

Learning Representations of Electronic Health Records for Synthetic Data Generation

Van Minh Nguyen¹

¹Mathematical Science
Florida Tech

Florida Tech, April 2022

1 Problem Background

- Electronic Health Records (EHR) and Patients' Privacy
- Researches using EHR

2 Problem Statement

- Synthetic Data Generation
- Goals and Objective

3 Prior Works

- Graph-based Method
- Deep Learning Method

4 Methods and Approaches

- (Learnable) EHR Embeddings
- Generative Model Approach

5 Evaluation of Synthetic Data

- Usability of Synthetic and Real Data
- Qualitative Evaluation
- Privacy Preservation of Synthetic Data

6 References

1 Problem Background

- Electronic Health Records (EHR) and Patients' Privacy
- Researches using EHR

2 Problem Statement

- Synthetic Data Generation
- Goals and Objective

3 Prior Works

- Graph-based Method
- Deep Learning Method

4 Methods and Approaches

- (Learnable) EHR Embeddings
- Generative Model Approach

5 Evaluation of Synthetic Data

- Usability of Synthetic and Real Data
- Qualitative Evaluation
- Privacy Preservation of Synthetic Data

6 References

Electronic Health Records (EHR)

An Electronic Health Records (EHR) is a digital version of patient's paper chart, maintained by health care providers. It includes **medical and treatment history of patients** and other information such as immunization logs, allergies, laboratory results, ...

The screenshot displays a "Rendered CCD Example" from the "Good Health Health Summary" section of a CDA document. The interface is designed to look like a medical chart. Key sections include:

- Document ID:** from the "U.S. Realm" Header (Document ID element).
- Header:** Mr. Adam Emerman, November 25, 1954, Male, 5'8", 180 lbs.
- Address:** 17 Davis St., NY, 10028, US, Tel: (212)333-1234.
- Document ID:** 300011108495111200013.
- Author:** Dr. Paula Physician, Dr. Paula Physician, Henry Doctor.
- Context info:** Dr. Paula Physician, Dr. Paula Physician, Henry Doctor.
- Allergies, Medications & Problems:** Implemented to meet "CCD" and Transition of Care Objective requirements.
- Laboratory, Adverse Reactions, Results:** Shows a table with rows for Adverse, Missing, Active, and a table for Medications and Problems.

Annotations highlight specific fields: a green box around the Document ID, another green box around the "Allergies", "Medications" & "Problems" sections, and a callout box pointing to the "Good Health Health Summary" header with the text: "'Good Health Health Summary' from the 'U.S. Realm' Header (Document Title element)".

Figure: Sample of an EHR document¹

¹HealthIT.gov, "Implementing Consolidated-Clinical Document Architecture (C-CDA) for Meaningful Use".

EHR Privacy Standard

HIPAA (US Only)

- Health Insurance Portability and Accountability Act of 1996 (HIPAA) is a federal law that protects the privacy of patients' health information.

GDPR (EU Only)

- General Data Protection Regulation (GDPR) is a regulation that requires businesses to protect the privacy of their customers' personal data.



Figure: GDPR vs HIPAA²

²Mooney, *Is HIPAA Compliant with the GDPR?*

Personal Data covered by HIPAA and GDPR

Personal Identifiable Information - PII

Any information that permits the identity of an individual to be directly or indirectly inferred, including any information that is linked or linkable to that individual.

- Examples: Name, address, phone number, SSN, medical record number, etc.

Protected Health Information - PHI

Any information, including demographic data, that relates to the individual's past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual, and that can be used to identify the individual.

- Examples: Weight, Height, Medical History, Allergies, Lab Results, etc.

Note that GDPR spectrum of cover is **any personal info**, covers both PII and PHI.

How researchers get access to EHR data

- Affiliate with healthcare providers or 3rd party companies with access to data.
- Institutionalized EHR dataset
 - PCORnet: while evaluation is free, to use and query the dataset, ones need to pay for the infrastructure cost just to access the data.
 - All of Us Research Program: freely available to participated institutions. Sponsored by NIH
- Restricted Access free EHR datasets. A few well known datasets:
 - Medical Information Mart for Intensive Care (MIMIC-III): Most well-known EHR dataset. Contains EHR data from 46,520 patients and 58,976 ICU admissions.³
 - i2b2/n2c2 Challenges Datasets (2006-2018): An annual Medical NLP challenge from at Harvard Medical School. This dataset sole focus is on NLP tasks on Clinical Notes⁴

³A. Johnson, Pollard, and Mark, *MIMIC-III Clinical Database*.

⁴Kumar et al., "Creation of a New Longitudinal Corpus of Clinical Narratives".

Limitation of restricted EHR access

- Researchers need to go over NDAs, security training, and maybe other legal contracts to get access to EHR data from healthcare providers and 3rd party. Risk of exposure to PHI/PII is almost certain.
- For institutionalized EHR dataset, researcher needs to be affiliated with participating institutions, or sponsored by them to get access.
- Even with freely available EHR datasets, at the minimum, researcher needs to go over a (rated) 50-hour-long training course on HIPAA, and needs a supervisor/guarantor to sign a confidentiality agreement, before getting access to the database. (I and Dr. White already completed this training)

Limitation of restricted EHR access

EHR researchers, especially independent researchers, with limited resources are at a massive disadvantage compare with those affiliated with institutions in terms of access to data including quality and quantity.

- MIMIC-III contains about 50,000 patients admitted to Beth Israel Deaconess Medical Center in Boston, Massachusetts.⁵ The data is not usable immediately and needs a lot of engineering and preprocessing to a standardized format.
- All of Us Research Program has EHRs on 100,000+ participants from 34 sites as of 2019⁶
- As of 2018, PCORnet has records of 80 millions of patient collected from 337 hospitals in the US. Data is rigorously screened for quality (Comformance, Completeness, Plausibility).⁷

⁵ A. Johnson, Pollard, and Mark, *MIMIC-III Clinical Database*.

⁶ "The "All of Us" Research Program".

⁷ Forrest et al., "PCORnet® 2020".

Limitation of restricted EHR access

Researchers affiliated with healthcare providers directly or via 3rd party companies can have access to a massive amount of EHRs, some of which contains PHI/PII. While this increases the chance of leaking sensitive info, researchers also has a better chance to gain insight from a more close-to-real world data (as usually deidentification process eliminates some information that may be useful for research).
e.g.: month + year of birth → age group

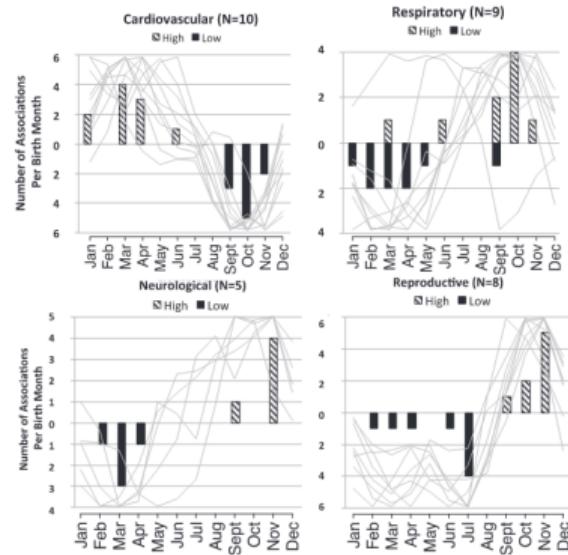


Figure: Disease risk by birth month⁸

⁸Boland et al., "Birth Month Affects Lifetime Disease Risk".

