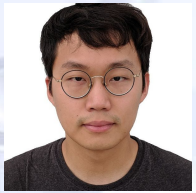
NLP SUMMIT HEALTHCARE

# EHR-Safe: Generating High-fidelity and Privacy-Preserving Synthetic Data Generation


**Sercan O Arik**
Research Scientist
Google Cloud



**Jinsung Yoon**
Research Scientist
Google Cloud

# Introduction

- Applying AI in healthcare has critical barrier: Privacy concerns limit data sharing
- Limitations of current solutions:
  - **De-identification**: High re-identification risks with ad-hoc datasets
  - **Public datasets**: Very few exist, limited in scope

**Synthetic data can be promising for sharing private healthcare data to unblock AI adoption**

- We envision this tech enabling "**Digital Twin**" for research and analysis: generating synthetic samples that have the **same characteristics** as the original data

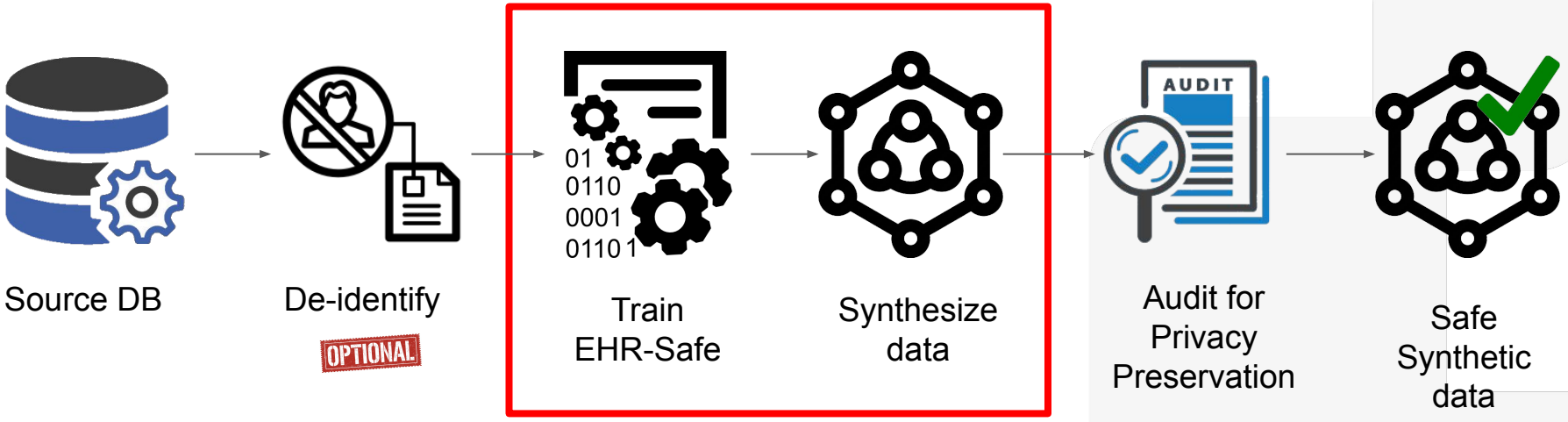
NLP SUMMIT

# TL;DR

- **Objective:**
  - Generating high-fidelity, privacy-preserving synthetic EHR data (**Digital twins**)
- **Method:**
  - Two-stage framework: Sequential encoder decoder + GAN models
- **Results - Fidelity:**
  - The predictive **performance on training on real vs. synthetic is almost the same**
  - The proposed method **significantly outperforms alternatives**
- **Results - Privacy:**
  - **Almost-perfect robustness** across three different privacy attacks

NLP SUMMIT

# High-level architecture




NLP SUMMIT

# Technical challenges

- **Measuring privacy** is not straightforward
  - Even though we generate synthetic data (no straightforward one-to-one mapping), it does not mean that we are free from privacy issues
- Generating **statistically-similar** synthetic dataset is challenging
  - Complex correlations, different length of stays, missingness
- **Mixed types of data**
  - Generating static, temporal, and categorical features simultaneously

NLP SUMMIT

# EHR-Data - components

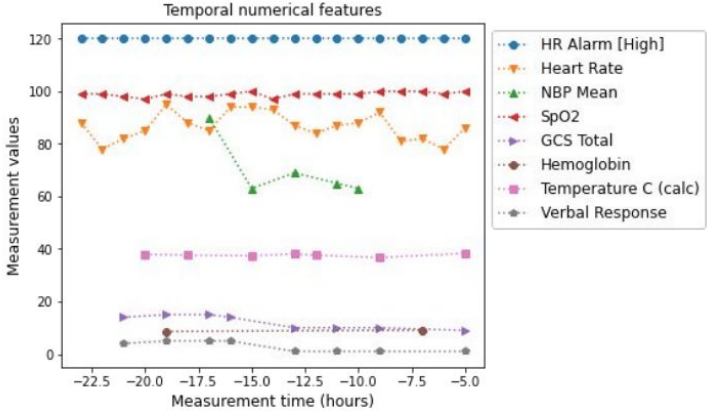
| trigger_id | hours_since_admit | gender | mortality | age_years | religion | marital_status | condition_code | HR Alarm [Low] | Heart Rate | Position Change | Restraints Evaluated | Heart Rhythm |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Time | | Static numerical | | | Static categorical | | Temporal numerical | | | Temporal categorical | |
| 100016 | -23 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 86.0 | None | None | SR (Sinus Rhythm) |
| 100016 | -22 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 92.0 | Done | Reapplied | SR (Sinus Rhythm) |
| 100016 | -21 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 92.0 | None | None | SR (Sinus Rhythm) |
| 100016 | -20 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 79.0 | Done | Reapplied | SR (Sinus Rhythm) |
| 100016 | -19 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 94.0 | None | None | SR (Sinus Rhythm) |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 199994 | -4 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 85.0 | Done | Behavior Conts | Normal Sinus |
| 199994 | -3 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 86.0 | None | None | Normal Sinus |
| 199994 | -2 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 81.0 | None | None | Normal Sinus |
| 199994 | -1 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 86.0 | None | None | Normal Sinus |
| 199994 | 0 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 74.0 | None | None | Normal Sinus |

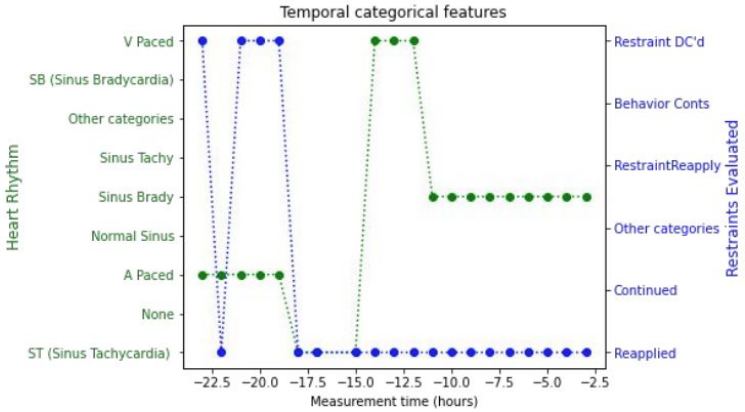
- **Measurement time**: Irregular measurement times
- **Static numerical features**: Age
- **Static categorical features:** Marital status, religion
- **Temporal numerical features**: Heart rate, blood pressure
- **Temporal categorical features**: Position change, heart rhythm

NLP SUMMIT

# Healthcare datasets are highly heterogeneous!

- We have both **static** and **temporal** features
- We have both **numerical** and **categorical** features
- EHR data has **high missing components** (not all the lab tests are measured)
- Each patient has different **length of sequences**




**Static numerical and categorical features**

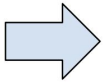
| Feature | Value |
|---|---|
| **Age** | 68 |
| **Gender** | Male |
| **Marital states** | Single |
| **Medical code** | '41071' |

# Data modeling

- **Temporal features**
  - Vital signs, lab tests
- **Mask features**
  - Binary vector representing the missingness of the features
- **Static features**
  - Age, gender, marital status
- **Measurement time**
  - Should be ordered in chronological way

