Stanford University | Center for Population Health Sciences

The IDEA Centre Launch
Synthetic Data: Why and How

Isabella Chu, MPH
Ayin Vala, MS
April 20, 2023

The Rationale for Synthetic Data

Stanford University | Center for Population Health Sciences

Isabella Chu, MPH
Associate Director, Data Core
Stanford Center for Population Health Sciences
Stanford University

PHS Data Core Mission

The mission of the PHS Data Core is to build a data ecosystem to make high-value health and behavioral data available to researchers easily, efficiently and safely.

BiB/Stanford Partnership

We concluded that the most effective path for long term transatlantic collaboration is through data science.

Stanford PHS ecosystem: secure compute environment, data managers, real world healthcare data and researchers

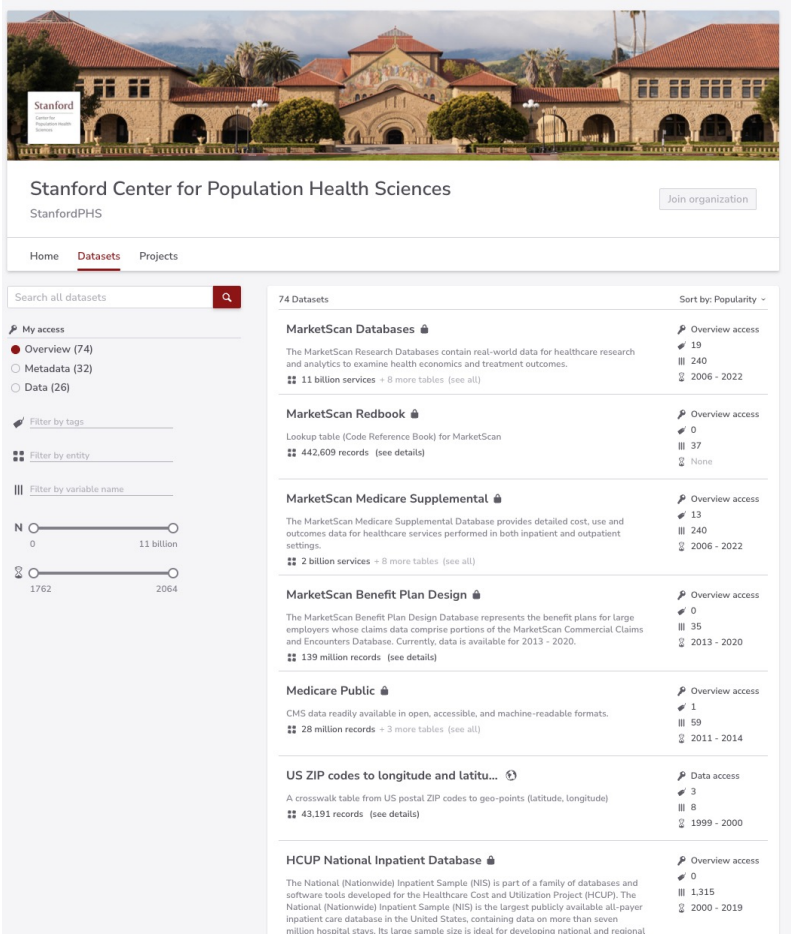
Stanford University | Center for Population Health Sciences

1. Secure Compute
    - Built over last 7 years using a grant from the Sloan Foundation
    - Built a secure academic data core on Google Cloud
    - Provides tools to analyze data at scale
    - Ability to run models on high risk data while developed on low risk data

