Generating Synthetic EHR Data with Adversarial Networks

Andrew Yale

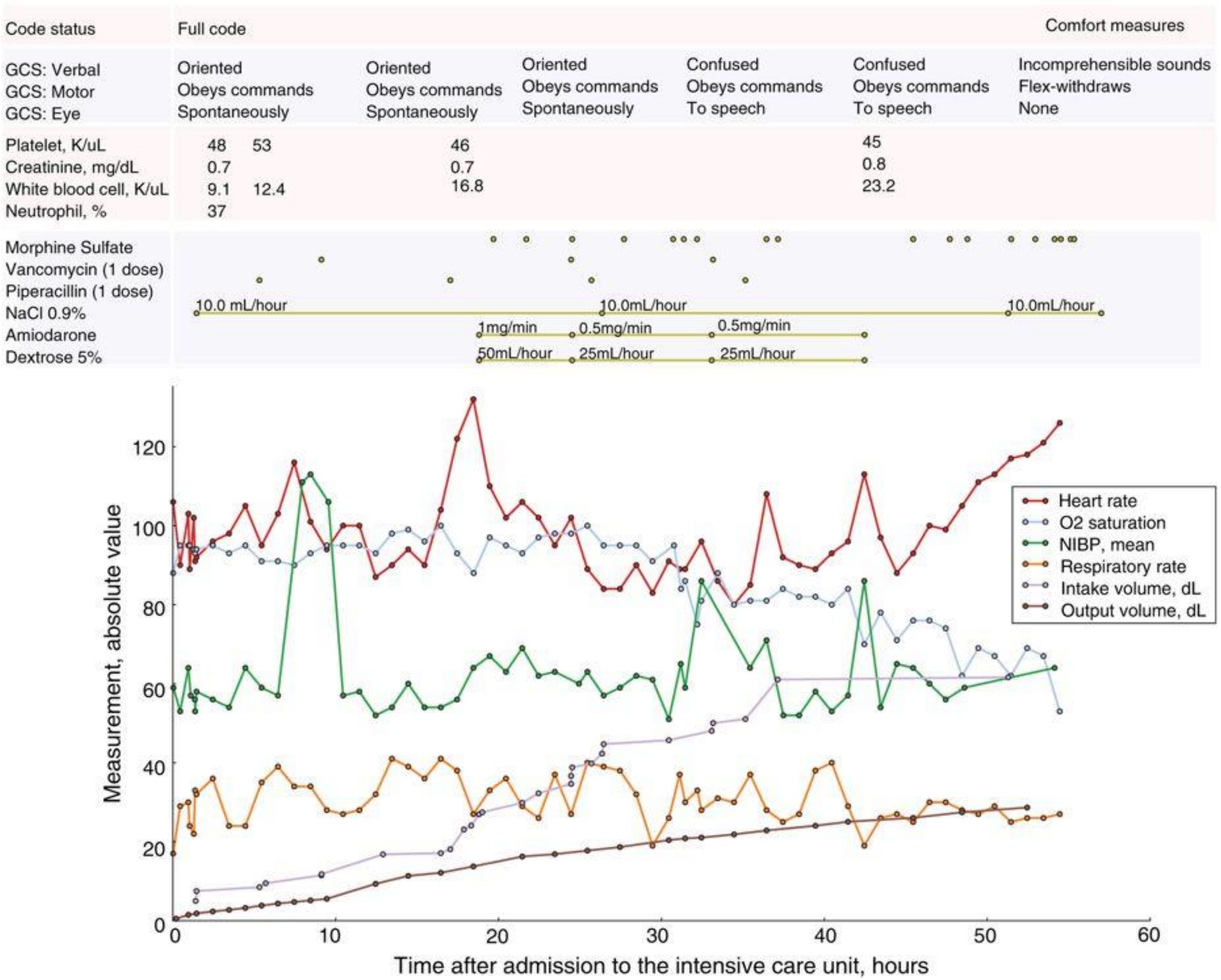
Phases of Development

- **Phase 1:** Synthesize patient level data directed towards a particular task
    - ICU Mortality Prediction using MIMIC data
    - Diabetic complications using OPTUMLabs data
- **Phase 2:** Synthesize longitudinal data for specific task
- **Phase 3:** Synthesize raw EHR data for any use

Goals of Synthetic Data

1. Create data that are similar to real data, and can be used in classes to teach students how to interact with medical data.
2. The data matches the real distributions sufficient enough to allow for workflows to be developed on the synthetic data and then brought into secure environments to test on real data to obtain results.
3. Create synthetic data that will be similar enough to the real data that models and results developed on synthetic data hold for the real data.