**Raw data**

| Time | Name | Value |
|---|---|---|
| - | Age | 64 |
| - | Mortality | 1 |
| 1 | SBP | 110 |
| 1 | HR | 67 |
| 3 | HR | 65 |
| 5 | SBP | 115 |
| 5 | HR | 67 |
| 6 | SBP | 107 |
| 6 | HR | 70 |
| 8 | HR | 71 |
| 9 | SBP | 112 |



**Measurement time**

| Time | 1 | 3 | 5 | 6 | 8 | 9 |
|---|---|---|---|---|---|---|

**Time-varying features**

| | | | | | | |
|---|---|---|---|---|---|---|
| SBP | 110 | N/A | 115 | 107 | N/A | 112 |
| HR | 67 | 65 | 67 | 70 | 71 | N/A |

**Mask time-varying features**

| | | | | | | |
|---|---|---|---|---|---|---|
| SBP | 1 | 0 | 1 | 1 | 0 | 1 |
| HR | 1 | 1 | 1 | 1 | 1 | 0 |

**Static features**

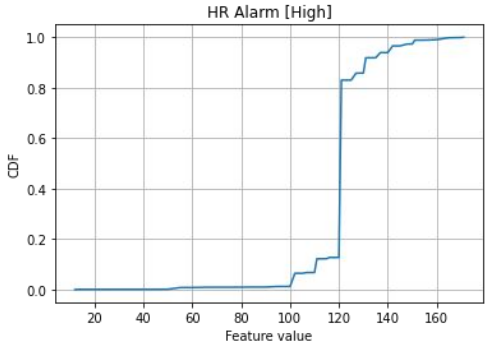
| | |
|---|---|
| Age | 64 |
| Mortality | 1 |

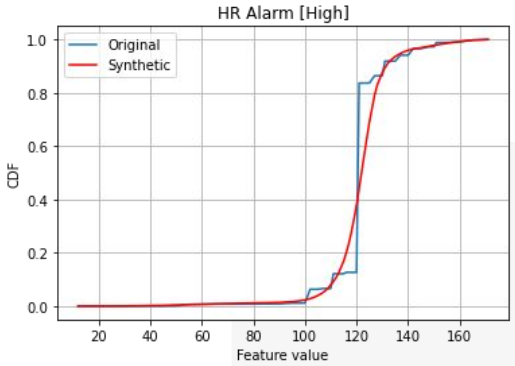
**Mask static features**

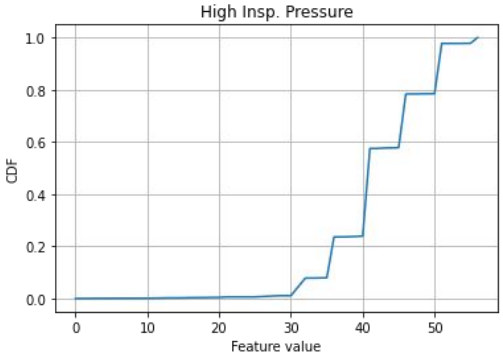
| | |
|---|---|
| Age | 1 |
| Mortality | 1 |

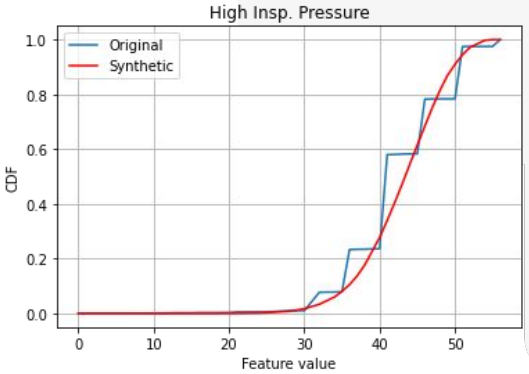
NLP SUMMIT

# Challenge 1: Generating highly non-uniform distributions



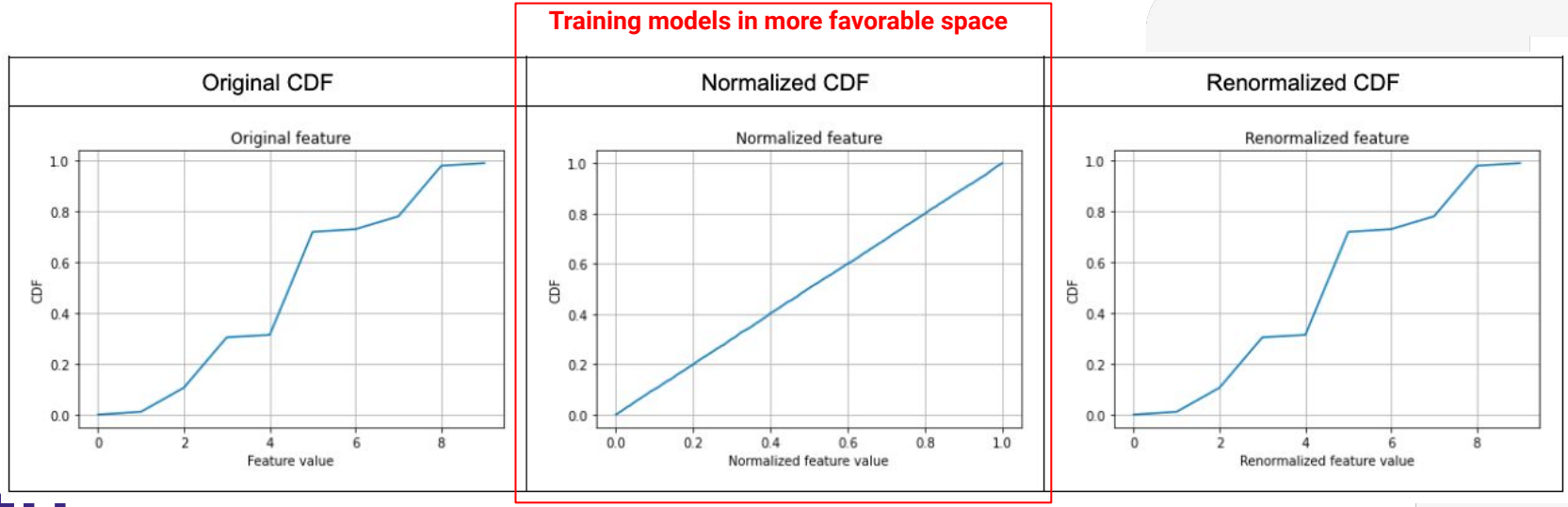




NLP SUMMIT

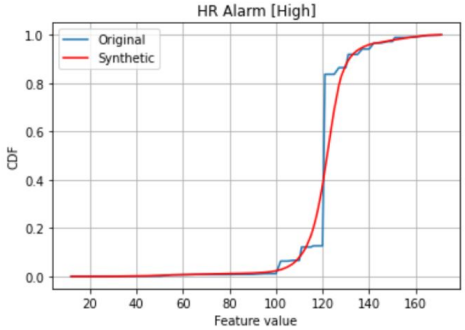
# Solution 1: Stochastic normalization

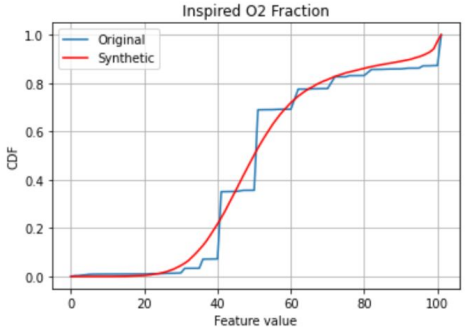
- **We propose mapping each feature value into a certain range with uniform distribution**
  - Feature values with higher frequency have wider range
  - More favorable for deep neural networks to generate synthetic data and reconstruction

