

# Bridging Open Science and Privacy Protection in research with Synthetic HIV Cohort Data

IWHOD 2024

Zhuohui Liang, Chao Yan, Yanink Caro-Vega, Peter F. Rebeiro, Bradley A. Malin,  
Stephany N. Duda, Bryan E. Shepherd

Department of Biostatistics, Vanderbilt University School of Medicine



VANDERBILT  
UNIVERSITY

# Introduction

Hypothetically, you found a procedure  $X$  is associated with survival with a small HIV cohort in Mexico, now, you want to generalize this finding to a larger cohort.

## How difficult it is to get a hand on the data?

- Application, Concept sheet, Approval waiting time...
- CITI, HIPAA training, Data center training
- Access the data only in facility or use VPN
- Rules of storing and sending the data



VANDERBILT  
UNIVERSITY

# Introduction

Journal or funding agency request the data you use for analysis as supplements:

## What about publishing the data?

- Re-identity risk/ Privacy concern/ Patient stigmatized
- HIPAA, General Data Protection Regulation (GDPR)
- Data sharing agreement
- Loss of information when de-identifying and adding noise



VANDERBILT  
UNIVERSITY

# Introduction

Currently, privacy regulation is a gap for open science

## How difficult it is to get a hand on the data?

- Application
- CITI, HIPPA training, Data center training
- Access the data only in facility or use VPN
- Rules of storing and sending the data

## What about publishing the data?

- Re-identity risk/ Privacy concern/ Patient stigmatized
- HIPAA, General Data Protection Regulation (GDPR)
- Data sharing agreement
- Loss of information when de-identifying and adding noise

... What can we do?



VANDERBILT  
UNIVERSITY

# Synthetic Data

## Synthetic data with Machine Learning

- No real record/patient is in the data, but it reflects the true data
- Since it is not real, it provides a workaround of current regulations

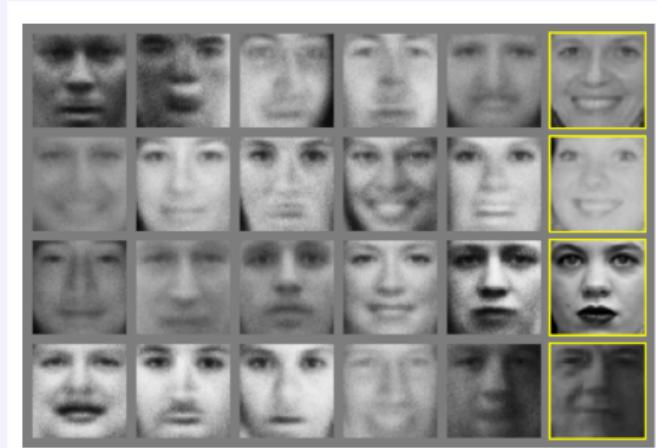


Figure 1: Goodfellow, Ian, et al. “Generative adversarial nets.” Advances in neural information processing systems 27 (2014).



VANDERBILT  
UNIVERSITY

# Synthetic Data

## Synthetic data with Machine Learning

- **No real record/patient is in the data**, but it reflects the true data
- **Since it is not real, it provide a workaround of current regulations**
- Minimum assumptions or constraints with respect to the associations between variables
- And there has been example that using both real and synthetic data together can increase accuracy
- There are already freely available synthetic database for research ie. The Synthetic Data Vault

VANDERBILT  
UNIVERSITY

# Synthetic Data

To Recap:

- No real patient information is in it
- It has most of the information in the real data

**So, can we use synthetic data to bridge the gap of open science and privacy protection?**



VANDERBILT  
UNIVERSITY

# Introduction



- Caribbean, Central and South America Network for HIV Epidemiology (CCASA net)
- 59208 participants from Argentina, Brazil, Chile, Haiti, Honduras, Mexico, Peru
- **To develop a synthetic dataset for the CCASA net that replicates the original dataset's structure and variable associations, facilitating its public use for**
  - Reproducible research and
  - **Hypothesis generation**