MIMIC

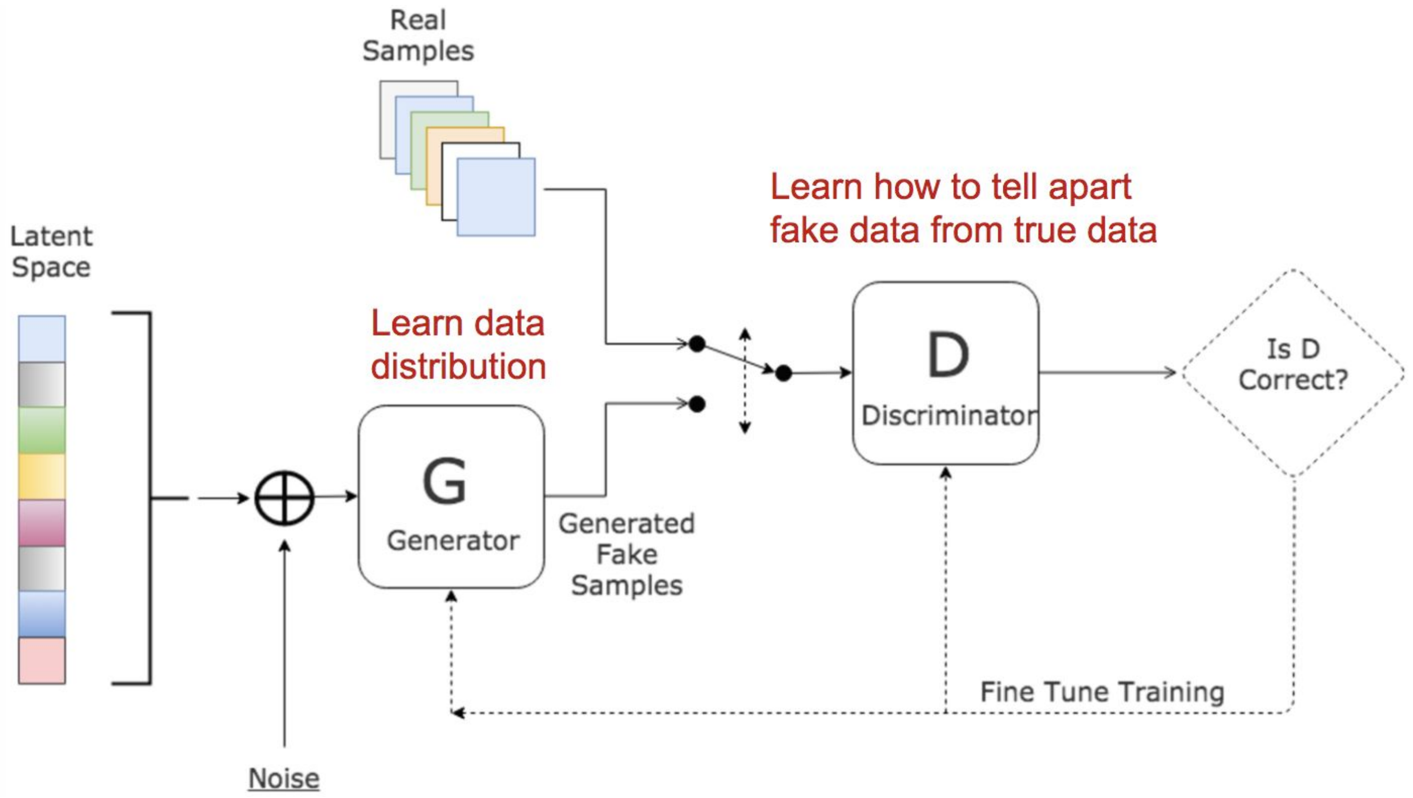


Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database."

