SYNTHESIZING ELECTRONIC HEALTH RECORDS FOR PREDICTIVE MODELS IN LMICS

Ghadeer O. Ghosheh and Tingting Zhu
Department of Engineering Sciences, University of Oxford
{ghadeer.ghosheh, tingting.zhu}@eng.ox.ac.uk

ABSTRACT

The spread of machine learning models, complemented by the increased adoption of electronic health records (EHRs) has opened the door for developing clinical decision support systems. However, despite the great promise of machine learning for healthcare in low-middle-income countries (LMICs), many data-specific limitations, such as the small size and irregular sampling, hinder the progress in such applications. Recently, deep generative models have been proposed to generate realistic-looking synthetic data, including EHRs, by learning the underlying data distribution without compromising patient privacy. In this study, we first use a deep generative model to generate synthetic data based on a small dataset (364 patients) from Ho Chi Minh City Hospital for Tropical Diseases, Vietnam. Next, we use synthetic data to build models that predict the onset of hospital-acquired infections (HAIs) based on minimal information collected at patient ICU admission. The performance of the diagnostic model trained on the synthetic data outperformed models trained on the original and oversampled data using techniques such as SMOTE. We also experiment with varying the size of the synthetic data and observed the impact on the models' performance. Our results show the promise of using deep generative models in enabling healthcare data owners to develop and validate models that serve their needs and applications, despite limitations in dataset size.

INTRODUCTION

Developing clinical decision support systems applications using electronic health records (EHRs) and machine learning (ML) techniques has gained increased interest from the research community (Xiao et al., 2018). Despite the promising results of many of these applications, the performance of ML models is highly dependent on the availability of training data (Jeni et al., 2013; van der Ploeg et al., 2014). ML models tend to be data hungry, where models could easily overfit and under-perform when trained on a small dataset (Jeni et al., 2013; van der Ploeg et al., 2014). The dependence on data hurdles the optimal development and utilization of clinical decision support systems, specifically in resource-constrained clinical settings. Many healthcare facilities in Low Middle-Income Countries (LMICs) suffer from limitations, such as high patient-doctor ratio, and financial and infrastructural constraints all of which impact the scale of research that could serve such populations.

While most of the world's population resides in LMICs, however, the patient-to-doctor ratio is lower than that of high-income countries by 10 folds (Schneider et al., 2004) resulting in an increased burden on the clinical staff. Moreover, healthcare systems in LMIC countries often experience infrastructural constraints, diagnostic capacity (Galindo-Fraga et al., 2018), frequent changes in strategic healthcare policies, and political instability (Mills, 2014), all of which could impact the quantity and quality of healthcare data collected from such clinical settings. These limitations make collecting large-scale data infeasible as it adds extra costs and burdens. Notwithstanding the relevance of ML healthcare work conducted in more advanced healthcare settings, many diseases and applications are specific and more prevalent in low-resource settings, resulting in an unmet need for machine learning research applications that are developed and validated for low-resource settings.

Many of the current medical statistics and data-driven models rely on methods such as SMOTE which oversample the training data, especially in imbalanced settings. Oversampling methods could introduce flawed correlations and dependencies between samples and result in limited data variability (Fernández et al., 2018), all of which could severely underperform in testing environments. Recent works in deep learning research proposed generative models that learn the underlying data distribution and generate realistic-looking data while preserving the privacy of the original samples. Those deep generative models, including Generative Adversarial Networks (GANs) and Variational AutoEncoders (VAEs) (Kingma & Welling, 2013; Goodfellow et al., 2014) have been originally proposed and validated for the imaging domain. Despite being very relevant and highly needed, using deep generative models for synthesizing EHRs for low-resource clinical applications is often not discussed nor motivated in most proposed works (Ghosheh et al., 2022).