# Method



VANDERBILT  
UNIVERSITY

# Generative Adversarial Network (GAN) Model

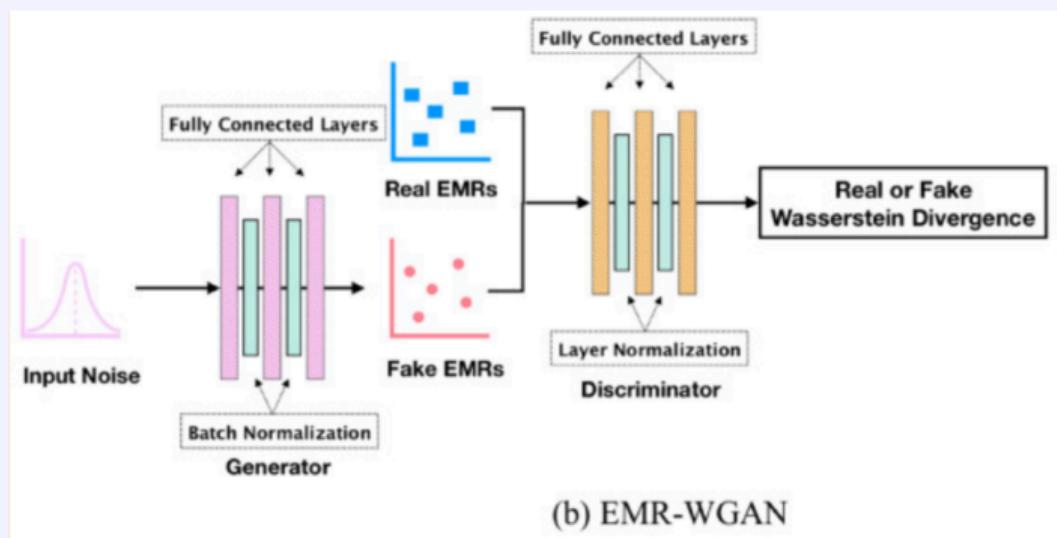
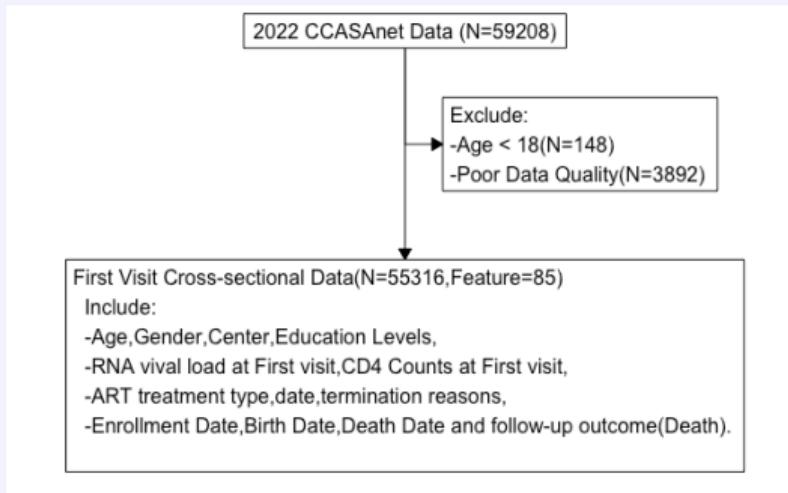


Figure 2: Zhang, Ziqi et al. JAMIA vol. 27,1 (2020)

- 1 A **Generator** is trained on the real data to **generate similar records**
- 2 A **Discriminator** is trained to **identify fake/generated records**
- 3 Generator **competes** with Discriminator until Discriminator cannot identify the fake data



# Training Data



- We use CCASAnet dataset enrolled before 2022.
- Need to transform into Machine Learning Format
- There is a detailed tutorial and code by Chao Yan available online

Figure 3: Training Data Set



VANDERBILT  
UNIVERSITY

# Evaluation

- **Utility:** How well the synthetic data reflect the real data: We will look into data distribution and associations
- **Privacy:** How well the synthetic data protect the real data: We will use two common used metrics to evaluating risk of information leakage and re-identification
- **Real Application:** How well the synthetic data can be used in real analysis

*Mortality and loss to follow-up among HIV-infected persons on long-term antiretroviral therapy in Latin America and the Caribbean*