MIMIC

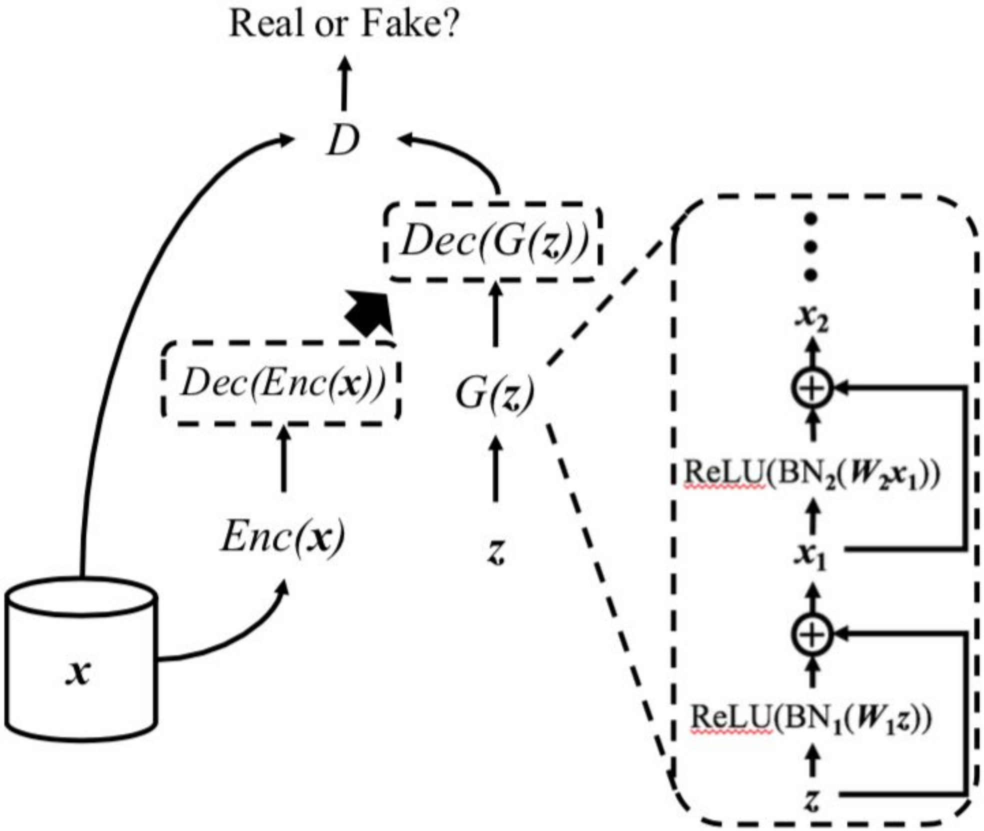
- Mixture of types of data
  - Time series measurements of vitals in the ICU
  - One time demographic information
  - Diagnoses and procedures corresponding to a specific hospital stay
- Converted to patient level features
  - First and second day average, minimum, and maximum for eight vital signs
  - Demographic data
  - ICD-9 diagnoses for that hospital stay
  - ICD-9 diagnoses rolled up into CCS codes

Generative Adversarial Networks (GANs)



MedGAN

- Created for MIMIC data
- Uses all binary data
- Adds an autoencoder
- Adds minibatch averaging
- Adds batch normalization