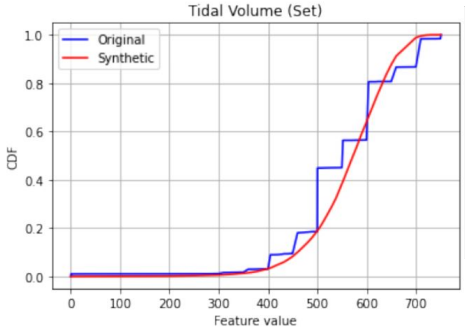


# Improvements with stochastic normalization

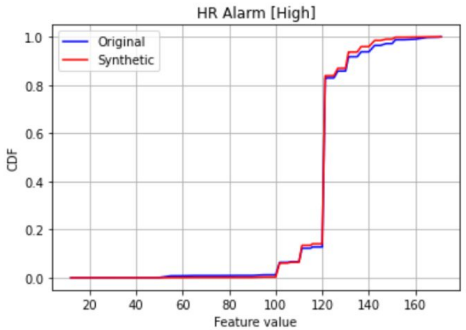
**Without** stochastic normalization

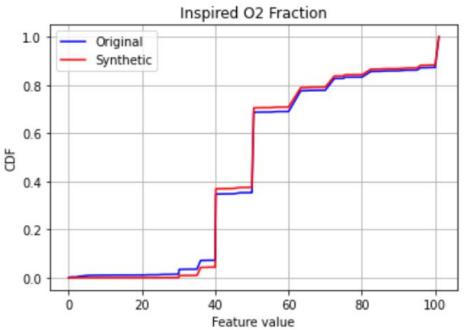


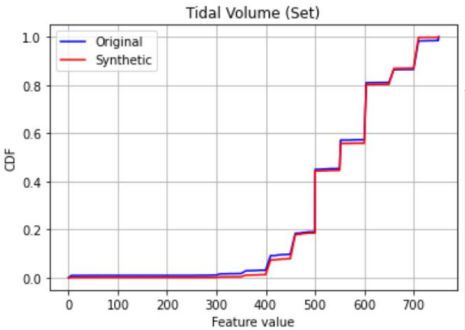




**With** stochastic normalization




NLP SUMMIT

# Challenge 2: Categorical features in EHR

- We **transform** categorical information into meaningful numerical space
- At inference, we **re-transform** to the original categorical data

### Static categorical features

| Variable | Value |
|---|---|
| Race | White |
| Marital status | Single |
| Religion | None |

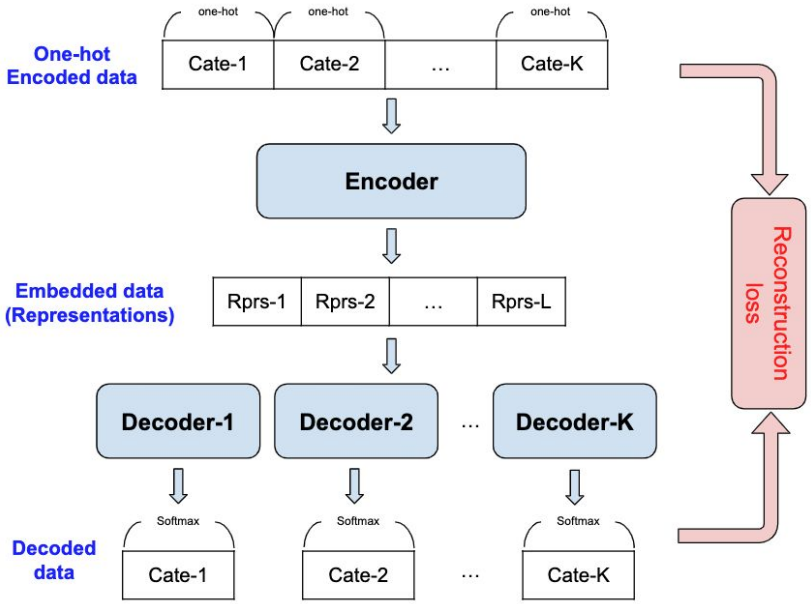
### Temporal categorical features

| Date | Variable | Value |
|---|---|---|
| 07.10.2015 | Oral care | Swab |
| 07.10.2015 | Heart rhythm | Sinus Tachy |
| 07.10.2015 | Restraint location | Both arms |
| 07.11.2015 | Restraints evaluated | Continued |
| 07.11.2015 | Ventilator mode | CPAP+PS |
| 07.11.2015 | Ventilator type | Drager |

NLP SUMMIT

# Solution 2: Handling categorical features

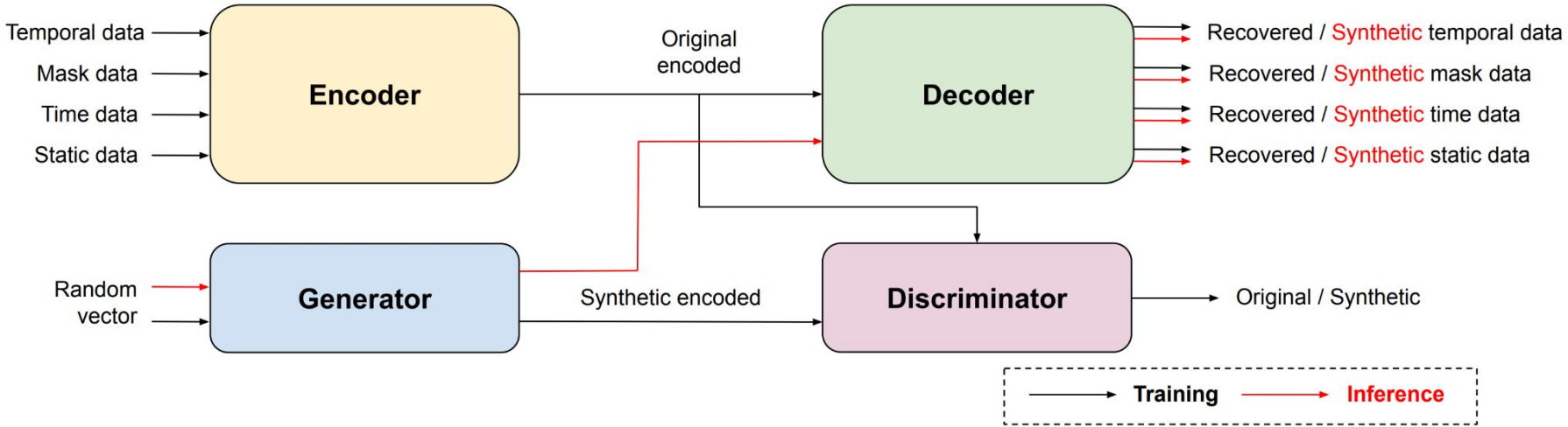
- We transform one-hot encoded categorical data into **low-dimensional latent space**
- We use **separate decoders** to generate the categorical outputs




NLP SUMMIT

# Overall architecture

- **Two-stage model training**
  - Encoder-decoder model training with reconstruction loss
  - GAN model training with WGAN-GP model on latent space