To this end, this paper proposes synthetic data as a solution for developing models based on small datasets collected from LMIC countries. To do so, we train a GAN-based model to learn the underlying data distribution and generate synthetic samples that could be utilized for training purposes. Specifically, we utilize a small dataset (364 patients) collected from an Intensive Care Unit in Vietnam (Thuy et al., 2018), with variables collected at admission and a binary outcome indicating if the patient got a hospital-acquired infection. With the increased burden of antimicrobial resistance, especially in LMICs, it's vital to develop risk scores to predict the probability of developing such infections. This could allow the clinical staff to take anti-septic measures, reduce unnecessary antibiotics prescriptions and introduce timely interventions to prevent prolonged lengths of stays.

Our contributions could be summarized as follows. For the first time, we demonstrate the feasibility of using generative models for synthesizing data that is used to develop ML models from small datasets from LMIC healthcare settings. Furthermore, we evaluate the utility of the synthetic data in comparison to other commonly used approaches such as data oversampling and showcase the impact of synthetic data size on the performance of the predictive model in a series of experiments where the synthetic data training size is varied. The proposed method provides a plausible solution that could be used for developing diagnostic models despite data scarcity in LMICs.

METHODS

DATASET

The data used in this work is collected from Ho Chi Minh City Hospital for Tropical Diseases, Vietnam, and released for open access (Thuy et al., 2018). The patients included in this study are 364 a total of patients who were all admitted to the ICU and stayed at least two days. The included variables are those readily available at the admission of ICU, which we categorize into comorbidities, demographics and admitting diagnosis. The outcome of interest is a binary label indicating if the patient acquired an infection during their stay. We describe the statistical distribution of our dataset in terms of outcomes and included features in Table 1.

SYNTHETIC DATA GENERATION

To evaluate the feasibility of using synthetic data as a training set, we apply a random train-test split for our data to obtain separate training and testing sets, with a split of 70% for training and 30% for the held-out set used for testing the performance of the model. The training set is used to train the GAN model for tabular data, namely medGAN (Choi et al., 2016). Upon training the GAN model the size of synthetic data is determined at inference time.

TASK AND BASELINES

The generated synthetic data is used to train a simple random forest model to predict hospital-acquired infections during the patient's ICU stay. The choice of a simple model is motivated by its relative simplicity, with often comparable performance to many advanced models, making it a good candidate for deployment in hospitals in LMICs. We compare the performance of the model trained on the synthetic data to those trained on the (1) original small training set and (2) oversampled training data using SMOTE. To better understand the impact of the synthetic data size on the predictive model performance, we train the GAN model to synthesize data of various sizes at inference.

Table 1: List of included patient features in terms of count and percentage prevalence in the population

Comorbidities (n, %)	
Diabetes	35 (9.62%)
Steroids	15 (4.12%)
Chronic Liver	55 (15.11%)
Chronic Kidney	3 (0.82%)
Demographics (n, %)	
Female	242 (66.48%)
Age	
16-45	133 (36.54%)
45-60	142 (39.01%)
65+	89 (24.45%)
Admission Diagnosis (n, %)	
Tetanus	17 (4.67%)
Sepsis	45 (12.36%)
Local Infections	75 (20.60%)
Dengue	204 (56.04%)
Internal Medicine	139 (6.32%)
Outcomes (n, %)	
Hospital Acquired Infections	86 (23.6%)

The synthesized data is then used to train the predictive model, where the performance is compared to that of models trained with original and oversampled data. The final performance is reported on the held-out test set in terms of Area Under the Receiver Operating Characteristic Curve (AUROC) (Hajian-Tilaki, 2013), The area under the Precision-Recall Curve (AUPRC) (Ozenne et al., 2015), and balanced accuracy with confidence intervals computed using bootstrapping with 1,000 iterations. An overview of the predictive modelling and evaluation of our approach is presented in Figure 1.