1 Problem Background

- Electronic Health Records (EHR) and Patients' Privacy
- Researches using EHR

2 Problem Statement

- Synthetic Data Generation
- Goals and Objective

3 Prior Works

- Graph-based Method
- Deep Learning Method

4 Methods and Approaches

- (Learnable) EHR Embeddings
- Generative Model Approach

5 Evaluation of Synthetic Data

- Usability of Synthetic and Real Data
- Qualitative Evaluation
- Privacy Preservation of Synthetic Data

6 References

What is Synthetic EHR Data?

Synthetic data is an **inexpensive, easy to access** alternative to real data.

- Most ML models, and almost all deep learning models, are not able to generalize well with a small amount of data.
- Synthetic EHR data, if done properly, will be completely devoid of PII (generated in compliance with HIPAA Safe Harbor method⁹), while PHI can be generated in a way that preserves the local structure/cohort of the data, but still protects the privacy of the patients.

⁹45 CFR § 164.514 - Other Requirements Relating to Uses and Disclosures of Protected Health Information ↗ ↘

Data Augmentation

Data Augmentation is one of the popular way of generating more data for ML model.

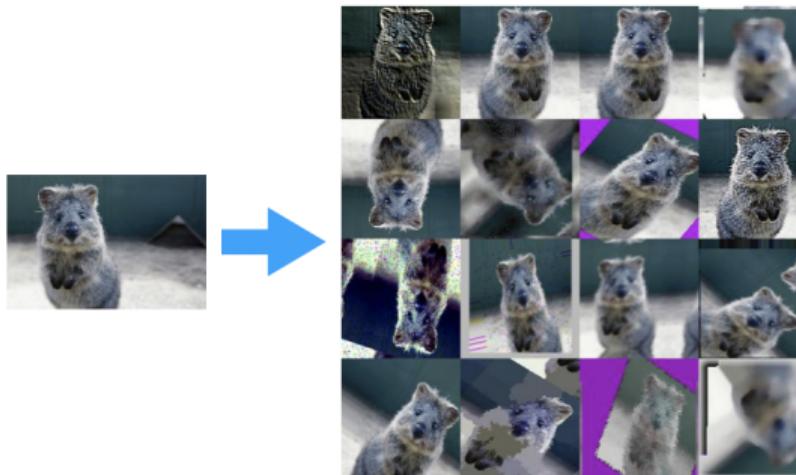


Figure credit: <https://github.com/aleju/imgaug>

Figure: Example of Image Data Augmentation

Augmentation w.r.t. EHR

- However, data needs to be anonymized before augmentation, which might not be completely in compliance with Privacy Rules. As such, those data are not classified as "Synthetic" and still have to go under the same scrutiny under HIPAA as original EHR data.
- Augmentation still requires access to the original EHR data, which is not easily accessible.
- Augmentation, in general, will improve the resilience of the model against noise, but might not allow the model to learn some new structure of the data as well as a model trained on a robust synthetic dataset.

Synthetic Data

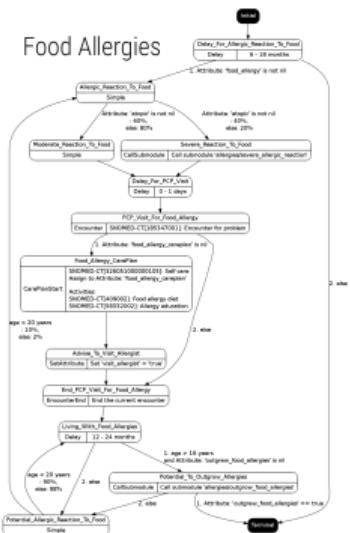


Figure: Synthetic data using Knowledge Graph¹

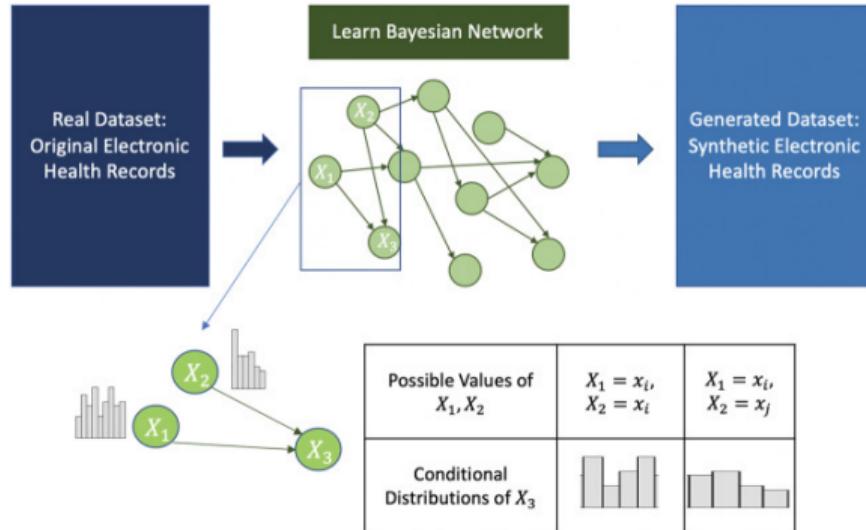


Figure: Synthetic data using Bayesian networks²

¹ Synthea by the Standard Health Record Collaborative.

² Kaur et al., "Application of Bayesian Networks to Generate Synthetic Health Data".

Synthetic Data

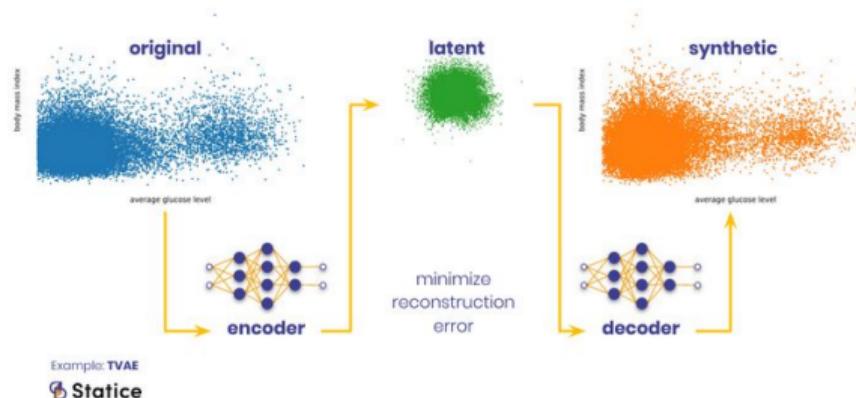


Figure: Synthetic data using Variational Auto Encoder (VAE)¹

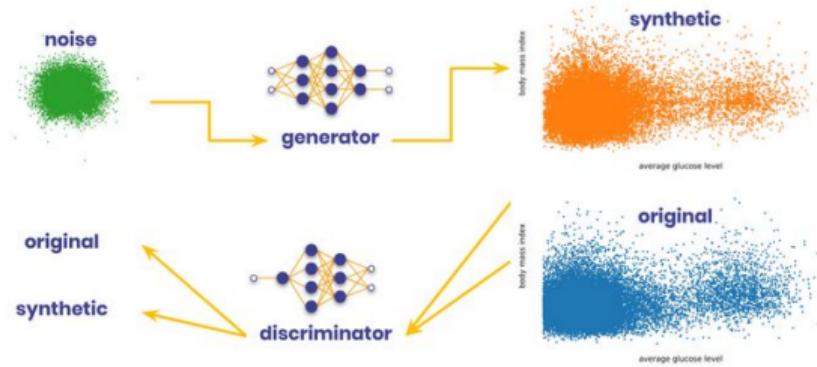


Figure: Synthetic data using Generative Adversarial Networks (GANs)¹

¹ [How Do You Generate Synthetic Data?](#), p. 1.