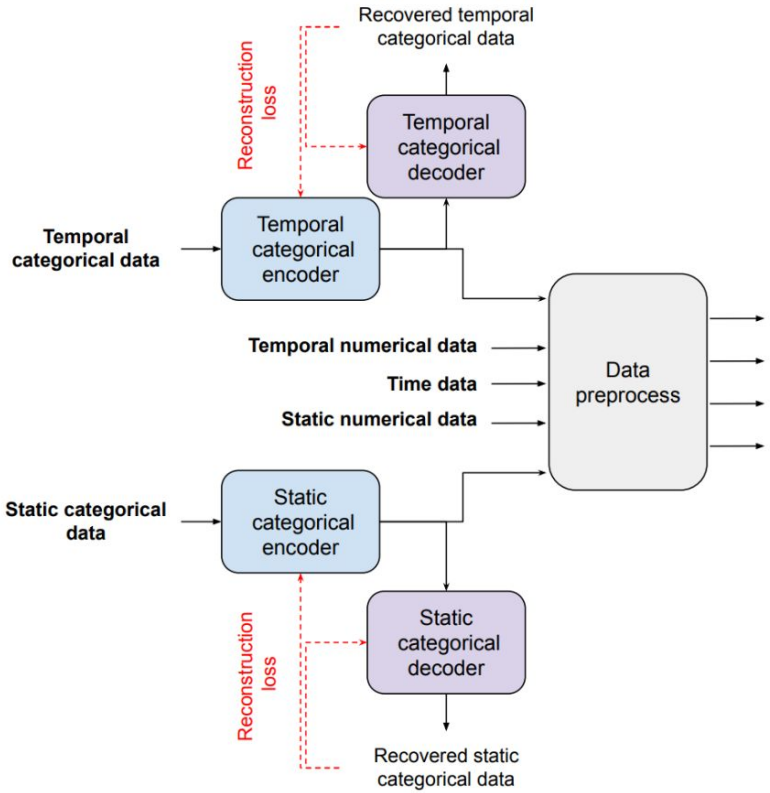

NLP SUMMIT

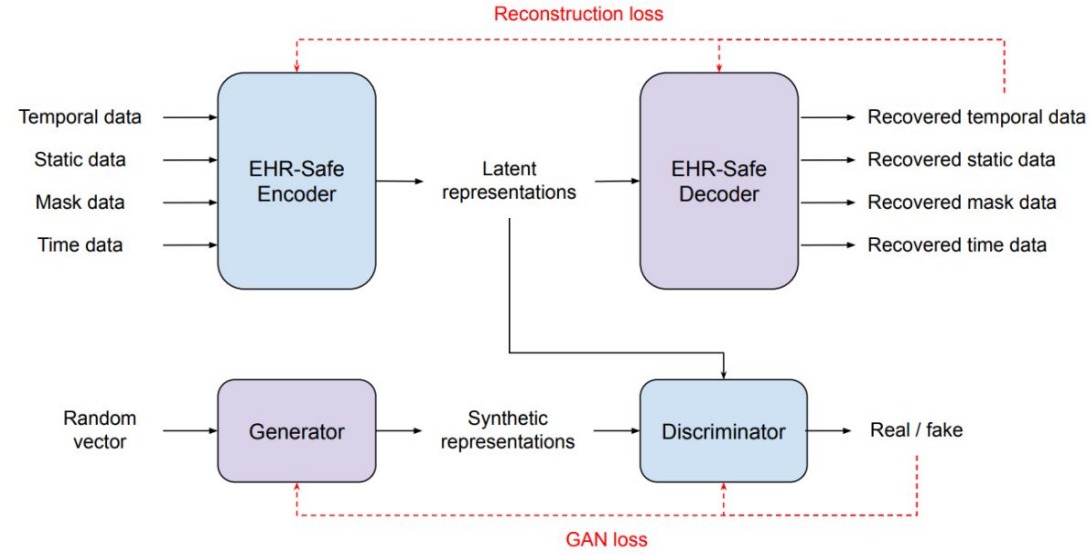
# Training details

At the end of training, we get **trained**

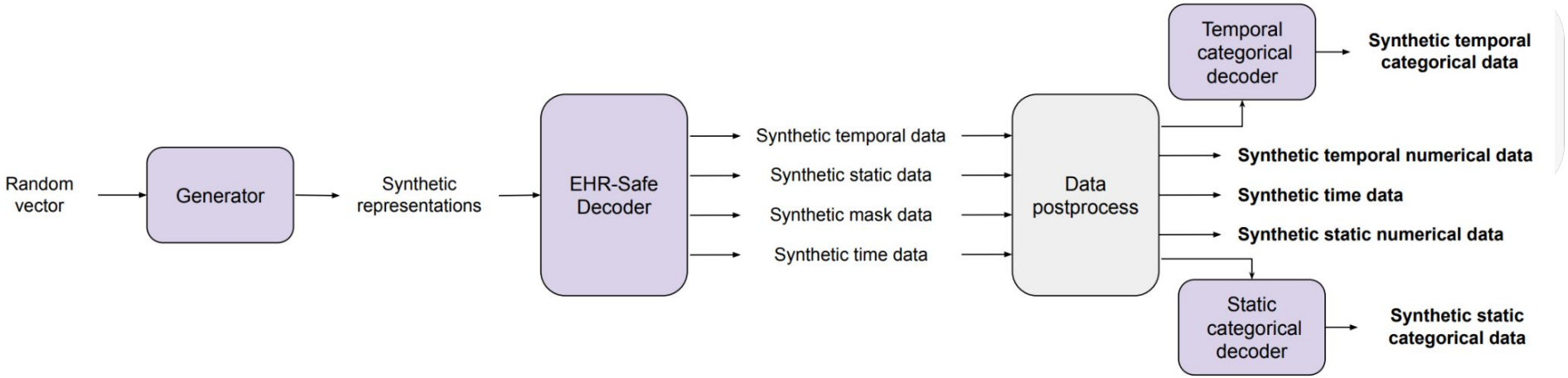
- **Generator, EHR-Safe decoder, Two categorical decoders**

Synthetic representations are **hard to distinguish** from original representations



# Inference details




Generated synthetic EHR data from **random vector** sampled from known distributions.

- Distributions of **generated heterogeneous synthetic EHR data** is very similar with the distributions of the original EHR data

NLP SUMMIT

# Examples of generated synthetic data

**Synthetic EHR**

| trigger_id | hours_since_admit | gender | mortality | age_years | religion | marital_status | condition_code | HR Alarm [Low] | Heart Rate | Position Change | Restraints Evaluated | Heart Rhythm |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | -23.0 | female | 0.0 | 66.0 | Protestant | Married | 34982 | NaN | 80.0 | NaN | NaN | Sinus Tachy |
| 0.0 | -22.0 | female | 0.0 | 66.0 | Protestant | Married | 34982 | NaN | 80.0 | NaN | NaN | Sinus Tachy |
| 0.0 | -21.0 | female | 0.0 | 66.0 | Protestant | Married | 34982 | NaN | 81.0 | NaN | NaN | Sinus Tachy |
| 0.0 | -20.0 | female | 0.0 | 66.0 | Protestant | Married | 34982 | NaN | 81.0 | NaN | NaN | Sinus Tachy |
| 0.0 | -19.0 | female | 0.0 | 66.0 | Protestant | Married | 34982 | NaN | 81.0 | NaN | NaN | Sinus Tachy |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 15955.0 | -6.0 | male | 0.0 | 45.0 | Adventist | Widowed | 4241 | 66.0 | 73.0 | Done | Behavior Conts | Sinus Tachy |
| 15955.0 | -5.0 | male | 0.0 | 45.0 | Adventist | Widowed | 4241 | 64.0 | 75.0 | Done | Behavior Conts | Sinus Tachy |
| 15955.0 | -4.0 | male | 0.0 | 45.0 | Adventist | Widowed | 4241 | NaN | 73.0 | NaN | NaN | NaN |
| 15955.0 | -3.0 | male | 0.0 | 45.0 | Adventist | Widowed | 4241 | NaN | NaN | NaN | NaN | NaN |
| 15955.0 | -2.0 | male | 0.0 | 45.0 | Adventist | Widowed | 4241 | NaN | NaN | NaN | NaN | NaN |

**Original EHR**

Time | Static numerical | Static categorical | Temporal numerical | Temporal categorical

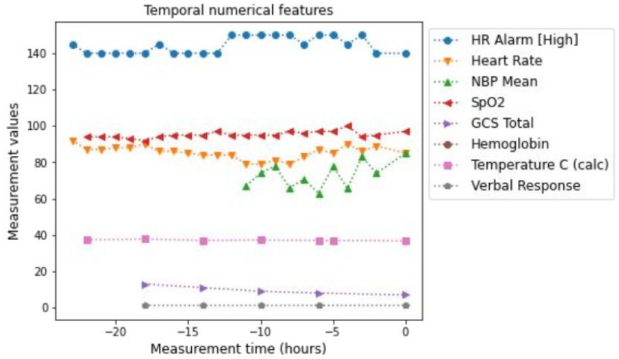
| trigger_id | hours_since_admit | gender | mortality | age_years | religion | marital_status | condition_code | HR Alarm [Low] | Heart Rate | Position Change | Restraints Evaluated | Heart Rhythm |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100016 | -23 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 86.0 | None | None | SR (Sinus Rhythm) |
| 100016 | -22 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 92.0 | Done | Reapplied | SR (Sinus Rhythm) |
| 100016 | -21 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 92.0 | None | None | SR (Sinus Rhythm) |
| 100016 | -20 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 79.0 | Done | Reapplied | SR (Sinus Rhythm) |
| 100016 | -19 | male | 0.0 | 56.0 | Protestant | Single | 5070 | NaN | 94.0 | None | None | SR (Sinus Rhythm) |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 199994 | -4 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 85.0 | Done | Behavior Conts | Normal Sinus |
| 199994 | -3 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 86.0 | None | None | Normal Sinus |
| 199994 | -2 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 81.0 | None | None | Normal Sinus |
| 199994 | -1 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 86.0 | None | None | Normal Sinus |
| 199994 | 0 | female | 0.0 | 58.0 | Roman Catholic Church | Single | 486 | NaN | 74.0 | None | None | Normal Sinus |