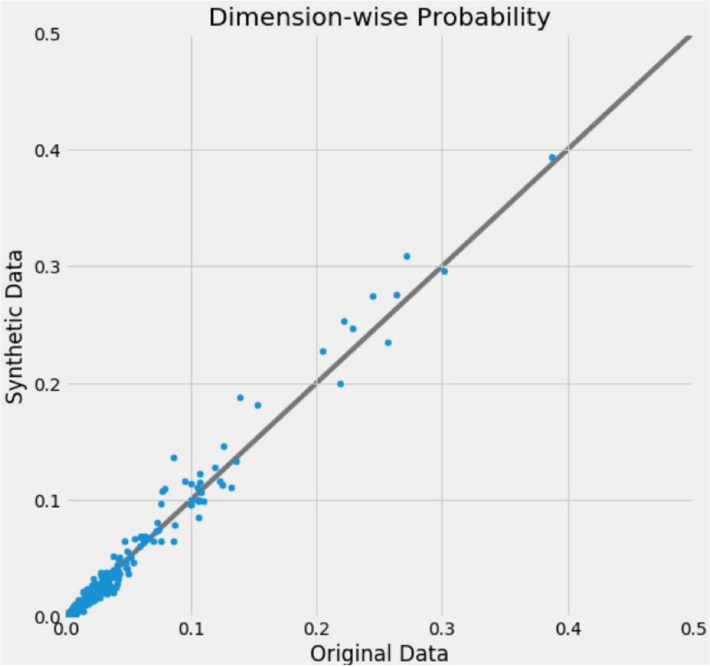


Real or Fake?
D
$Dec(G(z))$
$Dec(Enc(x))$
$G(z)$
$Enc(x)$
z
x
x_2
$\mathrm{ReLU}(\mathrm{BN}_2(W_2 x_1))$
x_1
$\mathrm{ReLU}(\mathrm{BN}_1(W_1 z))$
z

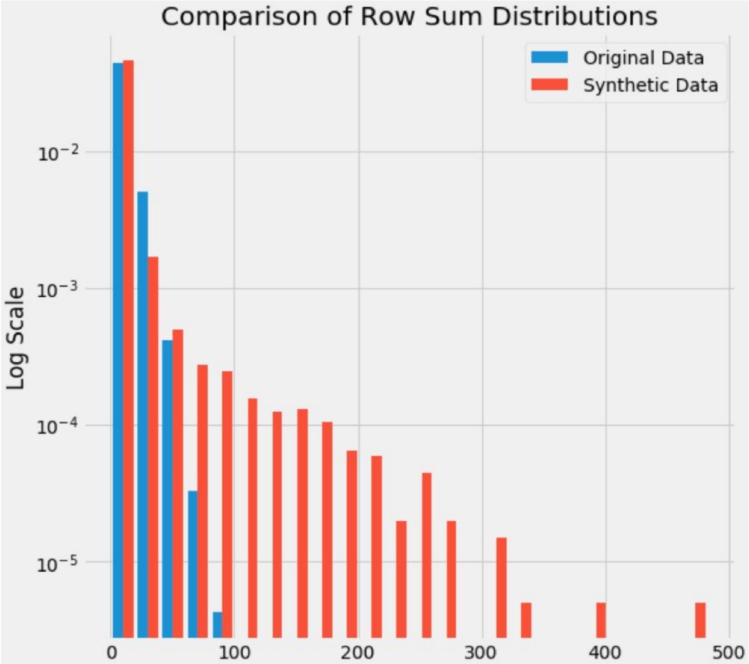
Choi, Edward, et al. "Generating multi-label discrete patient records using generative adversarial networks."