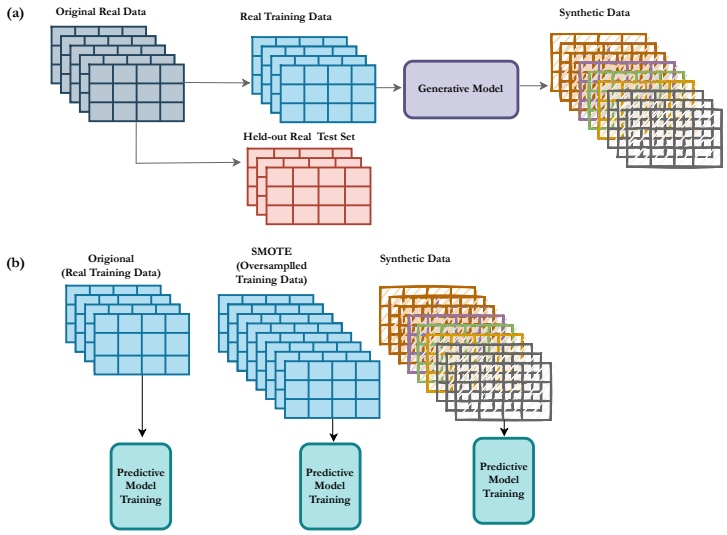


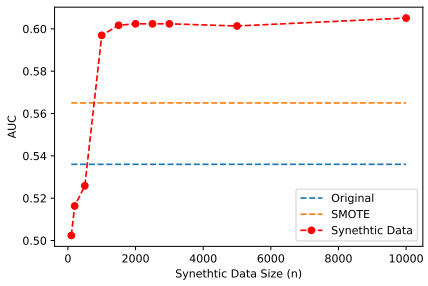
Figure 1: Overview of the proposed model trained on the synthetic data. (a) the dataset is split into training and a held-out test set. The training set is used to train the deep generative model which generates synthetic data. (b) a logistic regression model is trained in three different setups, (1) original, (2) SMOTE, (3) synthetic data which are evaluated on the held-out test set and compared in terms of the performance metrics

RESULTS

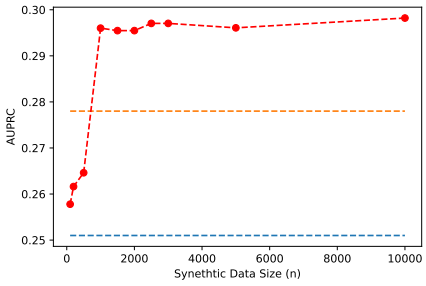
In Table 2, we present the results of the models trained on the original training data, the oversampled data, and synthetic data of various sizes. The original model achieved a performance of 0.536 in AUROC, compared to 0.565 for the SMOTE baseline. The models trained on synthetic data outperformed the other baselines, with a performance of 0.605 in terms of AUROC and 0.298 in terms of AURPC when using 10000 synthetic samples. The model trained on the original data and SMOTE are first outperformed by the model trained with 1000 synthetic samples in terms of AUROC and AUPRC, where it also achieved the highest balanced accuracy of 0.571. We notice that performance gains after increasing the synthetic data size from 1,500 to 10,000 are minimal, where the balanced accuracy did not change, with minor changes observed in AUROC and AUPRC scores. The results are also visualized in Figure 2.

Table 2: Results of the predictive model using the various baselines for training data. The results are reported in terms of AUROC, AUPRC and Balanced Accuracy.

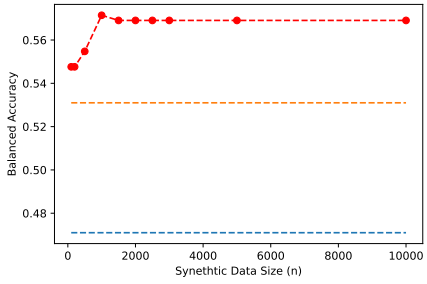
Model	AUROC	AURPC	Balanced Accuracy
Original	0.536 (0.393, 0.679)	0.251 (0.160, 0.389)	0.471 (0.442, 0.493)
SMOTE	0.565 (0.425, 0.704)	0.278 (0.168, 0.443)	0.538 (0.425, 0.642)
Synthetic 100	0.502 (0.352, 0.667)	0.258 (0.154, 0.424)	0.548 (0.455, 0.662)
Synthetic 200	0.516 (0.379, 0.665)	0.262 (0.159, 0.439)	0.548 (0.449, 0.659)
Synthetic 500	0.526 (0.389, 0.672)	0.265 (0.165, 0.435)	0.555 (0.461, 0.654)
Synthetic 1000	0.597 (0.471, 0.735)	0.296 (0.190, 0.480)	**0.571 (0.464, 0.674)**
Synthetic 1500	0.602 (0.465, 0.74)	0.295 (0.191, 0.478)	0.569 (0.476, 0.668)
Synthetic 2000	0.602 (0.465, 0.725)	0.295 (0.188, 0.475)	0.569 (0.473, 0.667)
Synthetic 2500	0.602 (0.467, 0.731)	0.297 (0.184, 0.488)	0.569 (0.477, 0.670)
Synthetic 10000	**0.605 (0.473, 0.737)**	**0.298 (0.185, 0.472)**	0.569 (0.476, 0.668)