Synthetic Data w.r.t. EHR

- PII can be randomly generated for each patient: Randomly combine vocab of first name, last name for fake name, randomly generated SSN starting 000, 666, 900-999¹⁰, randomly generated DOB, etc.
- PHI data can be generated from the original EHR data with added random noises from a distribution, but can also be randomly generated from the distribution of original EHR population, or can be generated using a generative model from random noises

¹⁰ Social Security Number Randomization.

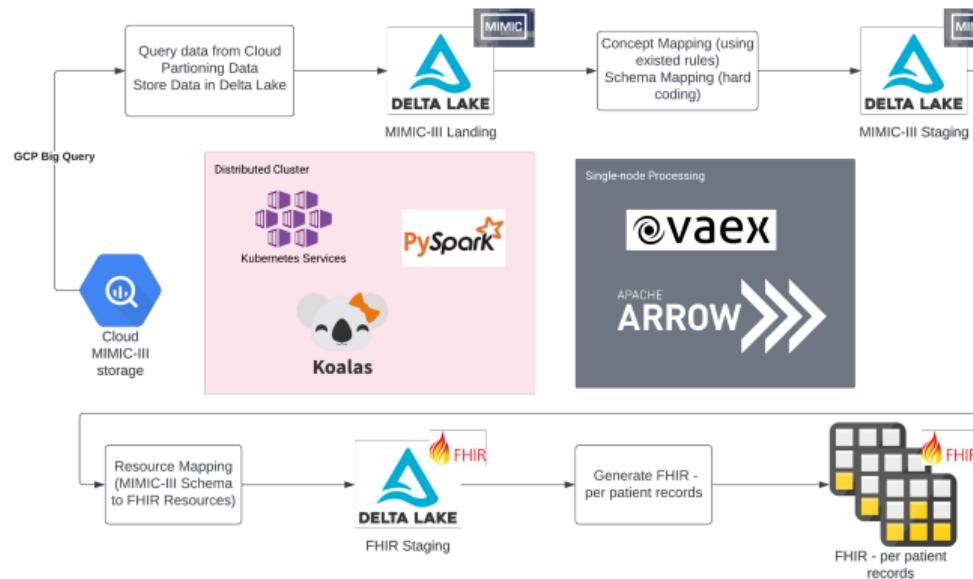
Long-term Goals

- Develop a fully distributed, scalable Extraction, Transformation, Loading (ETL) for (Learnable) Embeddings Generation
- Train data synthesis model using GANs that utilize the generated embeddings
 - Can be centralized for training/generation
- A scalable, parallelizable, and distributed data synthesis pipeline

Short-term Goals

Completed

- Build a scalable and distributed ETL pipeline using Spark and Koala.



Short-term Goals

Near Future

- Feature engineering of parameterizable EHR fields (embeddings)
 - Unsupervised embeddings or Semi-supervised embeddings
 - Scalable embeddings pipeline
- Ensure all information in original format can be extracted from embeddings

In 3-4 months

- Investigate data (embeddings) synthesis model with privacy preservation method

Other Goals

Further Study

- Measure usability of synthesized data
 - Distribution Similarity (Dimension-wise/Per-feature, Latent Distribution, Joint Distribution)
 - Results with common ML Models using synthesized data
 - Privacy Measurement via multiple methods and attacks
- Investigate the privacy of synthetic data

Not-in-consideration Goals

- Ontology Mapping (Concept Mapping/Normalization)
- Processing of Clinical Notes
 - Anonymization of PHI in Clinical Notes
 - Entity extractions from Notes.
 - Missing EHR data extraction from Notes.

1 Problem Background

- Electronic Health Records (EHR) and Patients' Privacy
- Researches using EHR

2 Problem Statement

- Synthetic Data Generation
- Goals and Objective

3 Prior Works

- Graph-based Method
- Deep Learning Method

4 Methods and Approaches

- (Learnable) EHR Embeddings
- Generative Model Approach

5 Evaluation of Synthetic Data

- Usability of Synthetic and Real Data
- Qualitative Evaluation
- Privacy Preservation of Synthetic Data

6 References

Synthea¹¹

- Probably the most well known synthetic data generator.
- Use a graph-based approach to generate synthetic data.
- Graph data is generated from a graph template, evaluated by clinicians.
- Can generate longitudinal data, however the validity of the data is not guaranteed.

¹¹ Synthea by the Standard Health Record Collaborative.

The Synthetic Health And Research Data (SHARED) Project^{13, 14}

- Generating Synthetic Longitudinal Patient Data with the PrivBayes Method (Bayesian Network)
- Failed to produce synthetic longitudinal patient data
- First batch of synthetic data will be released in a year¹²

¹² "Pie & AI: Aarhus - Synthetic Data for Health Research".

¹³ *The Shared Health and Research Project.*

¹⁴ Perkonoja, "Generating Synthetic Longitudinal Patient Data with the PrivBayes Method".



Simulacrum^{16, 17}

- Using Probability distribution to model the data and Chi-square to validate the synthesized data.
- Treat all continuous variable as categorical.
- Statistical sampling rather than randomizing the data.
- Downsides:
 - No clinical temporal features and relationships
 - No statistical analysis comparison with baseline data
 - Only for exploration
 - No privacy evaluation (and since it's sampled from actual population it can violate patient's privacy)
- Research on Bayesian Network for synthetic data in consideration of privacy¹⁵

¹⁵ "Pie & AI: Aarhus - Synthetic Data for Health Research".

¹⁶ *Simulacrum Home Page*.

¹⁷ *Simulacrum Informal White Paper - Methodology Overview*.

DeepSynthBody¹⁹

- Using GANs to generate synthetic **images** data.
 - Only GI tract image model so far.
 - For each target, a new model needs to be trained from scratch.
- Development of a framework rather than a model/algorithm.¹⁸

¹⁸ "Pie & AI: Aarhus - Synthetic Data for Health Research".

¹⁹ Thambawita et al., "DeepSynthBody".

StyleGAN2-ADA for generation of synthetic skin lesions²⁰

- Using StyleGAN2-ADA to generate synthetic **images** data for skin lesions.
- Privacy issue: reidentification of patient and privacy breaches on the dataset.
- An attacker with the original dataset and execute a membership inference attack

Figure: Example of skin lesions generated by StyleGAN2-ADA

²⁰Karras et al., "Training Generative Adversarial Networks with Limited Data".



Gretel AI - Gretel Synthetics^{21, 22}

- Using seq2seq RNN + tokenizer to generate data
- Use differential privacy in the framework
- Tested usability of dataset by comparing dimension-wise correlation with baseline data
- Evaluation of the privacy of the synthetic data with differential privacy approach

²¹ Deep Dive on Generating Synthetic Data for Healthcare.

²² GitHub Gretelai/Gretel-Synthetics.