MedGAN Results

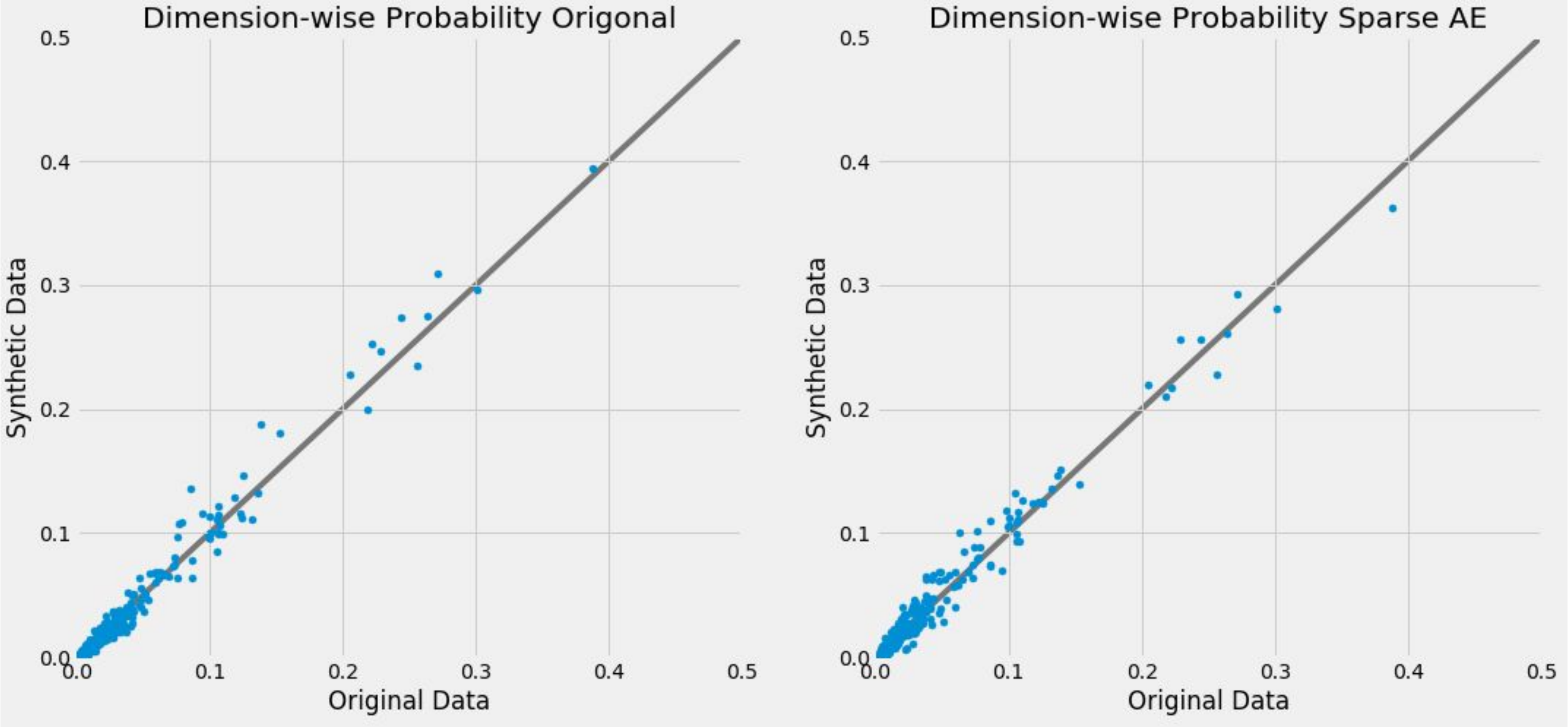
- Dimension-Wise Probability
  - Compare the column sparsity



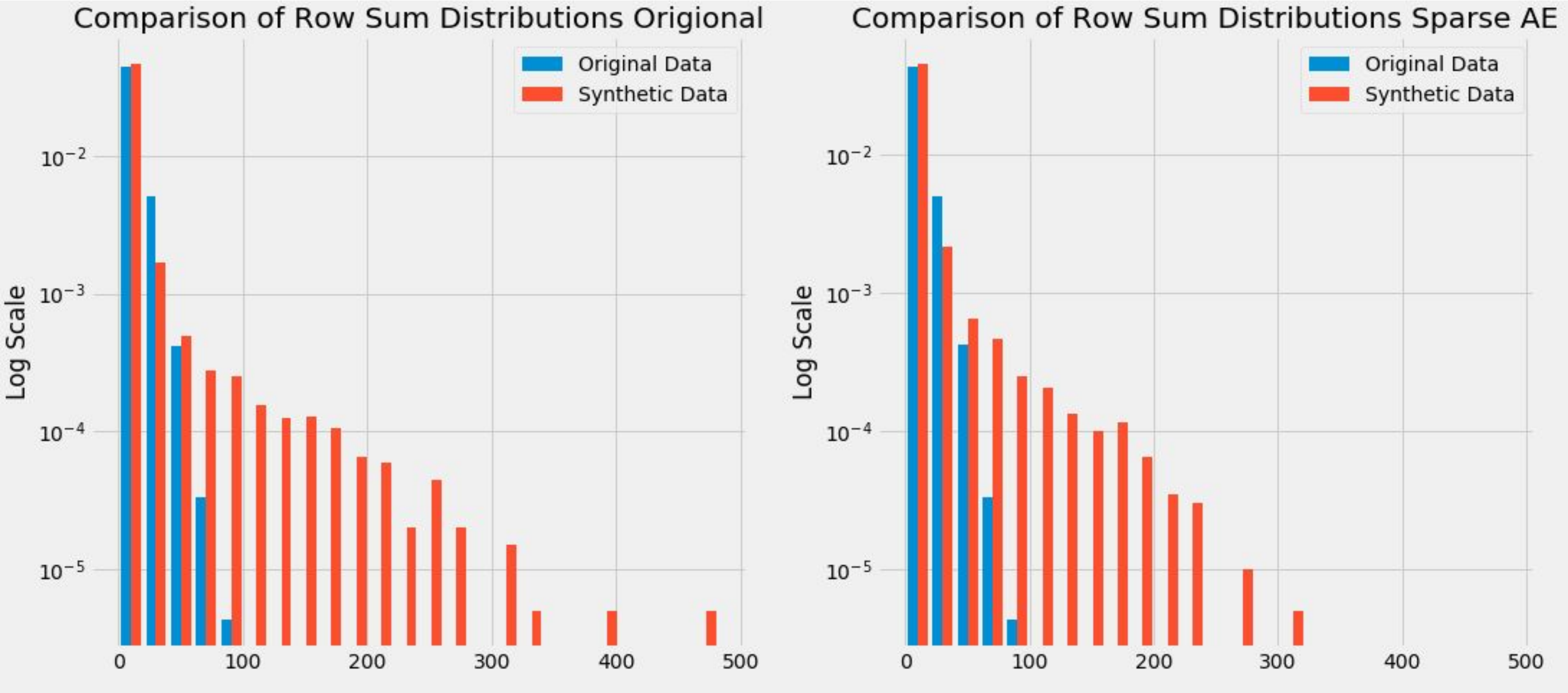
- Row Sum Distribution
  - Compare the sums of the rows



