05	46.03

Table 1: comparison mean and standard deviation

2. **Distribution analysis:** Density plots show similar patterns between original and synthetic data for both features



Figures 2 and 3: Comparison of distributions

3. **Statistical tests:** Kolmogorov-Smirnov tests indicate no significant difference in height ($p=0.0341$) and weight ($p=0.025$) for synthetic data

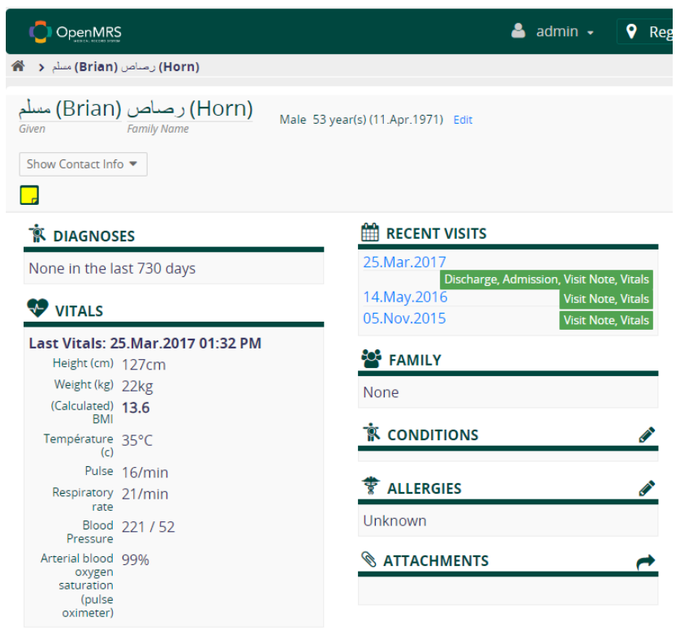
Feature	KS Statistic	KS p-value
Height (Inches)	0.042	0.0341125
Weight (Lbs)	0.066	0.025634

Table 2: K-S Statistical test result

Aims to Achieve

1. Provide a privacy-preserving solution for medical research and education, reducing reliance on real patient data by 75% while maintaining analytical utility
2. Generate realistic patient data that achieves a 95% match rate with key health indicators from original EHR data
3. Enable third-world countries to access high-quality synthetic medical data through integration with OpenMRS
4. Ensure 100% GDPR compliance in synthetic data generation, safeguarding patient privacy

Synthetic Data in Action: OpenMRS Integration



References

1. Rezende, Danilo, and Shakir Mohamed. "Variational inference with normalizing flows." International conference on machine learning. PMLR, 2015.
2. Seebregts, Christopher J., et al. "The OpenMRS implementers network." International journal of medical informatics 78.11 (2009): 711-720.

Scan for Further Information

SCAN ME

Acknowledgments

Thank you for your continuous support
Owen Johnson, Senior Teaching Fellow, UoL
Dr Luisa Cutillo, Associate Professor, UoL