1 Problem Background

- Electronic Health Records (EHR) and Patients' Privacy
- Researches using EHR

2 Problem Statement

- Synthetic Data Generation
- Goals and Objective

3 Prior Works

- Graph-based Method
- Deep Learning Method

4 Methods and Approaches

- (Learnable) EHR Embeddings
- Generative Model Approach

5 Evaluation of Synthetic Data

- Usability of Synthetic and Real Data
- Qualitative Evaluation
- Privacy Preservation of Synthetic Data

6 References

What is embedding?

Mathematical Definition

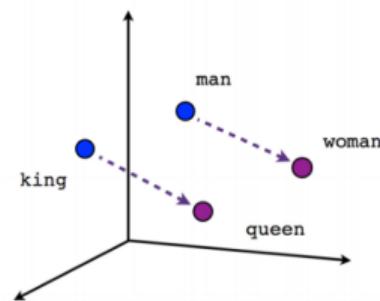
W.r.t. metric space, an embedding is an injective continuous mapping from a metric space to another metric space that preserves ratio between distances (with some distortion) (should I do this in terms of topology?)

Machine Learning/Deep Learning

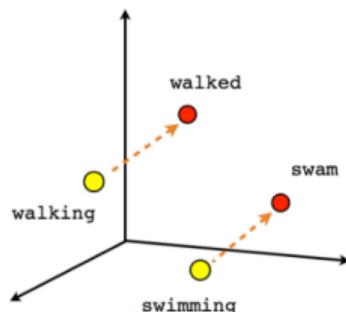
A vector/matrix output in latent space of a model that learned from data. (need source)
Since all ML algorithm deal with numbers on real line, function applies are injective and continuous, although the assumption of preserving distances ratio still need to be tested, the ML definition of embeddings is a good analog to the Mathematical definition of embeddings.

Embeddings example: Word embeddings

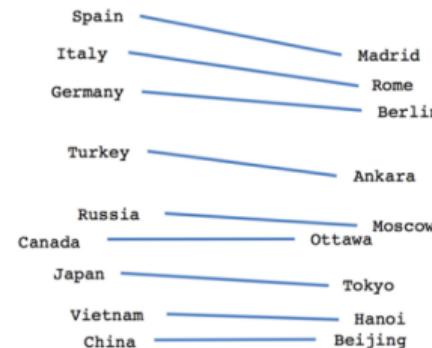
Embedding is a representation of data that is usable by ML models (in continuous space) that preserves structure of the data



Male-Female



Verb tense



Country-Capital

Figure: Example of word embeddings preserving meaning distance²³

²³Ratheesh, *Word Embeddings, WordPiece and Language-Agnostic BERT (LaBSE)*.

Embeddings example: PCA embeddings

Embedding can also be used as a representation of data in lower dimension space, which reduces noises and allows for more efficient learning of the model.

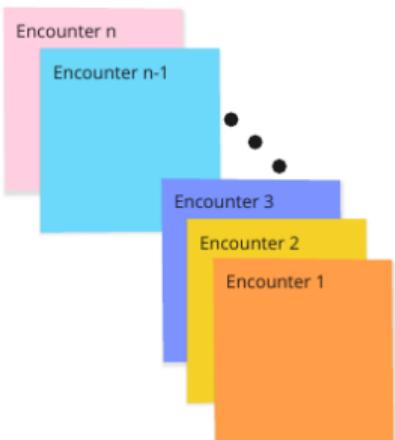


Figure: Original image and Reconstruction using PCA embeddings²⁴

²⁴ PCA Theory Examples - Rhea.

Embeddings Architecture

Current high level approach:



$T(\text{Patient}) = \text{Stack}(T(\text{Encounter}_0), T(\text{Encounter}_1), T(\text{Encounter}_2) \dots T(\text{Encounter}_n))$

- The goal of this embedding is a data storage that is learnable by ML models (in continuous space), and can be queried back to the original EHR format.
- The maximum number of layers in the embedding model is limited by the number of maximum encounter M
 - 75th percentile of number of encounters per patient in real EHR
 - $M = 3$ or $M = 4$ (number of channels in normal image)
- Non encounter is padded with zeroes (0).

Figure: Architecture Diagram - High Level

Embeddings Architecture

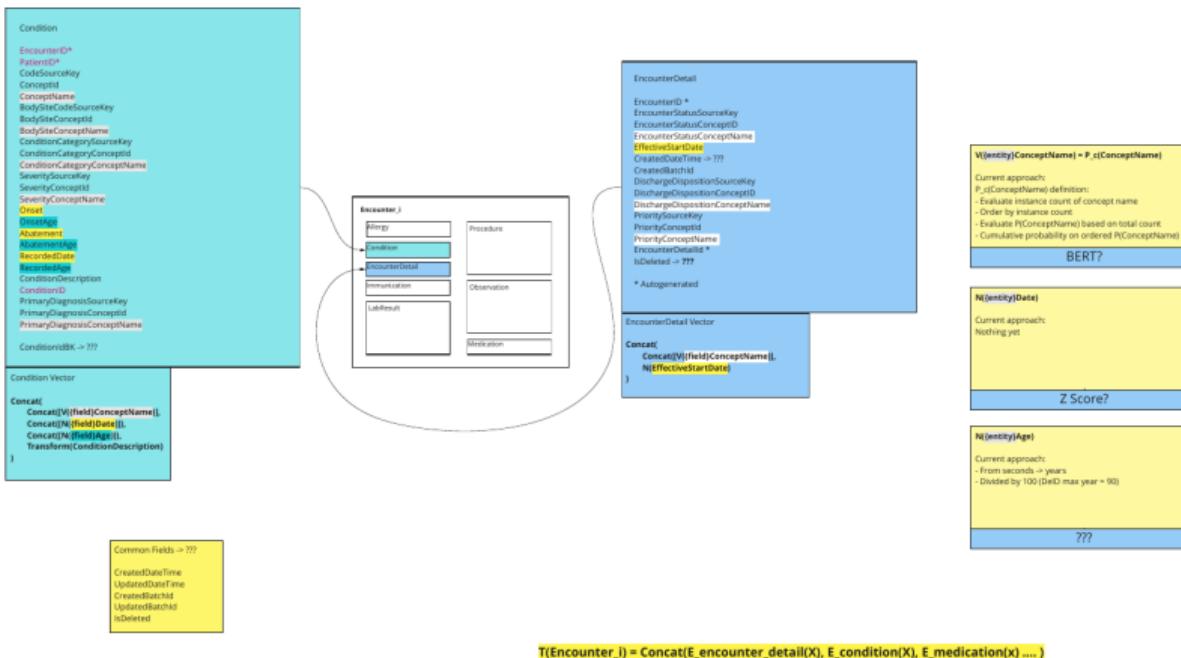


Figure: Architecture Diagram - Per Layer

Embeddings Architecture

- Each layer/channel is embedded data on a 784×784 ($784 = 28 \times 28$) matrix.
- Each table is designated a region of the matrix.
- For each categorical features:
 - Get the frequency distribution of the categorical values.
 - Generate an index table using cumulative frequency distribution.
 - Example: With feature F1, we have categories A,B,C with frequency distribution [0.2, 0.3, 0.5], then the index table is A:[0,0.2), B:[0.2,0.5), C:[0.5,1), with value of 1 indicate the category is not present in the encounter feature.
 - Categorical values are transformed into a continuous value by generating a random uniform number with indexed range.
 - Example: Categorical A is transformed to 0.12 ($0.12 \in Uniform([0, 0.2))$)
 - Continuous values will be passed directly to the matrix

Privacy-Preserving GANs Support Clinical Data Sharing²⁵

- Using AC-GAN to generate data
- Introduce differential privacy to GAN framework (proven to be robust against post-processing)
 - Applied Local differential privacy (DP for each output instead of aggregated output)
 - Implemented DP in Adam optimizer
- A set of statistical models to evaluate similarity of the population.
- Machine Learning models to evaluate if the synthetic data can be used for analysis.
- Evaluation of the privacy of the synthetic data.

