FEEC UNICAMP

Presentation 1 IA376N 2s2025

Generation and evaluation of privacy preserving synthetic health data

Henrique Parede de Souza - 260497

Overview

- Authors
- Problem
- Background & Motivation
- MedGAN
- HealthGAN
- Methodology
- Results
- Discussion & Conclusion

Neurocomputing 416 (2020) 244–255

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Generation and evaluation of privacy preserving synthetic health data

Andrew Yale[a,*], Saloni Dash[c], Ritik Dutta[d], Isabelle Guyon[b], Adrien Pavao[b], Kristin P. Bennett[a]

[a] Rensselaer Polytechnic Institute, Troy, New York, USA
[b] UPSud/INRIA Université Paris-Saclay, France
[c] BITS Pilani, Department of CSIS, Goa Campus, India
[d] IIT Gandhinagar, India

ARTICLE INFO

Article history:
Received 16 July 2019
Revised 16 December 2019
Accepted 16 December 2019
Available online 10 April 2020

Keywords:
Synthetic data
Health data
Generative adversarial networks
Privacy

ABSTRACT

We develop metrics for measuring the quality of synthetic health data for both education and research. We use novel and existing metrics to capture a synthetic dataset's resemblance, privacy, utility and footprint. Using these metrics, we develop an end-to-end workflow based on our generative adversarial network (GAN) method, HealthGAN, that creates privacy preserving synthetic health data. Our workflow meets privacy specifications of our data partner: (1) the HealthGAN is trained inside a secure environment; (2) the HealthGAN model is used outside of the secure environment by external users to generate synthetic data. This second step facilitates data handling for external users by avoiding de-identification, which may require special user training, be costly, or cause loss of data fidelity. This workflow is compared against five other baseline methods. While maintaining resemblance and utility comparable to other methods, HealthGAN provides the best privacy and footprint. We present two case studies in which our methodology was put to work in the classroom and research settings. We evaluate utility in the classroom through a data analysis challenge given to students and in research by replicating three different medical papers with synthetic data. Data, code, and the challenge that we organized for educational purposes are available.

1. Introduction

Teaching data analysis and doing research with actual patient level medical data such as electronic healthcare records (EHR) are greatly restrained by laws protecting patients' privacy, such as the Health Insurance Portability and Accountability Act (HIPAA) [1,2] in the United States and the General Data Protection Regulation (GDPR) [3] in the European Union. While beneficial, these laws severely limit access to patient level medical data thus stagnating innovation and limiting educational and research opportunities. The process of obfuscation of medical data is costly and time consuming with high penalties for accidental release. Research and education using EHR are highly skewed to a few shareable datasets such as MIMIC-III (Medical Information Mart for Intensive Care) [4], which consists of de-identified ICU (intensive care unit) longitudinal data from 2001 to 2012 that adheres to the HIPAA restrictions and therefore can be shared. The only requirement is that the user completes a "Data or Specimens Only Research" certification. Datasets like MIMIC protect patients' privacy with classical anonymization techniques consisting of removing or regrouping quasi-identifiers in higher level categories (such as broad geographical areas) and removing or obfuscating sensitive information. Hence data utility can be severely altered. While the MIMIC data is extremely useful and has generated many research papers, it is limited to ICU data. It does not give access to the entire medical history of patients, hence limiting the type of analyses that can be carried out. This paper addresses this problem by proposing to use generators of synthetic data. The balance that needs to be hit in this project is to create synthetic data with enough quality to be useful for teaching purposes and ideally even for research, while preserving the privacy of the real data. In order to be useful in an education setting, synthetic data must preserve the relationships that exist in real patient-level data, so that assignments and projects using or discovering these relationships can be taught to students with the privacy preserving synthetic data. Other synthetic data generators like Synthea [5] pursue a similar goal but are based on publicly available summary statistic data, and therefore do not provide the flexibility of creating generative models faithfully resembling real data.

* Corresponding author.
E-mail address: yalea@rpi.edu (A. Yale).

https://doi.org/10.1016/j.neucom.2019.12.136
0925-2312/

Authors

- Research primarily conducted at INRIA Université Paris-Saclay by Dr. Isabelle Guyon (4th).
- Collaboration with Rensselaer Polytechnic Institute (US), Birla Institute of Technology And Science and Indian Institute Of Technology Gandhinagar (India).

Main areas of study: Biometrics, Genomics, Proteomics, Cancer research, Healthcare, Pattern Classification

Problem

Health data is critical for research and education, but **highly sensitive**.

Electronic healthcare records (EHR)

- Greatly restrained by laws protecting patients' privacy (HIPAA in US and GDPR in Europe).
- Obfuscation of medical data is costly and time consuming with high penalties for accidental release.