2. Professionalize Data Management

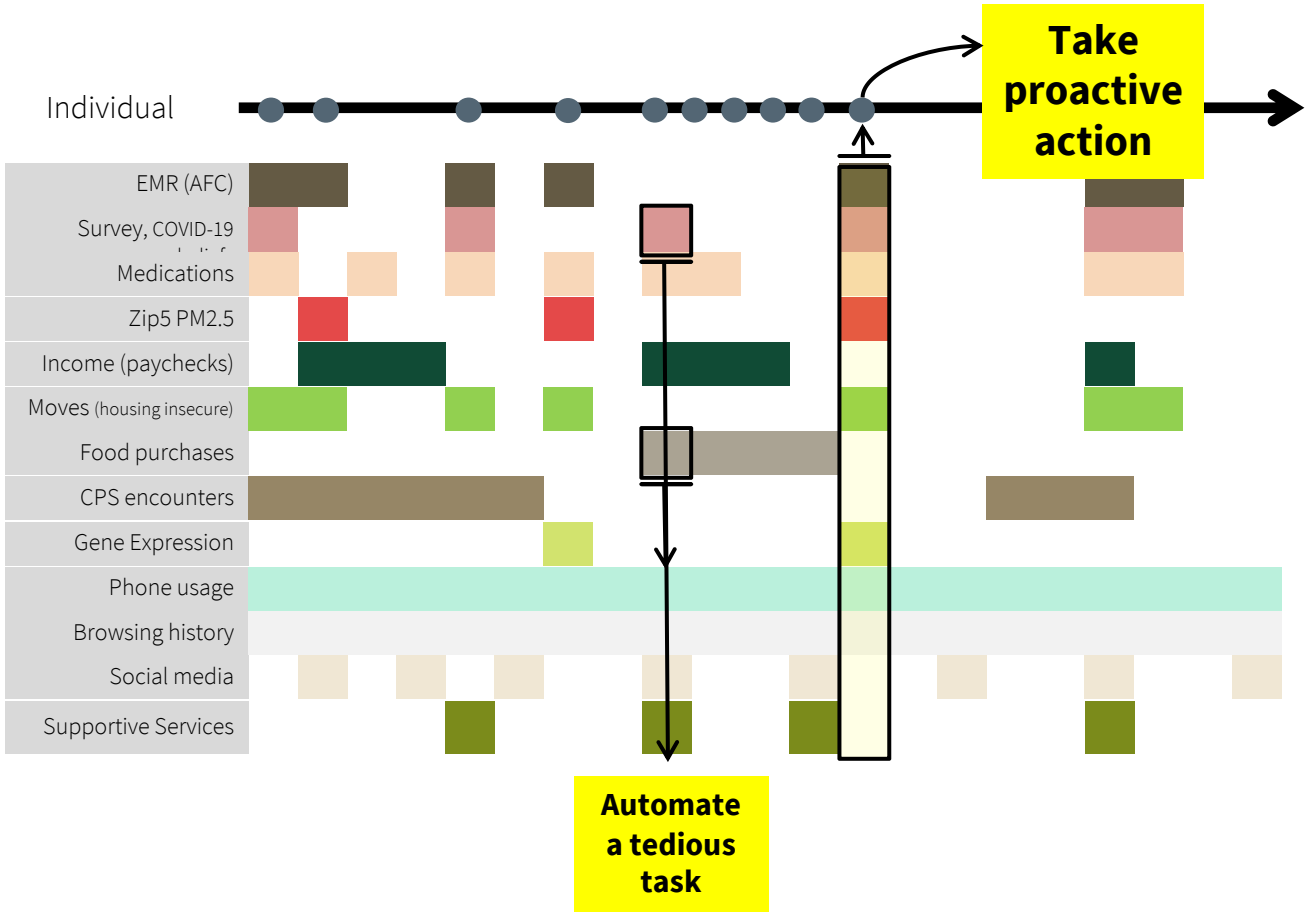
3. Data Hub
    - 74 Datasets
    - 80 Billion health records

4. Researchers
    - Serving +1600 members of Stanford community and beyond
    - Hosting +3000 projects led by students and faculty



Data utility for population health research: Large, Longitudinal and Linkable

Stanford University | Center for Population Health Sciences



Selected PHS Data Sources

Data Source	Sampling Frame	~N	Data Type	Smallest Geographic Unit	Years Available via PHS Data Portal	Time to Clean/Curate Data Updates*
American Family Cohort (AFC)	National, rural	~8.2 M	EMR	Census Block	2010 - Present	< 3 months
MarketScan	National, commercial & Medicare Part D insurance	183 M	Claims	Metropolitan Statistical Area	2007 – 2021 (through Q2 of 2022 coming soon)	~6 months
Medicaid	National, low income	(2011) 68.6 M (2015) 92.5 M (2018) 93.2 M	Claims	Census tract	2011, 2015, 2018**	~18 months
Medicare 20%	National, 65+	18.1 M	Claims	Census tract	2006 - 2020	~18 months

*Updates: Time between data generation (e.g., doctor visit), receipt, and completion of ETL, cleaning, validation & upload.
** Will have 2011 – 2019 by summer 2023.