²⁵ Beaulieu-Jones et al., "Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing".

Generative Adversarial Network

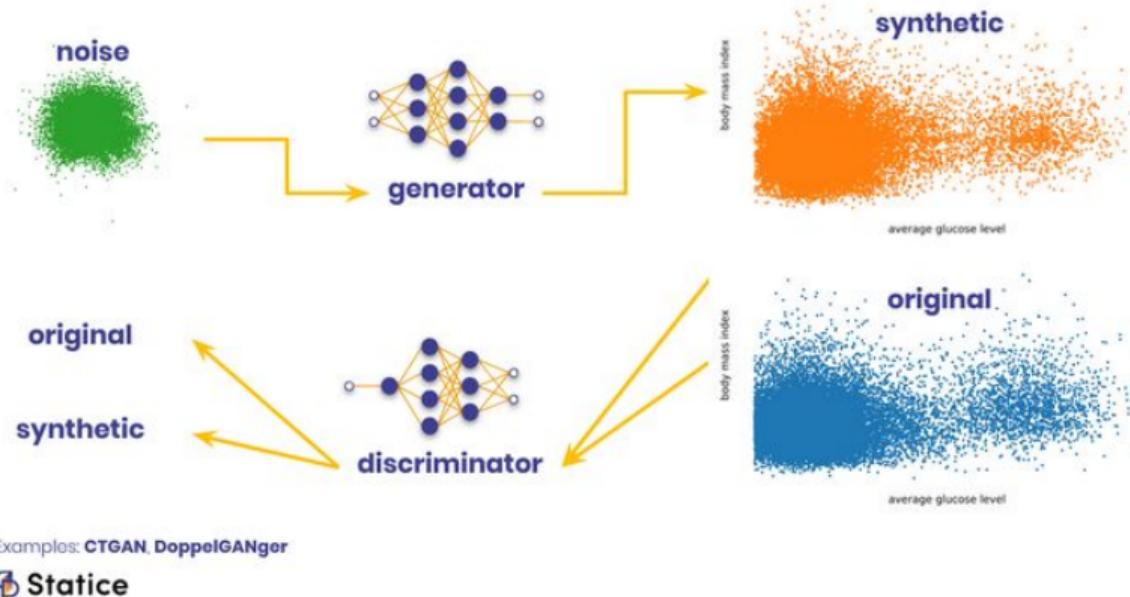


Figure: Synthetic data using Generative Adversarial Networks (GANs)²⁶

²⁶ How Do You Generate Synthetic Data?

Generative Adversarial Network²⁷

Generative Adversarial Network was first introduced by Ian Goodfellow in 2014.

Definition

A classical GAN consists of 2 models, a discriminator and a generator. Denote discriminator as a function \mathcal{D} and generator as \mathcal{G} .

- $\mathcal{G}(z)$ take in vector z generated from a distribution p_g , and map it to training data space. The goal of $\mathcal{G}(z)$ is to learn the structure of training data distribution p_{data} .
- $\mathcal{D}(x)$ outputs the probability that input x comes from the training data (real), as such, \mathcal{D} is a binary classifier.

Two models play a minimax game where discriminator tries to maximize the probability of correctly classifying the data as real or fake, while generator tries to minimize the probability of the discriminator detect its output being fake.

²⁷Goodfellow et al., "Generative Adversarial Nets".

Generative Adversarial Network²⁸

- For generator loss, intuitively we would want to minimize $L_G = \log(1 - \mathcal{D}(G(z)))$ but as demonstrated by Goodfellow, this doesn't provide sufficient gradient, so we'll maximize $\log(\mathcal{D}(\mathcal{G}(z)))$ (or minimize $L_G = -\log(\mathcal{D}(\mathcal{G}(z)))$) and use real label as ground truth instead.
- Since discriminator is a binary classifier, we'll use binary cross-entropy loss function for the output:

$$L_{\mathcal{D}} = -\log(\mathcal{D}(x)) - \log(1 - \mathcal{D}(\mathcal{G}(z))), \text{ for } x \sim p_{data}, z \sim p_g$$

We want to minimize this loss function

²⁸Goodfellow et al., "Generative Adversarial Nets".

ACGAN - Auxiliary Classifier Generative Adversarial Network

ACGAN is classical GAN with input for generative model is random noise and auxiliary label for downstream classification. This model allows generated data to contain structures that are relevant to the label.

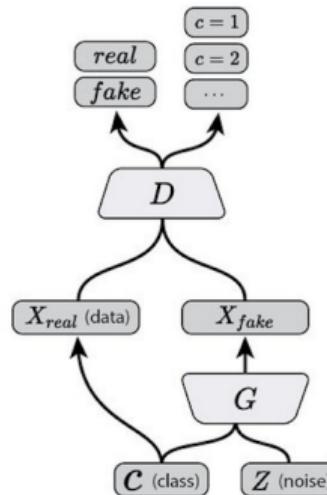


Figure: Synthetic data using Generative Adversarial Networks (GANs)²⁹

²⁹ Brownlee, *How to Develop an Auxiliary Classifier GAN (AC-GAN) From Scratch with Keras.*

1 Problem Background

- Electronic Health Records (EHR) and Patients' Privacy
- Researches using EHR

2 Problem Statement

- Synthetic Data Generation
- Goals and Objective

3 Prior Works

- Graph-based Method
- Deep Learning Method

4 Methods and Approaches

- (Learnable) EHR Embeddings
- Generative Model Approach

5 Evaluation of Synthetic Data

- Usability of Synthetic and Real Data
- Qualitative Evaluation
- Privacy Preservation of Synthetic Data

6 References

Statistical evaluation

Synthesized dataset needs to have a similar statistical distribution compared with read dataset.

Dimension-wise Distribution

Compare the distribution of each dimension (feature) of the synthetic data with the distribution of the real data.

Latent Distribution

Using an autoencoder to encode both real and synthesized data. Distribution of features in latent space must be similar.

Joint Distribution

Using Jensen-Shannon Divergence (JSD) to measure the distance of joint distributions between real and synthetic data.