MedGAN Results

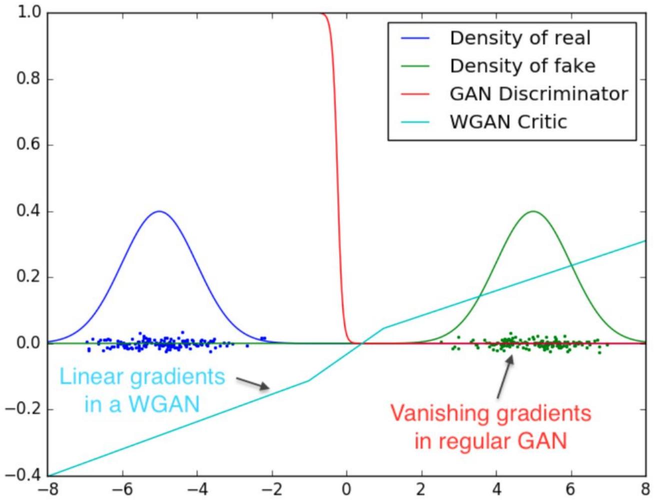


MedGAN Results



Wasserstein GAN

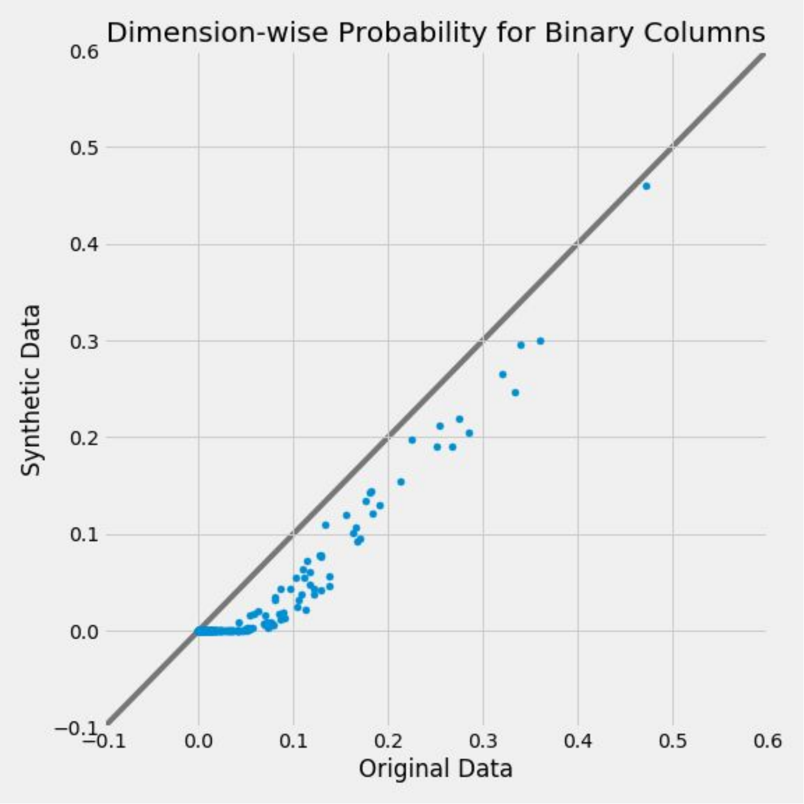
- Addresses flaws in the original GAN formulation
  - Mode collapse
  - Instability of the learning rate
  - Issues with KL divergence’s vanishing gradients