VANDERBILT  
UNIVERSITY

# Result



VANDERBILT  
UNIVERSITY

# Data Utilities

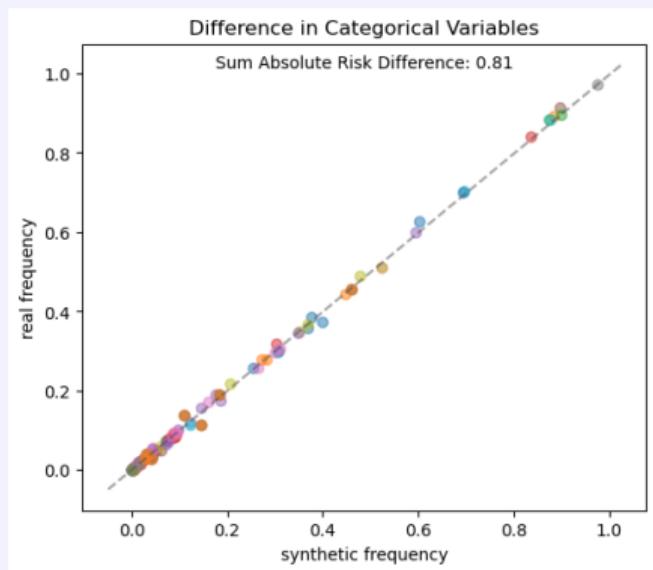


Figure 4: Absolute Difference in Prevalence of Catagorical Variables

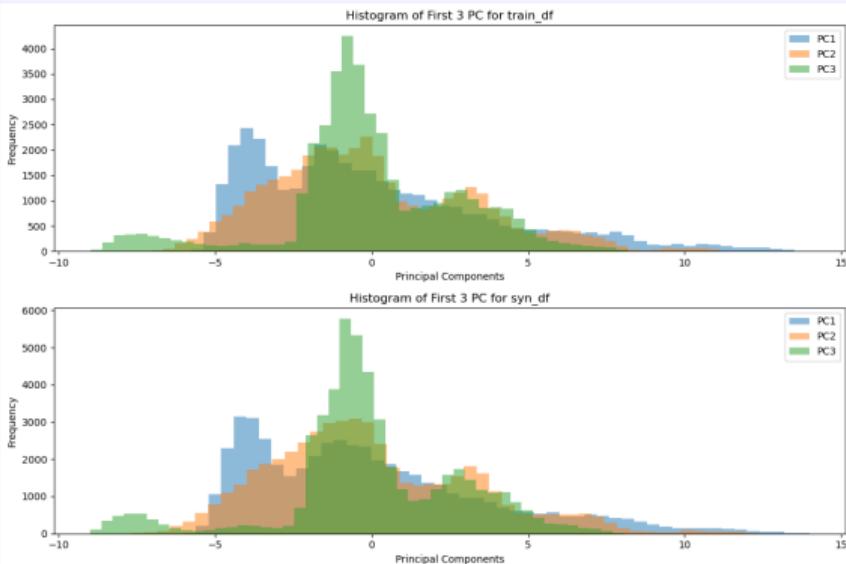


Figure 5: Principle Component

# Data Privacy

- Attribution Risk: How likely unknown attribute values of interest are predicted from a set of leaked attribute values with a trained model from the fake data
- Membership Risk: How likely a real record can be reconstruct from a leak generative model

Table 1: Privacy Risk

	Attribution Risk	Membership Risk
Synthetic	0.32	0.37
Real	0.51	0.83



VANDERBILT  
UNIVERSITY

# Study Reproducibility

Cox proportional hazards models (Time to Death) were stratified by CCASAnet site and adjust for:

sex, calendar year, age, CD4 count, clinical AIDS at ART initiation and ART regimen class

Variable	HR-Synthetic	HR-Training
100*CD4	0.902 (0.885-0.919)	0.856 (0.842-0.869)
10*Enrollment Year	0.793 (0.745-0.844)	0.710 (0.674-0.748)
10*Age	1.193 (1.155-1.232)	1.241 (1.206-1.277)
AID Enrollment	1.816 (1.678-1.967)	1.783 (1.671-1.903)
Other vs nnrti	2.336 (2.129-2.564)	1.857 (1.702-2.026)
Pi vs nnrti	1.029 (0.905-1.172)	1.070 (0.958-1.196)



# Study Reproducibility

- We modeled the first visit data, and use patient profile to predict their survival(time-to-event), this imposes the assumption that the survival is only related the first visit.