Probabilistic Validity of Synthetic EHRs in Feature Space \mathcal{F}

- Let \mathcal{F} be a feature space where EHRs are defined
 - Let $x_1, \dots, x_n \in \mathcal{F}$ be a sample of real EHRs
 - Let $z_1, \dots, z_m \in \mathcal{F}$ be a sample of synthetic EHRs
- Realistic synthetic EHRs should come from the same probability distribution over \mathcal{F} as real EHRs
- How can we compare distributions on an unwieldy, high dimensional feature space \mathcal{F} ?
 - Extract low-dimensional empirical joint distributions from the real and synthetic EHRs to compare
 - Map the samples into another space and compare the distributions in that space
- We propose to subject the empirical distributions to a battery of probabilistic evaluation techniques, including
 - Distance measures between distributions
 - Hypothesis tests that samples come from the same distribution

Probabilistic Validity of Synthetic EHRs in Feature Space \mathcal{F} - Distance Between Distributions

- Suppose we have selected k features and extracted them from the samples of real EHR and synthetic EHR
 - Let $f_X : \mathcal{F}_k \rightarrow [0, \infty)$ be the empirical joint probability density of the features extracted from real EHRs
 - Let $f_Z : \mathcal{F}_k \rightarrow [0, \infty)$ be the empirical joint probability density of the features extracted from synthetic EHRs
- f_X and f_Z should be similar if the real and synthetic EHRs come from the same distribution, so one test of validity is to measure the distance between them by a choice of distance metric, e.g.

$$D_{SSE}(f_X \| f_Z) = \sum_{x \in \mathcal{F}_k} \|f_X(x) - f_Z(x)\|_F \quad (\text{Sum of squared errors})$$

$$D_{KL}(f_X \| f_Z) = \sum_{x \in \mathcal{F}_k} f_X(x) \log \left(\frac{f_Z(x)}{f_X(x)} \right) \quad (\text{Kullback-Leibler divergence})$$

$$D_{JS}(f_X \| f_Z) = \frac{1}{2} D_{KL} \left(f_X \left\| \frac{f_X + f_Z}{2} \right. \right) + \frac{1}{2} D_{KL} \left(f_Z \left\| \frac{f_X + f_Z}{2} \right. \right) \quad (\text{Jensen-Shannon divergence})$$

Probabilistic Validity of Synthetic EHRs in Feature Space \mathcal{F} - Hypothesis Testing

- An alternative approach is to perform hypothesis tests of the null hypothesis:

The real and synthetic samples have the same distribution

- If the empirical distributions are one-dimensional and $\mathcal{F}_1 \subseteq \mathbb{R}$, the standard Kolmogorov-Smirnov test can be used, with test statistic

$$D_{n,m} = \sup_{x \in \mathcal{F}_1} |F_{X,n}(x) - F_{Z,n}(x)| \quad (F \text{ indicates an empirical CDF})$$

- Higher dimensional versions are well-known for two³⁰ and three³¹ dimensions while recent work³² suggests an approach for more dimensions.

³⁰ Two-Dimensional Goodness-of-Fit Testing in Astronomy — *Monthly Notices of the Royal Astronomical Society* — Oxford Academic.

³¹ Gosset, “A Three-Dimensional Extended Kolmogorov-Smirnov Test as a Useful Tool in Astronomy”.

³² Naaman, “On the Tight Constant in the Multivariate Dvoretzky–Kiefer–Wolfowitz Inequality”.

Probabilistic Validity of Synthetic EHRs in Embedding Space \mathcal{E}

- Let $\phi_\theta : \mathcal{F} \rightarrow \mathcal{E}$ be the function learned to map EHRs to embeddings
- The choice embedding space \mathcal{E} is a hyperparameter under investigation, but present work uses:
 - $\mathcal{E} = \mathbb{R}^{784 \times 784 \times M}$ where each depth channel corresponds to an encounter
- A reasonable embedding space \mathcal{E} is still a high-dimensional space, but the embedded EHRs have a common shape and size
- In this space, we propose the same distance metric-based comparisons on the empirical joint distributions of the embeddings
 - i.e. comparing empirical joint distributions of $\phi_\theta(x_1), \dots, \phi_\theta(x_n)$ and $\phi_\theta(z_1), \dots, \phi_\theta(z_m)$
- The high dimensionality of the embedding space and lack of interpretability of individual features rules out statistical testing in \mathcal{E} without unrealistic independence assumptions

Data Utilization

Train on Synthetic Test on Real

Utilizing multiple ML models to train on Synthetic data and test on Real data.

Train on Real Test on Synthetic

Similar as above but use Real data to train and Synthetic data to test.

Compare ML model performance on Real and Synthetic

On the same downstream task, we'll evaluate the performance of ML models on the real data and the synthetic data separately.

- Proposed models: Decision Tree/Random Forest, J48 (via python-weka-wrapper), Logistic Regression, XGBoost
- Compare performance and feature importance diagram between real and synthetic data

Human Blind Evaluation

While this is already done with a discriminator model in GANs, we'll also use human annotators to rate the "realism" of the synthetic data. A team of subject-matter experts will evaluate the synthetic data to determine if data "looks real".

- For this to be reasonable, the data generated should be from a specific cohort that all subject-matter experts have expertises.
- Rate from 1 to 10 (discrete), on how "real" the data is.
 - Need a guideline draft for scoring/rating
- Perform a non-parametric hypothesis test on the result
 - H0: for random selected records from synthetic and real data, the probability synthetic data being real is the same as real data being real (?)

Privacy Evaluation

- Since PII is generated in compliance with HIPAA Safe Harbor, we don't need to worry about PII leakage.
- Both Synthesized Health Data (including PHI) and Generative Model need to be in compliance with privacy standards and resilience against different types attacks:
 - Privacy Standard using Differential Privacy
 - Membership Inference attacks
 - Attribute Disclosure attacks

Differential Privacy^{33, 34}

We'll implement differential privacy and use their evaluation method for evaluating privacy of synthetic data.

Definition

Let $X = Z^m$ is the metric space of data set of $m \in \mathbb{N}$ rows, with metric $d(x, x') = |\{i \in \mathbb{N} : x_i \neq x'_i\}|$ for $x, x' \in X$

Let $\epsilon, \delta > 0$. A randomized algorithm/mechanism $\mathcal{M} : X \rightarrow Range(\mathcal{M})$ is **(ϵ, δ) -differentially private** if for all $S \subseteq Range(\mathcal{M})$ and for all $x, x' \in X$ s.t. $d(x, x') \leq 1$:

$$Pr[\mathcal{M}(x) \in S] \leq e^\epsilon Pr[\mathcal{M}(x') \in S] + \delta$$

Where ϵ is privacy budget (how much privacy risk taken), δ is the probability of leaking privacy info. If $\delta \rightarrow 0$, we say that \mathcal{M} is **ϵ -differentially private**

³³Pei, *A Tail of Two Densities*.

³⁴Dwork and Roth, "The Algorithmic Foundations of Differential Privacy".

Differential Privacy^{35, 36}

Theorem

The Gaussian Mechanism:

$$\mathcal{M}(x) = f(x) + Y$$

where $Y \sim \mathcal{N}(0, \sigma^2)$, for $\sigma \geq \Delta f \cdot \frac{\sqrt{\ln e^\epsilon \delta^{-2}}}{\epsilon}$, and $\Delta f = \sup_{x, x' \in X: d(x, x')=1} ||f(x) - f(x')||_2$ is the sensitivity of function f ,

preserves ϵ -differential privacy.

Note

Independent use of an ϵ_1 -differentially private and an ϵ_2 -differentially private, when taken together, is $(\epsilon_1 + \epsilon_2)$ -differentially private, which is weaker than each of the used algorithms.

³⁵Pei, A Tail of Two Densities.

³⁶Dwork and Roth, "The Algorithmic Foundations of Differential Privacy".