NLP SUMMIT

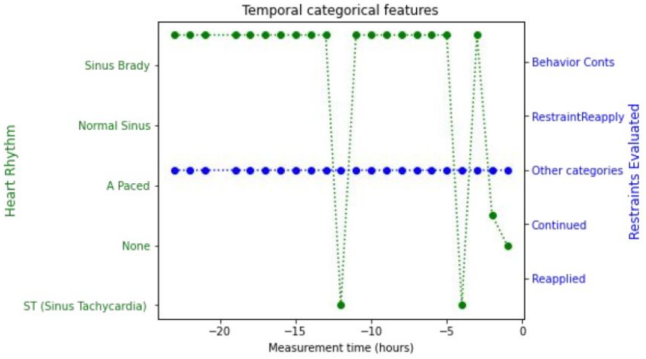
# Examples of generated synthetic data

**Synthetic EHR**

**Temporal numerical**




**Temporal categorical**



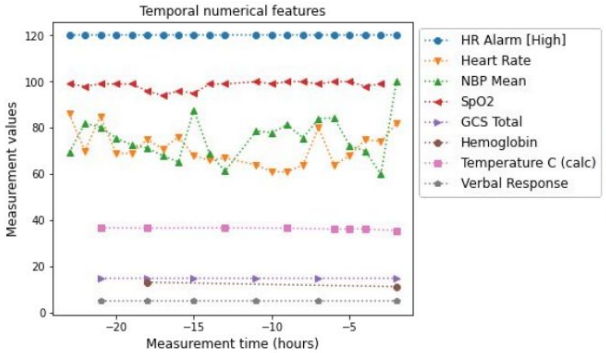

**Static**

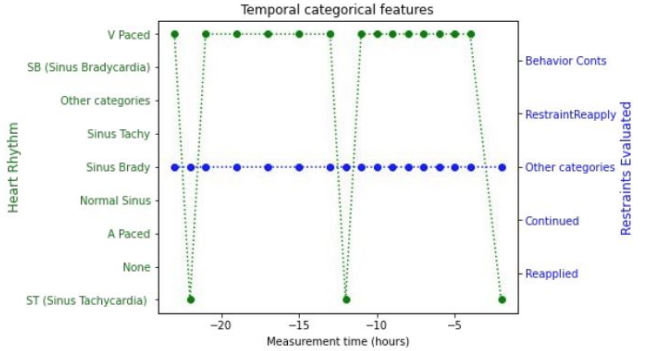
**Static numerical and categorical features**

| Feature | Value |
|---|---|
| Age | 73 |
| Gender | Male |
| Marital states | Married |
| Medical code | '0845' |

---

**Real EHR**




**Static numerical and categorical features**

| Feature | Value |
|---|---|
| Age | 76 |
| Gender | Male |
| Marital states | Married |
| Medical code | '0389' |

NLP SUMMIT

# Synthetic data evaluation

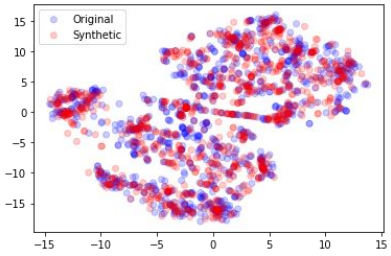
- **Fidelity Metrics**
  - How **statistically similar** is the synthetic data to the original data?
  - **Downstream model performance**: how well are the predictive properties of the original data preserved?

- **Privacy metrics**
  - Can we predict whether a patient's data **participated in generative model training?**
  - If an attacker had partial real data, would the synthetic data **reveal more information about a patient?**
  - Can we **predict sensitive features more accurately** with synthetic data?

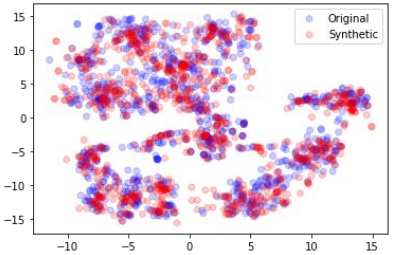
NLP SUMMIT

# Qualitative experiments - tSNE (Coverage)

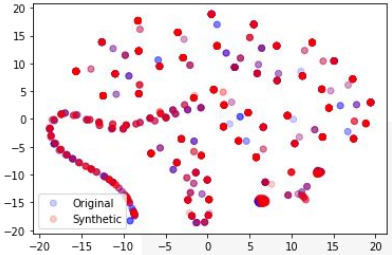
- Coverage of **synthetic data (red)** is almost overlapping with the **coverage of real data (blue)**