Challenges in Sharing and Collaborating

1. Access to rich Healthcare data is appropriately limited to select researchers and cumbersome to get access

2. Linking datasets increases the risk of reidentification

3. Difficult (impossible) to share rich healthcare data across nations and often organizations.

The Solution: Synthetic Data

High-fidelity synthetic data retains the properties and behaviors of the original data and is indistinguishable from its substrate with respect to variables, summary statistics and behavior in statistical models.

Synthetic data are not subject to HIPAA/GDPR rules as they do not contain information about any real person. They can be safely shared.

Synthetic Data: The Utility without the Risk

Stanford University | Center for Population Health Sciences

Research Identifiable Files (RIFs)	Limited Data Set (LDS)	De-identified	Synthetic
• Includes PHI and identifiable individuals, payer, providers.	• Removed PHI • Includes variables such as 5 digit zip code, DoB, etc.	• Many variables and records removed • Statistically de-identified	• Not subject to HIPAA/GDPR
Available to • U.S. based researchers	Available to • U.S. based researchers	Available to • U.S. based researchers • Foreign Institution researchers	Available to • U.S. based researchers • Foreign Institution researchers

The Rationale for Synthetic Data

Stanford University | Center for Population Health Sciences

Ayin Vala, MS
Associate Director, Cloud Computing and Data Science
Stanford Center for Population Health Sciences
Stanford University

The Rationale for Synthetic Data: Tension between researchers and data custodians/owners

- Get access to data quickly to gauge the value of the data
- Moral imperative to make the highest and best use of data
- Open science and spirit of discovery

- GDPR/HIPAA rules
- Optics/Public perception
- Costs of secure systems
- Risk aversion
- Data owner's business interests

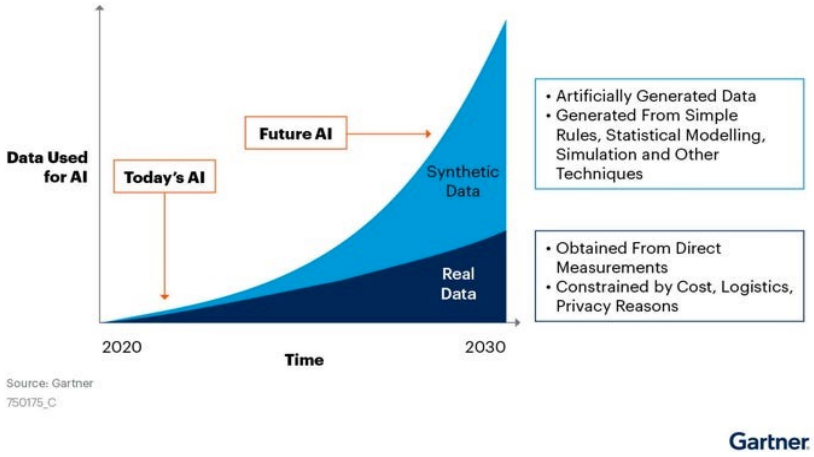
Why Now?
Rise of Synthetic Data Usage

Stanford University | Center for Population Health Sciences

1. Compute
Secure and scalable cloud environments.
i.e. AWS, GCP, Azure

2. Generative AI Algorithms
GANs, Transformers, and Diffusers making synthetic data. i.e. Dall-E, chatGPT

3. Democratization of Data Science
Much wider range of researchers need access to data



https://www.gartner.com/en/newsroom/press-releases/2022-08-22-gartner-data-analytics-summit-2022-orlando-day-1-highlights