Differential Privacy

Implementation

- To implement Differential Privacy in training GANs, we only need to apply noise during the training of discriminator as it's the only component that has access to the real data.
- We will be using Tensorflow's DP implementation (TensorFlow Privacy^a) or PyTorch's DP implementation (Meta Opacus^b) to apply differential privacy to discriminator's optimizer.
- TensorFlow Privacy implementations also has optimizer calculation of the privacy budget (ϵ, δ) for the dataset
- None of these implementation evaluate exact minimum σ but instead use gradient clipping as sensitivity Δf and "noise_multiplier" in place of $\frac{\sqrt{\ln e^\epsilon \delta^{-2}}}{\epsilon}$. We can utilize this knowledge to quickly minimize added noise while still preserving privacy according to the privacy budget.

^aImplement Differential Privacy with TensorFlow Privacy — Responsible AI Toolkit.

^bOpacus · Train PyTorch Models with Differential Privacy.

Membership Inference attacks

Assume attacker has a subset of original dataset, they attempt to determine if any of the patient in the breached dataset is used to train the model. Use metrics (cosine, Euclidean, ...) to evaluate the similarity (distance) between the real and synthetic dataset entries.

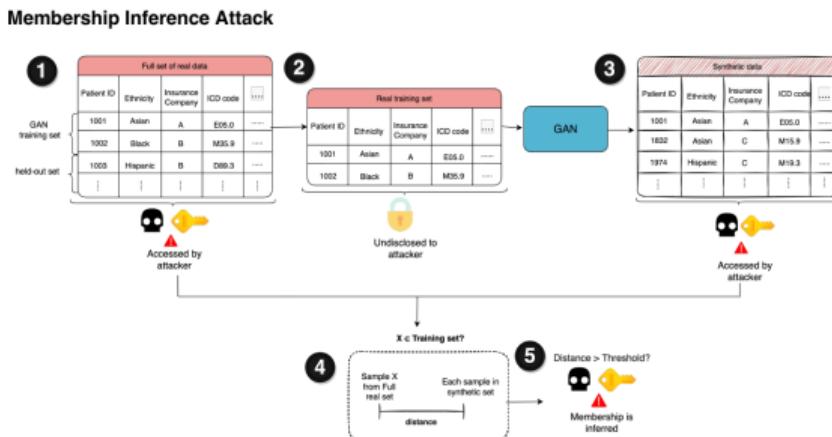


Figure: Diagram for Membership Inference attacks³⁷

³⁷ Ghosheh, Li, and Zhu, "A Review of Generative Adversarial Networks for Electronic Health Records".

Attribute Disclosure attacks

Attacker can infer additional attributes of a patient knowing a subset of other attributes by training a model on synthesized dataset and inferring on breached dataset

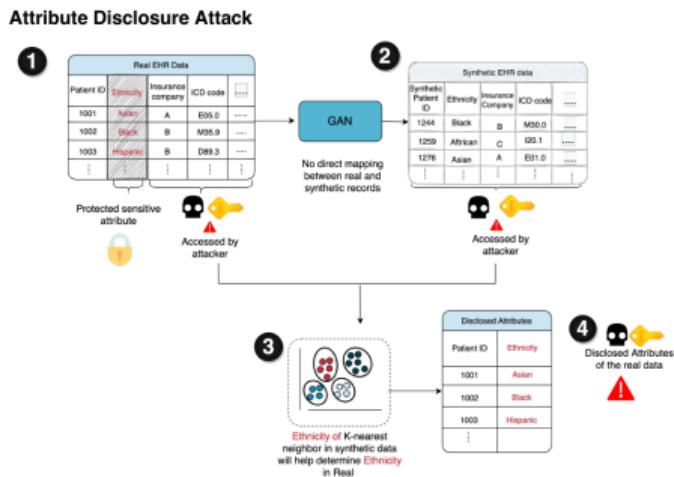


Figure: Diagram for Attribute Disclosure attacks³⁸

³⁸Ghosheh, Li, and Zhu, "A Review of Generative Adversarial Networks for Electronic Health Records".

1 Problem Background

- Electronic Health Records (EHR) and Patients' Privacy
- Researches using EHR

2 Problem Statement

- Synthetic Data Generation
- Goals and Objective

3 Prior Works

- Graph-based Method
- Deep Learning Method

4 Methods and Approaches

- (Learnable) EHR Embeddings
- Generative Model Approach

5 Evaluation of Synthetic Data

- Usability of Synthetic and Real Data
- Qualitative Evaluation
- Privacy Preservation of Synthetic Data

6 References

Further Reading I

-  Baowaly, Mrinal Kanti et al. "Synthesizing Electronic Health Records Using Improved Generative Adversarial Networks". In: *Journal of the American Medical Informatics Association : JAMIA* 26.3 (Dec. 7, 2018), pp. 228–241. ISSN: 1067-5027. DOI: 10.1093/jamia/ocy142. pmid: 30535151. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647178/> (visited on 04/10/2022).
-  Choi, Edward et al. "Generating Multi-label Discrete Patient Records Using Generative Adversarial Networks". Jan. 11, 2018. arXiv: 1703.06490 [cs]. URL: <http://arxiv.org/abs/1703.06490> (visited on 04/10/2022).
-  "Common Data Model (CDM) Specification, Version 6.0". In: (), p. 193.

Further Reading II

-  Henry, Sam et al. "2018 N2c2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records". In: *Journal of the American Medical Informatics Association* 27.1 (Jan. 1, 2020), pp. 3–12. ISSN: 1527-974X. DOI: 10.1093/jamia/ocz166. URL: <https://doi.org/10.1093/jamia/ocz166> (visited on 04/09/2022).
-  Iannucci, Stefano et al. "A Comparison of Graph-Based Synthetic Data Generators for Benchmarking Next-Generation Intrusion Detection Systems". In: *2017 IEEE International Conference on Cluster Computing (CLUSTER)*. 2017 IEEE International Conference on Cluster Computing (CLUSTER). Sept. 2017, pp. 278–289. DOI: 10.1109/CLUSTER.2017.54.

Further Reading III

-  Johnson, Alistair E. W. et al. "MIMIC-III, a Freely Accessible Critical Care Database". In: *Scientific Data* 3.1 (1 May 24, 2016), p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35. URL: <https://www.nature.com/articles/sdata201635> (visited on 04/09/2022).
-  Mironov, Ilya. "Renyi Differential Privacy". In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* (Aug. 2017), pp. 263–275. DOI: 10.1109/CSF.2017.11. arXiv: 1702.07476. URL: <http://arxiv.org/abs/1702.07476> (visited on 04/11/2022).

Cited Sources I

- 45 CFR § 164.514 - Other Requirements Relating to Uses and Disclosures of Protected Health Information. LII / Legal Information Institute. URL:
<https://www.law.cornell.edu/cfr/text/45/164.514> (visited on 04/09/2022).
- Beaulieu-Jones, Brett K. et al. "Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing". In: *Circulation: Cardiovascular Quality and Outcomes* 12.7 (July 2019), e005122. DOI: 10.1161/CIRCOUTCOMES.118.005122. URL:
<https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.118.005122> (visited on 03/27/2022).
- Boland, Mary Regina et al. "Birth Month Affects Lifetime Disease Risk: A Phenome-Wide Method". In: *Journal of the American Medical Informatics Association* 22.5 (Sept. 1, 2015), pp. 1042–1053. ISSN: 1527-974X, 1067-5027. DOI: 10.1093/jamia/ocv046. URL:
<https://academic.oup.com/jamia/article/22/5/1042/930268> (visited on 04/08/2022).