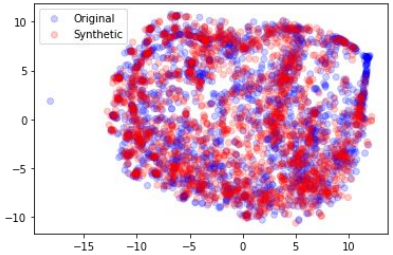

**MIMIC-III Temporal data**



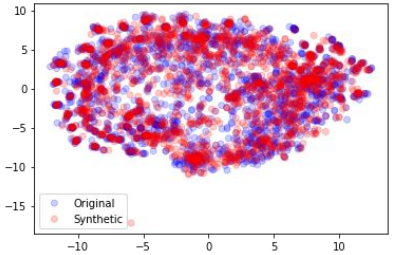

**MIMIC-III Mask data**



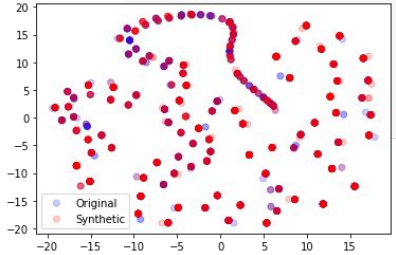

**MIMIC-III Static data**




**eICU Temporal data**




**eICU Mask data**



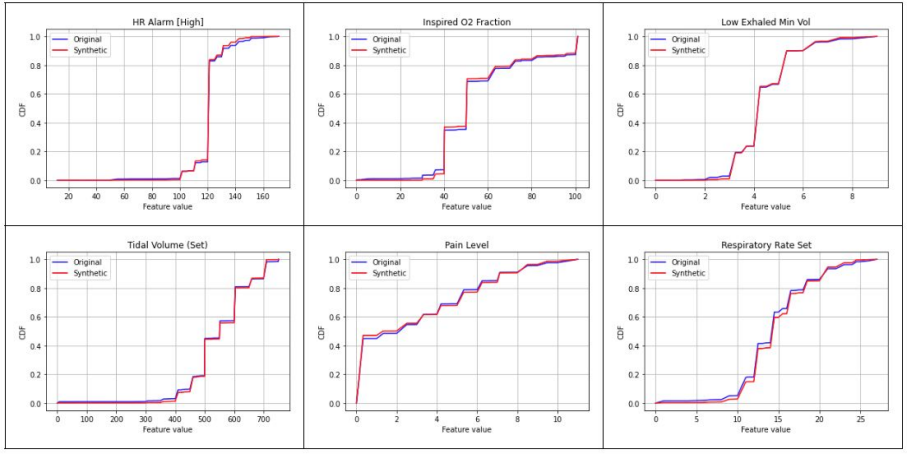

**eICU Static data**

NLP SUMMIT

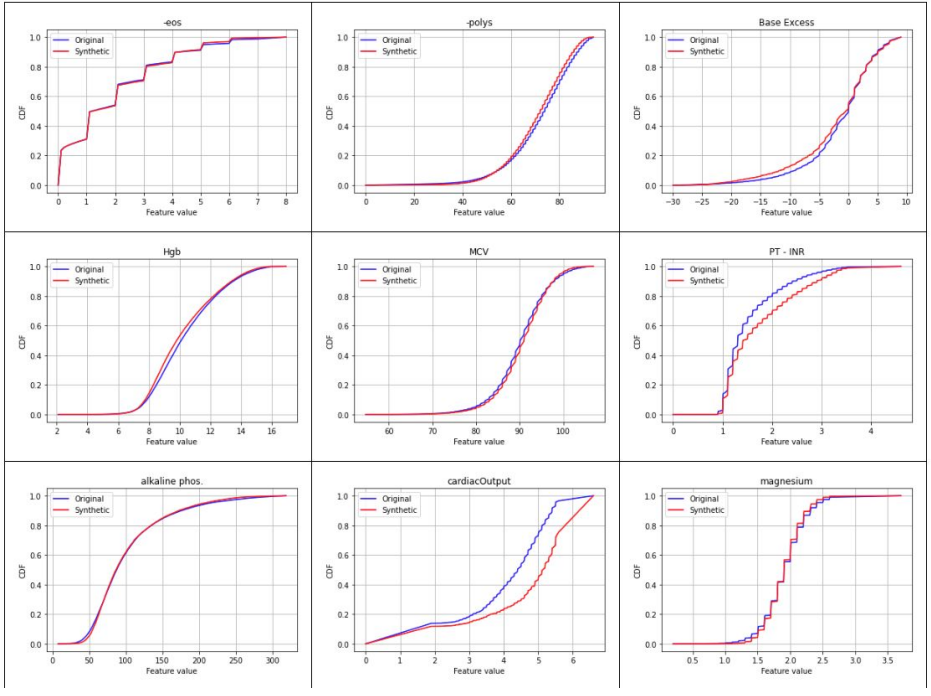
# Qualitative experiments - CDFs

- **Similar feature distributions** across **synthetic** & **real**

## MIMIC-III




## eICU




NLP SUMMIT

# Statistical Analyses

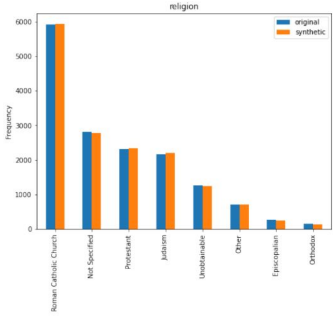
- **Similar feature statistics** across **synthetic** & **real**

**MIMIC-III Dataset**

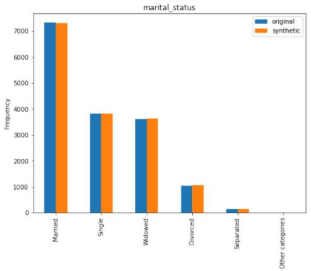
| Feature type | Feature name | Original data | | | Synthetic data | | | KS-Stat |
|---|---|---|---|---|---|---|---|---|
| | | Mean | Std | Miss rate | Mean | Std | Miss rate | |
| Temporal | Heart Rate | 82.56 | 17.34 | 36.53 | 82.20 | 15.91 | 35.74 | 0.01 |
| | Respiratory Rate | 18.85 | 5.31 | 37.88 | 18.29 | 4.54 | 36.72 | 0.04 |
| | calprevflg | 1.00 | 0.00 | 66.68 | 1.00 | 0.00 | 66.65 | 0.00 |
| | SpO2 | 97.30 | 3.38 | 67.99 | 97.39 | 2.27 | 67.39 | 0.03 |
| | O2 saturation pulseoxymetry | 96.89 | 3.12 | 70.63 | 96.97 | 2.41 | 70.00 | 0.02 |
| | NBP [Systolic] | 119.86 | 22.78 | 78.24 | 117.53 | 19.77 | 79.20 | 0.04 |
| | NBP [Diastolic] | 56.64 | 14.75 | 78.26 | 56.89 | 13.20 | 79.29 | 0.03 |
| | NBP Mean | 76.01 | 14.82 | 78.60 | 75.16 | 13.51 | 79.90 | 0.03 |
| | HR Alarm [Low] | 54.21 | 8.39 | 79.43 | 53.98 | 5.13 | 79.43 | 0.02 |
| | HR Alarm [High] | 120.28 | 11.86 | 79.48 | 120.15 | 8.75 | 79.44 | 0.01 |
| Static | Age | 91.33 | 67.41 | 0 | 93.05 | 70.15 | 0 | 0.02 |
| | Gender | 0.51 | 0.49 | 0 | 0.52 | 0.49 | 0 | 0.00 |
| | Mortality | 0.10 | 0.30 | 0 | 0.09 | 0.29 | 0 | 0.01 |

**eICU Dataset**

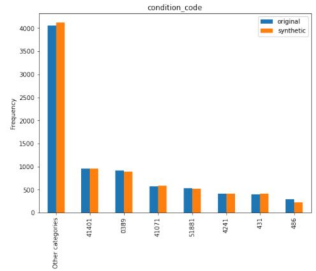
| Feature type | Feature name | Original data | | | Synthetic data | | | KS-Stats |
|---|---|---|---|---|---|---|---|---|
| | | Mean | Std | Miss rate (%) | Mean | Std | Miss rate (%) | |
| Temporal | nonInvasiveMean | 81.65 | 16.48 | 50.47 | 82.39 | 15.16 | 48.61 | 0.03 |
| | nonInvasiveSystolic | 121.97 | 22.62 | 50.57 | 121.79 | 20.60 | 48.62 | 0.02 |
| | nonInvasiveDiastolic | 65.34 | 14.59 | 50.57 | 65.80 | 13.02 | 48.67 | 0.03 |
| | bedside glucose | 150.86 | 59.10 | 81.44 | 149.28 | 49.85 | 84.62 | 0.04 |
| | potassium | 3.98 | 0.55 | 91.02 | 3.92 | 0.48 | 91.98 | 0.04 |
| | Hgb | 10.35 | 2.14 | 91.98 | 10.47 | 2.10 | 92.17 | 0.04 |
| | glucose | 130.45 | 48.72 | 91.98 | 132.15 | 47.56 | 92.26 | 0.03 |
| | sodium | 138.01 | 4.98 | 91.66 | 138.26 | 4.36 | 92.37 | 0.02 |
| | creatinine | 1.35 | 1.20 | 92.07 | 1.34 | 1.11 | 92.42 | 0.01 |
| | Hct | 31.49 | 6.19 | 92.10 | 31.76 | 6.06 | 92.43 | 0.03 |
| | BUN | 24.37 | 17.55 | 92.12 | 23.23 | 16.67 | 92.89 | 0.04 |
| | calcium | 8.42 | 0.71 | 92.43 | 8.39 | 0.70 | 92.66 | 0.03 |
| | bicarbonate | 25.44 | 4.81 | 92.46 | 25.21 | 4.31 | 93.02 | 0.03 |
| | platelets x 1000 | 215.19 | 104.12 | 92.74 | 207.75 | 94.41 | 93.30 | 0.02 |
| | WBC x 1000 | 10.39 | 4.83 | 92.81 | 10.00 | 4.16 | 93.53 | 0.02 |
| Static | Age | 63.05 | 17.07 | 0.00 | 64.25 | 16.82 | 0.00 | 0.03 |
| | Gender | 0.54 | 0.49 | 0.00 | 0.54 | 0.49 | 0.00 | 0.00 |
| | Mortality | 0.049 | 0.21 | 0.00 | 0.048 | 0.21 | 0.00 | 0.00 |