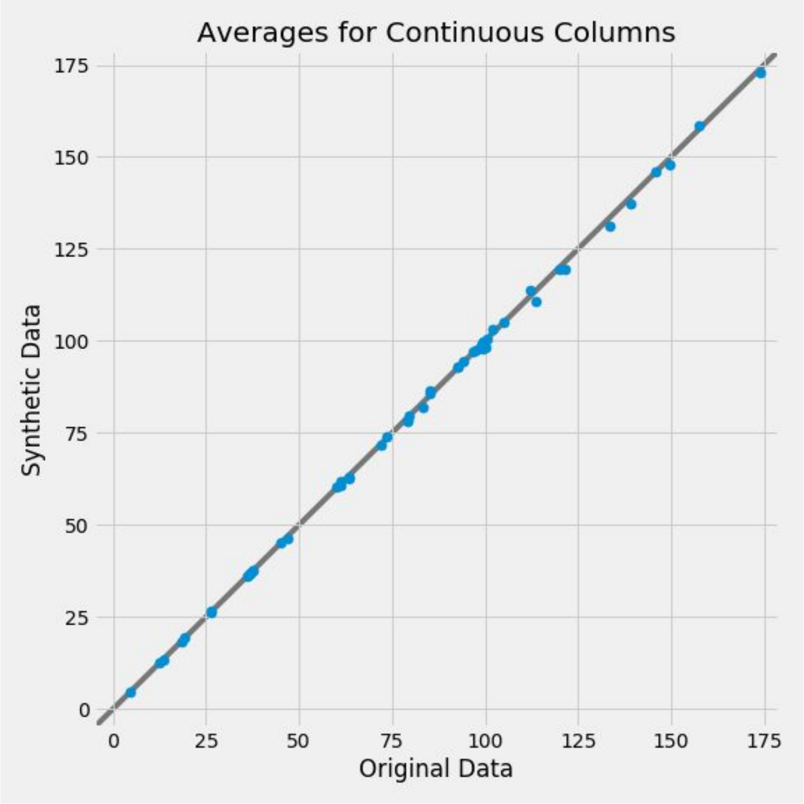


$$KL(\mathbb{P}_r\|\mathbb{P}_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x)d\mu(x)\ , \qquad W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma\in\Pi(\mathbb{P}_r,\mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\,\|x-y\|\,\big]\ ,$$

Arjovsky, Martin, et al. "Wasserstein gan."

WGAN Results





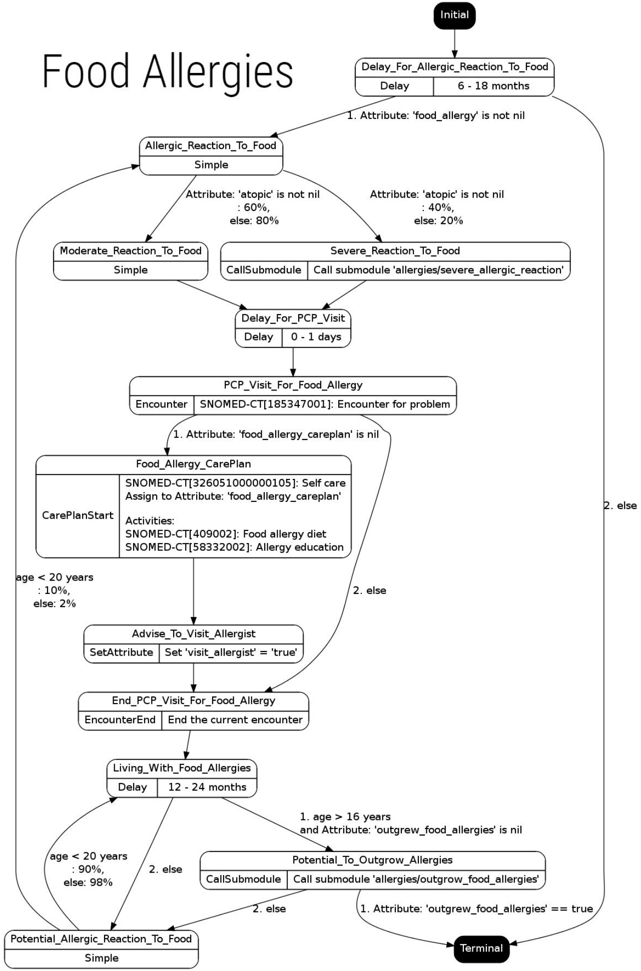
Synthetic Data Vault (SDV)