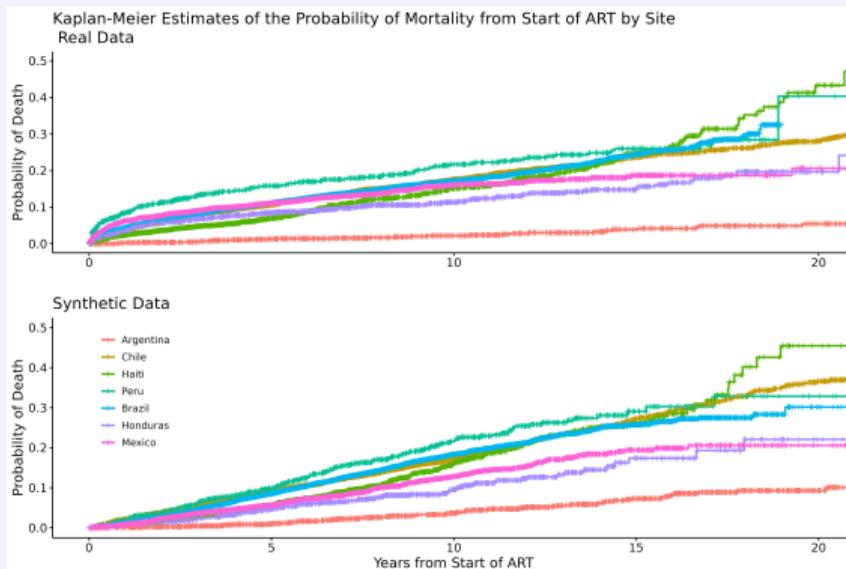


Figure 6: K-M estimate of Survival(Top: Real vs Bottom: Synthetic Data(best performance))

# Roadmap



VANDERBILT  
UNIVERSITY

# Adapt to longitudinal models

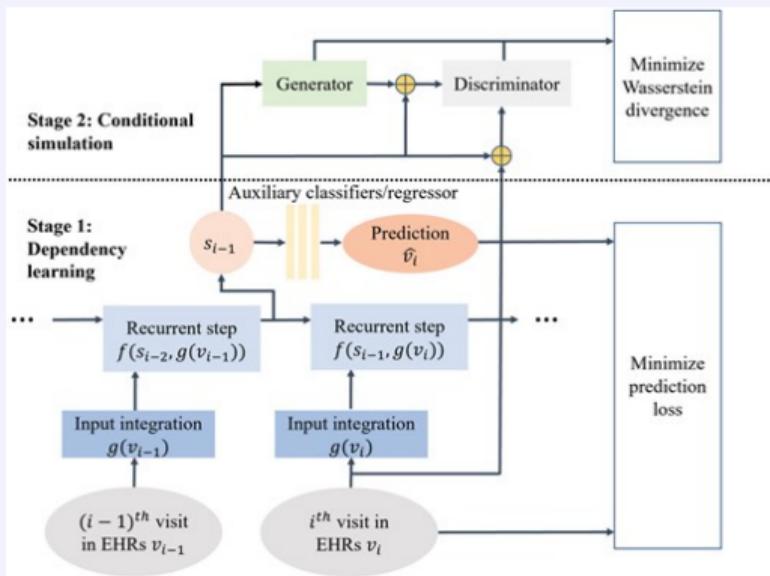


Figure 7: Zhang, Ziqi, et al. JAMIA 28.3 (2021): 596-604.



VANDERBILT  
UNIVERSITY

## Future works

- Missingness is an attribute in EHR data, we want our synthetic data to reflect the missing pattern in the real data when publish
- So far, because of the limitation of algorithm and computation time, we use only a single imputation method to prepare the data
- Rare Event ( $p<0.001$ ) cannot be properly captured by the model
- It is encouraged to check the data for abnormalities post-generation (ie. male record with Cervical cancer)
- Further improve privacy protection
- Develop a evaluation method that shows confidence of any hypothesis generated from the synthetic data

VANDERBILT  
UNIVERSITY

# Summary

- We developed a version of synthetic dataset for the CCASAnet that replicates the original dataset's structure and variable associations
- It allows researchers to use the data to reproduce the study, identify problems, and produce more results
- It shows a path forward for open science, but one must also keep in mind the quality and privacy risk of the data when published and used.

VANDERBILT  
UNIVERSITY

## Other Methods

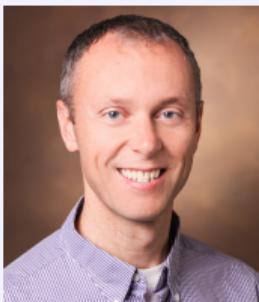
We chose the GAN model because our team has experts in this field and it (SynTEG) is used as a comparison standard in most recent papers.

We want to point out that other methods are available and produce similar or better results depending on the characteristics of the data.

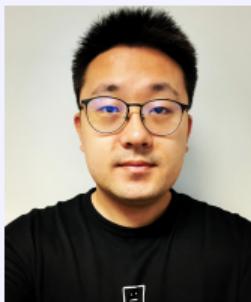
- 1 Theodorou, B., Xiao, C. & Sun, J. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. Nat Commun 14, 5305 (2023). <https://doi.org/10.1038/s41467-023-41093-0>
- 2 Clinical knowledge base simulation
- 3 Conditional Model for research specific data



# QUESTION?



Bryan  
E.Shepherd, PhD



Chao Yan, PhD



Stephany  
N.Duda, PhD



Bradley Malin,  
PhD

**Thank you**

NIH and CCASAnet fund this projects

Include a link to project here



VANDERBILT  
UNIVERSITY