Potential Solution

Usage of **synthetic EHR data** generators.

Goal

- Create synthetic data with enough quality to be useful for teaching and research, while preserving the privacy of the real data.

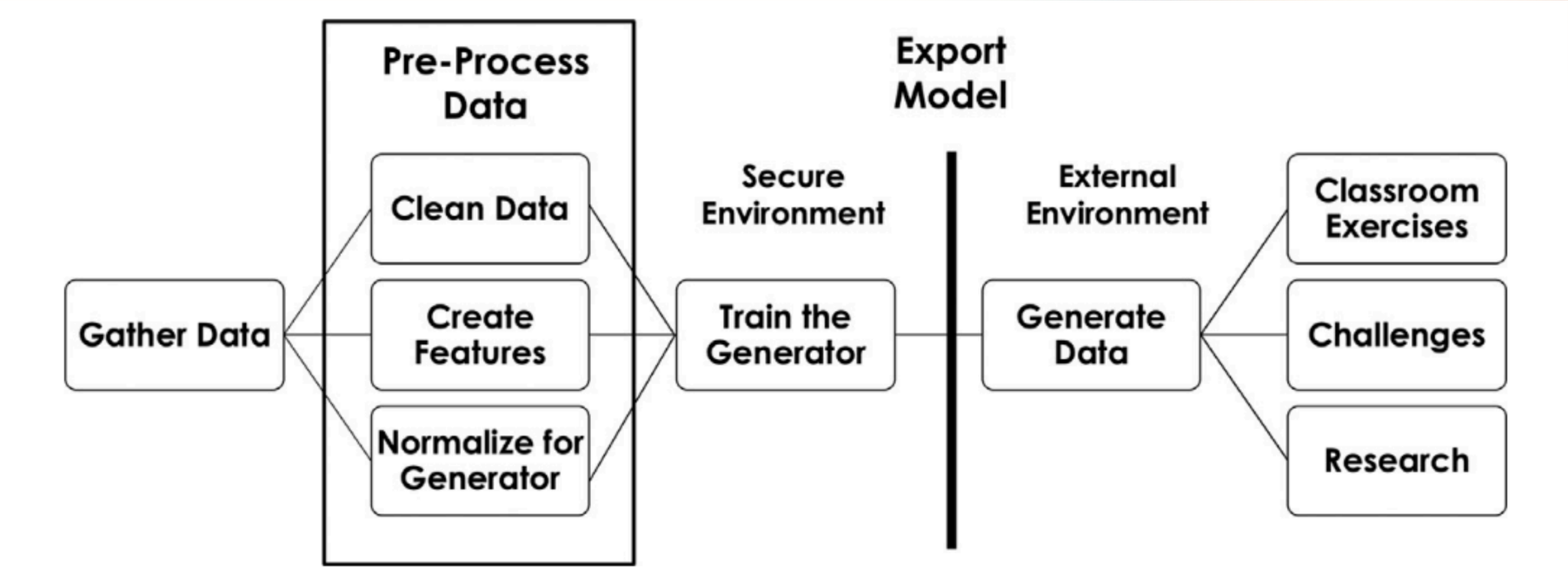
Synthetic data must preserve relationships

- Garantes the quality and the usability in educational and research settings.

Proposal

- Workflow for training a generative model, using real data in a secure sand-boxed environment, exporting the model to the outside, and then synthesizing data.
- HealthGAN for privacy-preserving synthetic health data (based on medGAN).

Workflow



1) The data is processed and used to train the generator model **inside the secure environment.**

2) Model is exported to an **external environment**, allowing its usage for multiple types of applications.

Background & Motivation

Traditional anonymization techniques have limitations, including **risk of re-identification** and **loss of data utility.**

Usage of GANs

- MedGAN (Choi et al. 2017) had already proposed the usage of Adversarial Networks for realistic yet private synthetic data generation.

Need for method balancing metrics

- **Resemblance**: data generated are sufficiently close to the real data.
- **Privacy**: data generated are significantly different from training samples.
- **Utility**: data generated preserves some utility (for research and education purposes).
- **Footprint**: model may not require real data to generate data and should not be on the order of the real data in size.

MedGAN Limitations

- **Binary-only data:** cannot handle continuous and categorical features.
- **Column-wise resemblance only:** matches marginal probabilities but fails to capture patient-level patterns.
- **Row-sum distortion:** generates patients with unrealistic numbers of diagnoses, not seen in real data.
- **Spurious comorbidities: i**ntroduces medically implausible combinations (e.g., male + female-specific diagnoses).
- **Overfitting risks:** optimized for one dataset, not robust across varied health data.

HealthGAN Implementation

GAN framework tailored for **categorical and numerical** health data.