(a) AUC



(b) AUPRC



(c) Balanced Accuracy

Figure 2: Performance of the predictive model trained using synthetic data of various sizes, SMOTE, and original training set

DISCUSSION AND CONCLUSION

Despite the increased research interest in using deep generative models, there exists a gap in identifying the opportunities and limitations such models have in machine learning applications for low-source settings. To the best of our knowledge, this work is the first to investigate the use of deep generative models for generating EHRs from LMICs, where the datasets often come with small sizes and feature sets. Another contribution of this work is demonstrating the impact of the size of the generated data on the performance of the predictive model, which we believe is an understudied area of research. While the results of the predictive model could be improved by using neural networks, we prefer using a simpler model to simulate a close setup to the target application setting in resource-constrained settings. Furthermore, the model used to generate the synthetic EHRs is medGAN (Choi et al., 2017) which is one of the simpler and earlier works of GANs for EHRs. We believe that results could be improved by using conditional variants of GANs (Li et al., 2021), and those with more stable loss functions (Baowaly et al., 2019).

The promising results of using synthetic data for training purposes will open the door for new research directions in building ML models for LMIC despite data scarcity, which could pave the way for new research and clinical decision support systems that best fit LMIC settings. Furthermore, in the absence of protection guidelines and regulations such as HIPAA (for Disease Control et al., 2003) and GDPR (Voigt & Von dem Bussche, 2017) that are specific to low-resource settings, we believe that using deep generative models could encourage data owners in low-resource settings to share synthetic data for international research without compromising the privacy of patients coming from low-resource settings.

REFERENCES

Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.

Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*, 2016.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pp. 286–305. PMLR, 2017.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

Centers for Disease Control, Prevention, et al. Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services. *MMWR: Morbidity and mortality weekly report*, 52(Suppl 1):1–17, 2003.

Arturo Galindo-Fraga, Marco Villanueva-Reza, and Eric Ochoa-Hein. Current challenges in antibiotic stewardship in low-and middle-income countries. *Current Treatment Options in Infectious Diseases*, 10(3):421–429, 2018.

Ghadeer Ghosheh, Jin Li, and Tingting Zhu. A review of generative adversarial networks for electronic health records: applications, evaluation measures and data sources. *arXiv preprint arXiv:2203.07018*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014.

Karimollah Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pp. 245–251. IEEE, 2013.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Li, Benjamin J Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *arXiv preprint arXiv:2112.12047*, 2021.

Anne Mills. Health care systems in low-and middle-income countries. *New England Journal of Medicine*, 370(6):552–557, 2014.

Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, 68(8):855–859, 2015.

Helen Schneider, Duane Blaauw, Lucy Gilson, Nzapfurundi Chabikuli, and Jane Goudge. Health systems strengthening and art scaling up: Challenges and opportunities'. *Centre for Health Policy, School of Public Health, University of Witwatersrand, Johannesburg, December*, 2004.

Duong Bich Thuy, James Campbell, Le Thanh Hoang Nhat, Nguyen Van Minh Hoang, Nguyen Van Hao, Stephen Baker, Ronald B Geskus, Guy E Thwaites, Nguyen Van Vinh Chau, and C Louise Thwaites. Hospital-acquired colonization and infections in a vietnamese intensive care unit. *PLoS One*, 13(9):e0203600, 2018.

Tjeerd van der Ploeg, Peter C Austin, and Ewout W Steyerberg. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC medical research methodology*, 14(1):1–13, 2014.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10:3152676, 2017.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.