- Data preprocessing
  - Methods for representing NA values in data
  - Converts categorical data into values from 0 to 1
- Calculate representations for each column
  - Uses Gaussian copulas
  - Recursively calculated using the relationship between tables in a database
  - Use the representations to generate data

Patki, Neha, et al. "The synthetic data vault."

Synthea

- Models the progression and treatment of diseases
- Each patient is generated independently using the selected modules
- Modules can be created for any disease or condition
- Currently modelling Massachusetts residents



Food Allergies

Initial

Delay_For_Allergic_Reaction_To_Food
Delay | 6 - 18 months

1. Attribute: 'food_allergy' is not nil

2. else

Allergic_Reaction_To_Food
Simple

Attribute: 'atopic' is not nil : 60%, else: 80%

Attribute: 'atopic' is not nil : 40%, else: 20%

Moderate_Reaction_To_Food
Simple

Severe_Reaction_To_Food
CallSubmodule | Call submodule 'allergies/severe_allergic_reaction'

Delay_For_PCP_Visit
Delay | 0 - 1 days

PCP_Visit_For_Food_Allergy
Encounter | SNOMED-CT[185347001]: Encounter for problem

1. Attribute: 'food_allergy_careplan' is nil

2. else

Food_Allergy_CarePlan
CarePlanStart | SNOMED-CT[326051000000105]: Self care
Assign to Attribute: 'food_allergy_careplan'

Activities:
SNOMED-CT[409002]: Food allergy diet
SNOMED-CT[58332002]: Allergy education

age < 20 years : 10%, else: 2%

Advise_To_Visit_Allergist
SetAttribute | Set 'visit_allergist' = 'true'

End_PCP_Visit_For_Food_Allergy
EncounterEnd | End the current encounter

Living_With_Food_Allergies
Delay | 12 - 24 months

1. age > 16 years and Attribute: 'outgrew_food_allergies' is nil

age < 20 years : 90%, else: 98%

2. else

Potential_To_Outgrow_Allergies
CallSubmodule | Call submodule 'allergies/outgrow_food_allergies'

2. else

1. Attribute: 'outgrew_food_allergies' == true

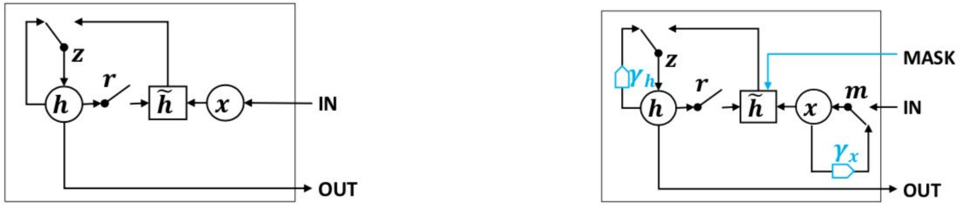
Potential_Allergic_Reaction_To_Food
Simple

Terminal

Walonoski, Jason, et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record."

RNN for Multivariate Time Series

- Created for MIMIC data
- Models time series data using recurrent layers such as GRU and GRU-D
- Handles the missingness in the MIMIC data



(a) GRU (b) GRU-D

Figure 3: Graphical illustrations of the original GRU (left) and the proposed GRU-D (right) models.

Che, Zhengping, et al. "Recurrent neural networks for multivariate time series with missing values."

Next Steps

- This week we will be starting our OPTUM access
- Create the precisely defined cohort in the OPTUM environment
- Adapt the final WGAN formulation to the OPTUM dataset
- Define the correct hyperparameters for the OPTUM dataset