Architecture

- MedGAN combined with Wasserstein GAN gradient penalty (WGAN-GP).
- Similar to the original GAN with a generator network with three layers, and a discriminator network with four layers.

Training Setup

- Real health datasets used to train generator and discriminator.
- A large batch size is used to ensure that outliers and rare values are captured in each batch and therefore learned by the generator

Evaluation Methodology

Comparison with baselines

- Gaussian multivariate.
- Parzen Windows.
- Differential Privacy.
- Aditive Noise Model.
- Copy of real data.

Evaluation Metrics

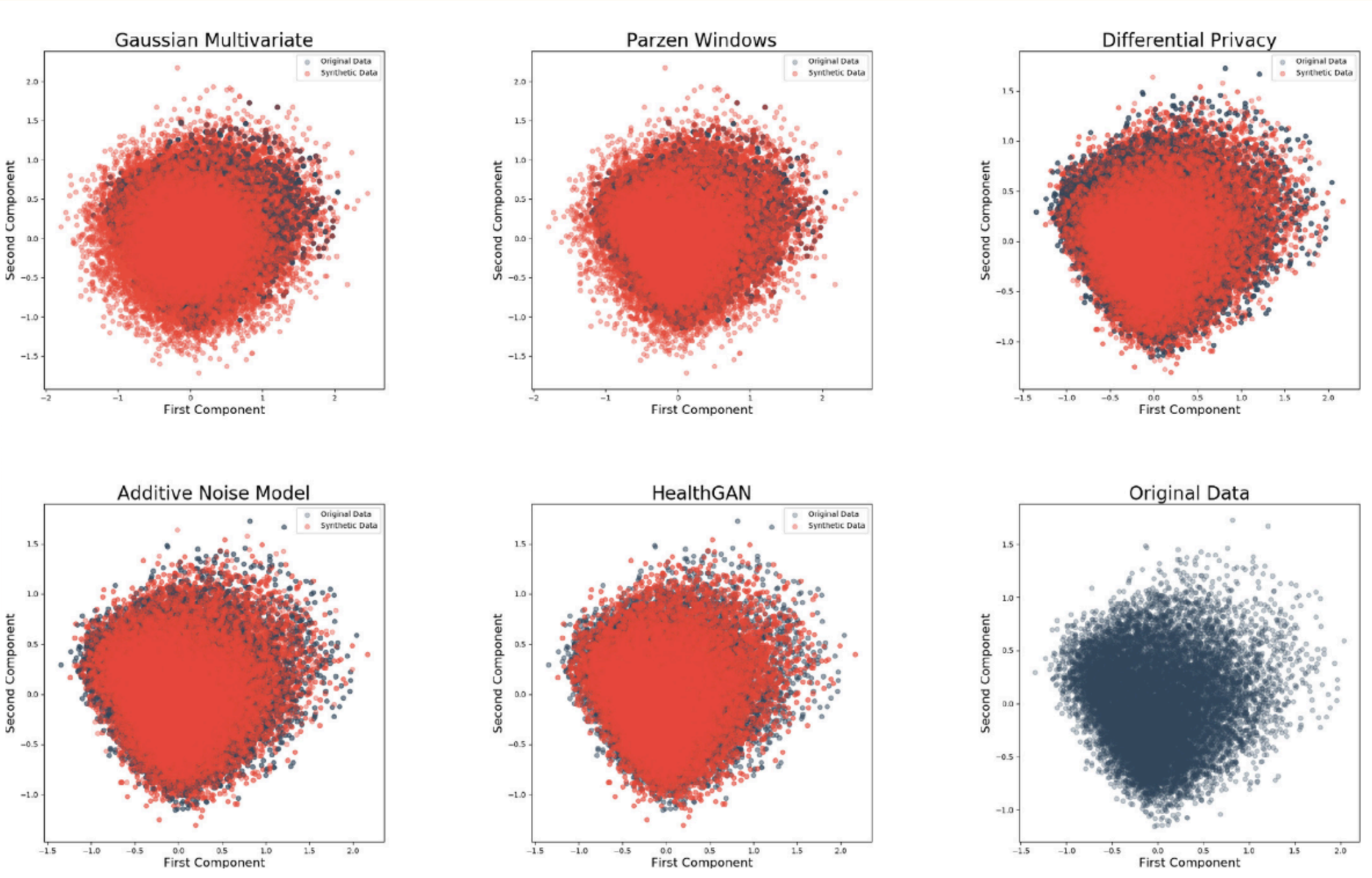
- Resemblance x privacy: Nearest Neighbor Adversarial Accuracy (AA).
- Utility: performance of predictive models trained on synthetic vs real data.
- Privacy footprint: risk of re-identification and membership inference attacks.
- PCA dimension reduction.
- Custom Losses.