Cited Sources II

-  Brownlee, Jason. *How to Develop an Auxiliary Classifier GAN (AC-GAN) From Scratch with Keras*. Machine Learning Mastery. July 18, 2019. URL:
<https://machinelearningmastery.com/how-to-develop-an-auxiliary-classifier-gan-ac-gan-from-scratch-with-keras/> (visited on 04/11/2022).
-  Deep Dive on Generating Synthetic Data for Healthcare. URL: <https://gretel.ai/blog/deep-dive-on-generating-synthetic-data-for-healthcare> (visited on 04/10/2022).
-  Dwork, Cynthia and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4 (2013), pp. 211–407. ISSN: 1551-305X, 1551-3068. DOI: 10.1561/0400000042. URL:
<http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042> (visited on 04/11/2022).

Cited Sources III

-  Forrest, Christopher B. et al. "PCORnet® 2020: Current State, Accomplishments, and Future Directions". In: *Journal of Clinical Epidemiology* 129 (Jan. 2021), pp. 60–67. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2020.09.036. pmid: 33002635. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7521354/> (visited on 04/09/2022).
-  Ghosheh, Ghadeer, Jin Li, and Tingting Zhu. "A Review of Generative Adversarial Networks for Electronic Health Records: Applications, Evaluation Measures and Data Sources". Mar. 14, 2022. arXiv: 2203.07018 [cs]. URL: <http://arxiv.org/abs/2203.07018> (visited on 04/08/2022).
-  GitHub Gretelai/Gretel-Synthetics. GitHub. URL: <https://github.com/gretelai/gretel-synthetics> (visited on 04/10/2022).

Cited Sources IV

-  Goodfellow, Ian et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014. URL: <https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html> (visited on 04/11/2022).
-  Gosset, E. "A Three-Dimensional Extended Kolmogorov-Smirnov Test as a Useful Tool in Astronomy". In: *Astronomy and Astrophysics* 188.1 (Dec. 1987), pp. 258–264. ISSN: 0004-6361. URL: <https://ui.adsabs.harvard.edu/abs/1987A&A...188..258G/abstract> (visited on 04/11/2022).
-  HealthIT.gov. "Implementing Consolidated-Clinical Document Architecture (C-CDA) for Meaningful Use". In: (), p. 88. URL: https://www.healthit.gov/sites/default/files/c-cda_and_meaningfulusecertification.pdf.

Cited Sources V

-  *How Do You Generate Synthetic Data? - Statice.* URL:
<https://www.statice.ai//post/how-generate-synthetic-data> (visited on 04/10/2022).
-  *Implement Differential Privacy with TensorFlow Privacy — Responsible AI Toolkit.* TensorFlow. URL: https://www.tensorflow.org/responsible_ai/privacy/tutorials/classification_privacy (visited on 04/11/2022).
-  Johnson, Alistair, Tom Pollard, and Roger Mark. *MIMIC-III Clinical Database*. Version 1.4. PhysioNet, 2015. DOI: 10.13026/C2XW26. URL:
<https://physionet.org/content/mimiciii/1.4/> (visited on 04/09/2022).
-  Karras, Tero et al. “Training Generative Adversarial Networks with Limited Data”. Oct. 7, 2020. arXiv: 2006.06676 [cs, stat]. URL: <http://arxiv.org/abs/2006.06676> (visited on 04/10/2022).

Cited Sources VI

-  Kaur, Dhamanpreet et al. "Application of Bayesian Networks to Generate Synthetic Health Data". In: *Journal of the American Medical Informatics Association: JAMIA* 28.4 (Mar. 18, 2021), pp. 801–811. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa303. pmid: 33367620.
-  Kumar, Vishesh et al. "Creation of a New Longitudinal Corpus of Clinical Narratives". In: *Journal of Biomedical Informatics*. Supplement: Proceedings of the 2014 I2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data 58 (Dec. 1, 2015), S6–S10. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2015.09.018. URL: <https://www.sciencedirect.com/science/article/pii/S1532046415002129> (visited on 04/09/2022).
-  Mooney, Greg. *Is HIPAA Compliant with the GDPR? - Ipswitch*. URL: <https://www.ipswitch.com/blog/is-hipaa-compliant-with-the-gdpr> (visited on 04/09/2022).

Cited Sources VII

-  Naaman, Michael. "On the Tight Constant in the Multivariate Dvoretzky–Kiefer–Wolfowitz Inequality". In: *Statistics & Probability Letters* 173 (June 1, 2021), p. 109088. ISSN: 0167-7152. DOI: 10.1016/j.spl.2021.109088. URL: <https://www.sciencedirect.com/science/article/pii/S016771522100050X> (visited on 04/11/2022).
-  Opacus · Train PyTorch Models with Differential Privacy. URL: <https://opacus.ai/> (visited on 04/11/2022).
-  PCA Theory Examples - Rhea. URL: https://www.projectrhea.org/rhea/index.php/PCA_Theory_Examples (visited on 04/10/2022).
-  Pei, Yuchen. A Tail of Two Densities. URL: <https://ypei.org/posts/2019-03-13-a-tail-of-two-densities.html>.

Cited Sources VIII

-  Perkonoja, Katariina. "Generating Synthetic Longitudinal Patient Data with the PrivBayes Method". In: (), p. 89.
-  "Pie & AI: Aarhus - Synthetic Data for Health Research". (Steno Diabetes Center Aarhus). Mar. 14, 2022. URL: <https://www.eventbrite.com/e/pie-ai-aarhus-synthetic-data-for-health-research-tickets-269303432817>.
-  Ratheesh, Bijula. *Word Embeddings, WordPiece and Language-Agnostic BERT (LaBSE)*. MLearning.ai. Feb. 20, 2021. URL: <https://medium.com/mlearning-ai/word-embeddings-wordpiece-and-language-agnostic-bert-labse-98c7626878c7> (visited on 04/10/2022).
-  *Simulacrum Home Page*. simulacrum.healthdatainsight.org.uk. URL: <https://simulacrum.healthdatainsight.org.uk/> (visited on 04/10/2022).
-  *Simulacrum Informal White Paper - Methodology Overview*. Nov. 18, 2021.

Cited Sources IX

-  *Social Security Number Randomization.* URL:
<https://www.ssa.gov/employer/randomization.html> (visited on 04/09/2022).
-  *Synthea by the Standard Health Record Collaborative.* URL:
<https://synthetichealth.github.io/synthea/> (visited on 04/10/2022).
-  Thambawita, Vajira et al. "DeepSynthBody: The Beginning of the End for Data Deficiency in Medicine". In: *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. 2021 International Conference on Applied Artificial Intelligence (ICAPAI). May 2021, pp. 1–8. DOI: 10.1109/ICAPAI49758.2021.9462062.
-  "The "All of Us" Research Program". In: *New England Journal of Medicine* 381.7 (Aug. 15, 2019), pp. 668–676. ISSN: 0028-4793. DOI: 10.1056/NEJMsr1809937. pmid: 31412182. URL: <https://doi.org/10.1056/NEJMsr1809937> (visited on 04/09/2022).
-  *The Shared Health and Research Project.* URL: <https://shared.landen.co/> (visited on 04/10/2022).

Cited Sources X

-  *Two-Dimensional Goodness-of-Fit Testing in Astronomy — Monthly Notices of the Royal Astronomical Society — Oxford Academic.* URL:
<https://academic.oup.com/mnras/article/202/3/615/967854> (visited on 04/11/2022).