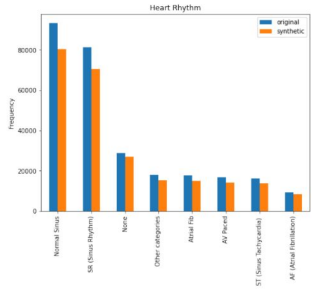

(a) Religion




(b) Marital status




(c) Medical code




(d) Heart Rhythm

NLP SUMMIT

# Fidelity metric - Train on Real / Synthetic

- Training on **synthetic** vs. **real** are **highly similar in terms of AUC**.
- On MIMIC-III, the best model (GBDT) on synthetic data is **2.6% worse** than the best model on real.
- On eICU, the best model (RF) on synthetic data is **0.9% worse** than the best model on real.

| Target | Models | Metrics | MIMIC-III | | eICU | |
|---|---|---|---|---|---|---|
| | | | Train on Real | Train on Synth | Train on Real | Train on Synth |
| Mortality | GBDT | AUC | **0.762** | **0.736** | 0.943 | 0.938 |
| | | AP | **0.304** | **0.261** | 0.600 | 0.534 |
| | RF | AUC | 0.723 | 0.710 | **0.954** | **0.945** |
| | | AP | 0.276 | 0.251 | **0.600** | **0.580** |
| | GRU | AUC | 0.728 | 0.667 | 0.937 | 0.938 |
| | | AP | 0.278 | 0.193 | 0.567 | 0.528 |
| | LR | AUC | 0.712 | 0.680 | 0.872 | 0.818 |
| | | AP | 0.233 | 0.207 | 0.323 | 0.260 |
| | Average | AUC | 0.731 | 0.689 | 0.926 | 0.909 |
| | | AP | 0.272 | 0.228 | 0.522 | 0.475 |

NLP SUMMIT

# Fidelity metric - Comparison with alternatives

- EHR-Safe **significantly outperforms alternative methods**
  - Alternatives are much **worse in addressing the real-world data challenges**

| Models | Fidelity | | | |
|---|---|---|---|---|
| | MIMIC-III | | eICU | |
| | AUC | AP | AUC | AP |
| Upper bound – using real data | 0.723 | 0.276 | 0.954 | 0.600 |
| EHR-Safe | 0.710 | 0.251 | 0.945 | 0.580 |
| TimeGAN | 0.576 | 0.147 | 0.726 | 0.241 |
| RC-GAN | 0.554 | 0.129 | 0.684 | 0.245 |
| C-RNN-GAN | 0.567 | 0.146 | 0.671 | 0.229 |

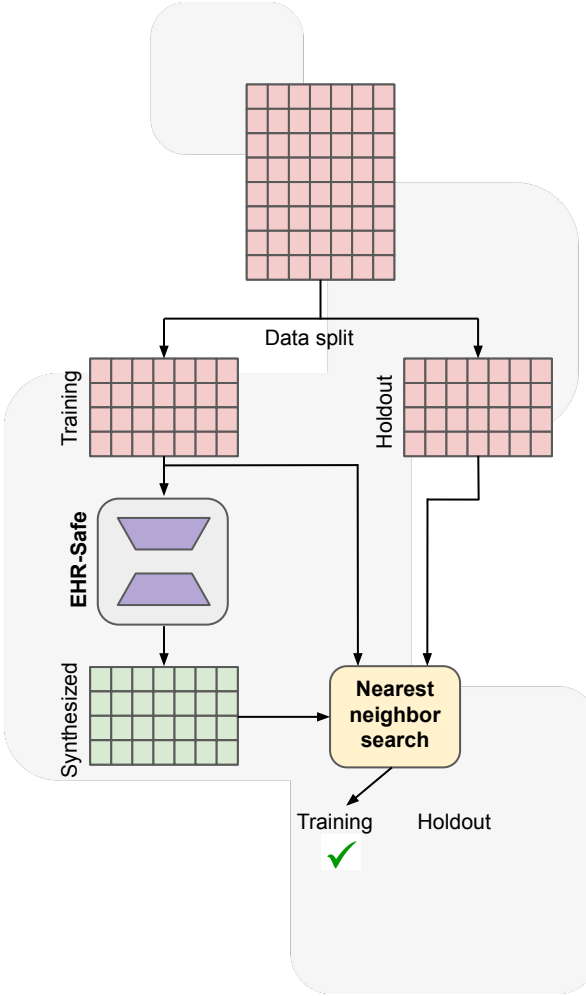
NLP SUMMIT

# Privacy metrics

- **Membership inference attacks:**
  - Finding out whether an adversary can understand a data is used for training the generative model.
  - **Ideal value** would be the same as random coin tossing (0.5)

| Datasets | **No privacy risk** | **EHR-Safe** |
|---|---|---|
| MIMIC-III | 0.5 | 0.496 |
| eICU | 0.5 | 0.489 |

**Near-perfect privacy preservation on this metric**




NLP SUMMIT

# Future work

- Extend EHR-Safe from tabular and time-series to **multi-modal datasets (text and/or image)**
- Demonstrations for **out-patient data (e.g. Insurance)**
- Incorporate **theoretically-guaranteed privacy (e.g., differential privacy)** on top

NLP SUMMIT


# NLP SUMMIT HEALTHCARE

Presented by

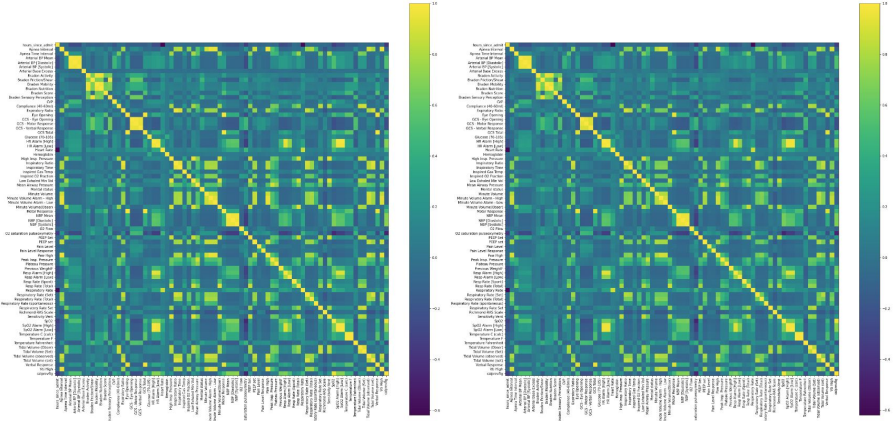
John Snow LABS

www.nlpsummit.org

# Qualitative experiments - Correlations

- Preserving **relationships across the features** is important for the fidelity of synthetic data.
- **Pairwise correlations** are well preserved in synthetic data:

**MIMIC-III**

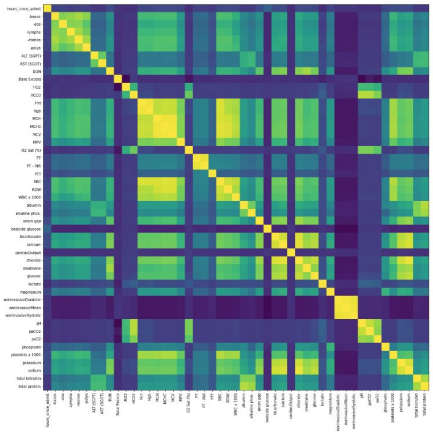


(a) MIMIC-III - Original data

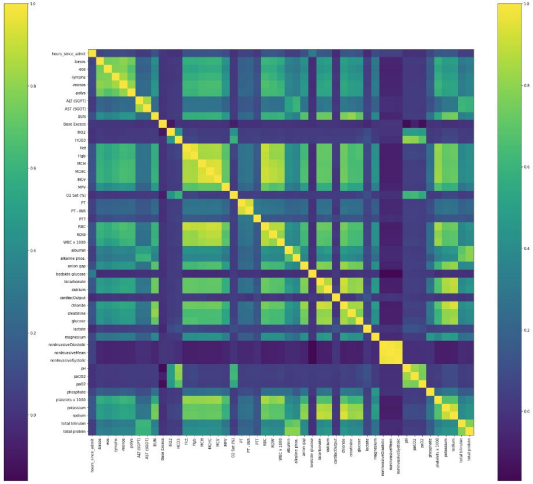


(b) MIMIC-III - Synthetic data

**eICU**



(c) eICU - Original data

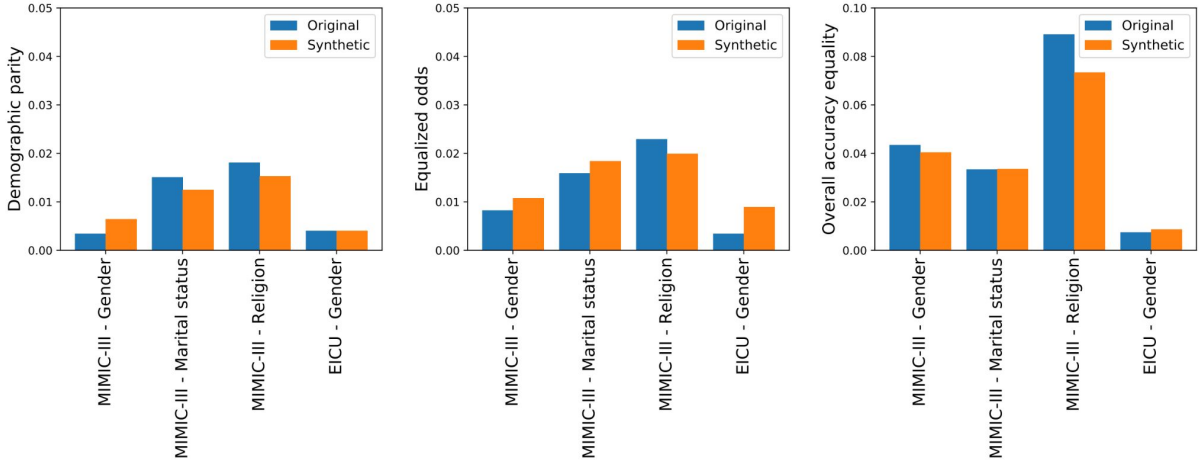


(d) eICU - Synthetic data

NLP SUMMIT

# Fidelity metric - Algorithmic fairness

- **No amplification of algorithmic bias** compared to original data




(a) Demographic parity (b) Equalized odds (c) Overall accuracy equality

NLP SUMMIT

# Privacy metrics

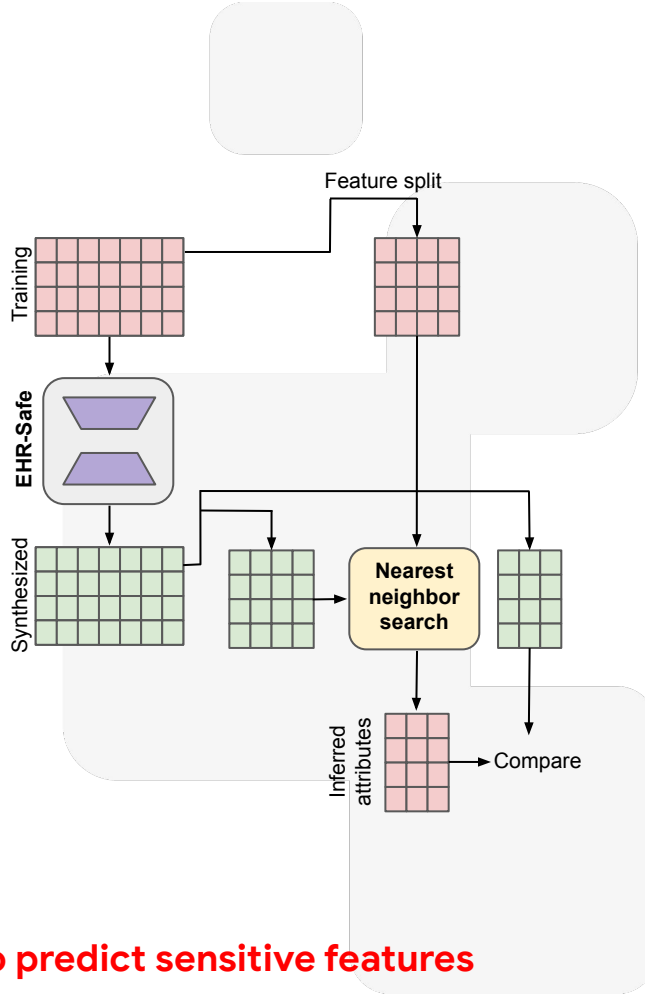
- **Attribute inference attacks:**
  - Whether the attacker can predict the sensitive features more accurately with synthetic data?

**MIMIC-III**

| | Specific attributes | With original data | EHR-Safe |
|---|---|---|---|
| **Attribute inference** | Gender | 0.696 | 0.681 |
| | Marital status | 0.628 | 0.620 |
| | Religion | 0.639 | 0.619 |

**eICU**

| | Specific attributes | With original data | EHR-Safe |
|---|---|---|---|
| **Attribute inference** | Gender | 0.678 | 0.669 |
| | Marital status | - | - |
| | Religion | - | - |

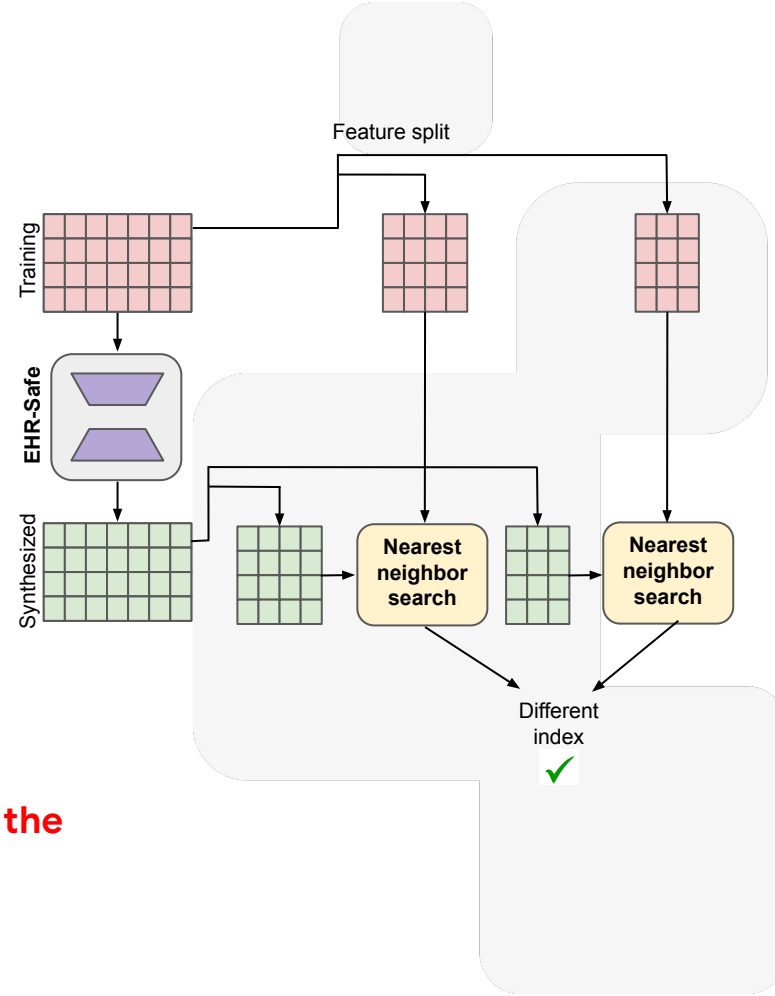



**Synthetic data does not increase ability to predict sensitive features**

# Privacy metrics

- **Re-identification ratios:**
  - If the partial information is given, can we identify the rest of the feature information?

| Datasets | No privacy risk | EHR-Safe |
|---|---|---|
| MIMIC-III | 0.049 | 0.061 |
| eICU | 0.068 | 0.085 |




**Only very marginally increased risk for identifying the information for the rest of the features**

NLP SUMMIT