Custom Losses

$$\textbf{TrResemblLoss}\ (Train\ Adversarial\ Acc.) = E[\mathcal{AA}_{RtrA_1}]$$
$$\textbf{TeResemblLoss}\ (Test\ Adversarial\ Acc.) = E[\mathcal{AA}_{RteA_2}]$$
$$\textbf{PrivacyLoss}\ (Test\ \mathcal{AA}\ -\ Train\ \mathcal{AA}) = E[\mathcal{AA}_{RteA_2} - \mathcal{AA}_{RtrA_1}]$$

Results

PCA analysis



Quick understanding of resemblance of the real test data.

Results

AA Analysis

Closer to 0.5 is better.

Table 1
Comparison of models with respect to various metrics. Blue: Excellent; Yellow: Good; Orange: Poor. Our advocated method marked with (*) performs best. Train $\mathcal{AA}$ and Test $\mathcal{AA}$ measure resemblance loss. $PrivacyLoss = Test\,\mathcal{AA} - Train\,\mathcal{AA}$. Utility measures test accuracy of predicting mortality. Footprint indicates whether we can export a small footprint model out of the secure area.

		Methods					
		HealthGAN (*)	**Gaussian Multivariate**	Fitted Parzen Win.	Overfitted ANM	Differential Privacy (DP)	Copy (CP)
Adversarial Accuracy	Train $\mathcal{AA}$	0.50	0.53	0.50	0.00	0.05	0.00
	Test $\mathcal{AA}$	0.51	0.55	0.50	0.50	0.52	0.50
	Privacy Loss	0.00	0.02	0.00	0.50	0.47	0.50
Utility	Area Under the Curve	0.66	0.62	0.77	0.74	0.87	0.88
Footprint	Up-to-specs	yes	yes	no	no	no	no

Results

Metrics

- Resemblance: HealthGAN generated data closest to real.
- Utility: Predictive models trained on HealthGAN data performed nearly as well as those on real data.
- Privacy: Strong resistance to adversarial/membership inference attacks compared to baselines.
- Case studies:
  - Education (training ML models without access to real patient data).
  - Research (enabling collaboration while preserving privacy).

Discussion & Conclusions

- HealthGAN provides a privacy-utility tradeoff better than existing methods.

Challenges

- Model biases can still propagate.
- Data quality depends on training set quality.
- Ethical concerns about over-reliance on synthetic data.

Future work

- Scaling to larger datasets, integration with differential privacy for stronger guarantees.

Summary

HealthGAN is a promising approach for generating privacy-preserving and usabe synthetic health data.

Difficult points

Understand statistical metrics

As métricas dependiam de conhecimentos estatísticos prévios, fazendo-me gastar bastante tempo nesta parte.

AI Perspective

ChatGPT

🔑 Main Contributions

- **Problem addressed:** Limited access to real patient data (HIPAA, GDPR) restricts research & education.
- **Proposed solution:**
  - Developed **HealthGAN** (a Wasserstein GAN) to generate synthetic electronic health records (EHR).
  - Designed a **workflow**: train GAN in a secure environment, export only the model (no real data leaves), generate synthetic data externally.
- **New evaluation metrics:** Based on nearest-neighbor adversarial accuracy to measure:
  - **Resemblance** (realism),
  - **Privacy** (non-disclosure of real records),
  - **Utility** (can data be used for tasks like prediction/replication),
  - **Footprint** (whether the model leaks real data).
- **Case studies:**
  - **Education:** Students used synthetic data in prediction challenges. Performance close to real data (~0.77 vs 0.80).
  - **Research:** Replicated 3 medical studies with synthetic data — results were generally consistent.

AI Perspective

ChatGPT

⚠️ Limitations & Open Questions

- **Scope:** Tested mainly on ICU data (MIMIC-III). Limited variety of medical contexts.
- **Data types:** Current HealthGAN doesn't fully support **time-series** or highly imbalanced datasets.
- **Resemblance issues:** Some rare conditions or long-tail features not well captured.
- **Evaluation gap:** While utility is good for education, research replication sometimes diverged (e.g., long-term mortality studies).
- **Privacy-utility tradeoff:** Balancing realism and privacy remains challenging.

🚀 Future Directions

- **Expand beyond ICU:** Apply HealthGAN to diverse medical datasets (e.g., outpatient, imaging, genomics).
- **Temporal modeling:** Extend GANs to handle **longitudinal health records**.
- **Bias & fairness:** Ensure synthetic data does not replicate or amplify biases from real datasets.
- **Integration in practice:** Create public repositories of synthetic datasets to accelerate medical AI training while preserving privacy.
- **Hybrid approaches:** Combine GANs with **differential privacy** or causal modeling for stronger guarantees.

Thank you!