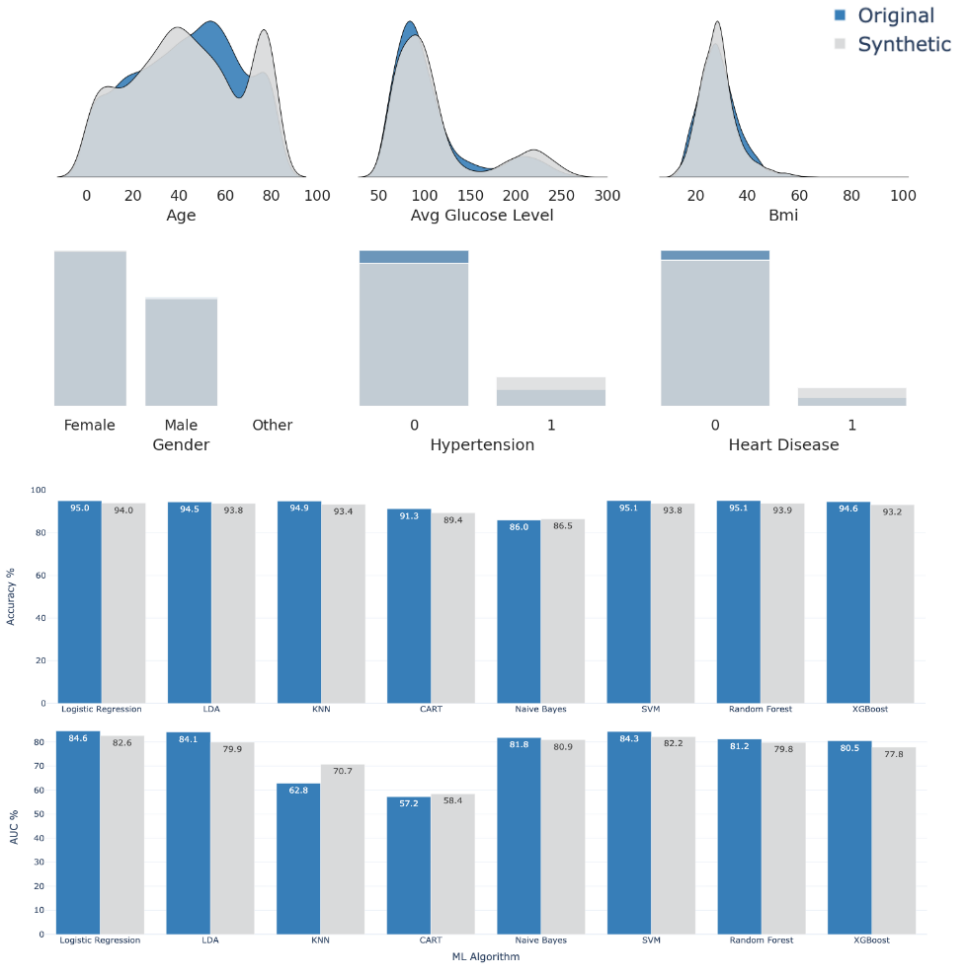
Example

Stanford University | Center for Population Health Sciences

Example: Stroke Prediction Dataset*

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
36922	Male	56.0	0	0	Yes	Private	Rural	62.68	18.4	never smoked	0
65258	Male	53.0	0	0	Yes	Private	Urban	86.73	26.1	Unknown	0
32445	Female	78.0	0	0	Yes	Self-employed	Urban	79.55	21.1	formerly smoked	0
68438	Female	51.0	0	0	Yes	Private	Rural	90.78	32.3	never smoked	0
57419	Male	59.0	0	0	Yes	Private	Rural	96.16	44.1	Unknown	1
68089	Female	44.0	0	0	Yes	Private	Urban	121.46	40.4	Unknown	0
51809	Female	60.0	0	0	Yes	Self-employed	Rural	103.17	32.1	formerly smoked	0
72132	Male	16.0	0	0	No	children	Urban	102.30	21.9	Unknown	0
44010	Female	3.0	0	0	No	children	Urban	57.33	16.8	Unknown	0
37349	Female	61.0	0	0	Yes	Private	Rural	123.36	33.4	never smoked	0

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
27296	Female	2.36	0	0	No	children	Rural	109.82	21.3	Unknown	0
39470	Male	29.62	0	0	Yes	Private	Rural	223.77	36.4	never smoked	0
7601	Female	53.33	0	0	Yes	Private	Rural	76.98	22.6	formerly smoked	0
31163	Male	2.07	0	0	No	children	Rural	87.45	25.8	Unknown	0
32784	Male	9.14	0	0	No	children	Rural	159.00	30.3	never smoked	0
8374	Female	14.06	0	0	No	Private	Urban	119.78	30.7	Unknown	0
15418	Male	27.99	0	0	No	Private	Rural	98.07	NaN	Unknown	0
49135	Female	20.83	0	0	Yes	Private	Rural	71.51	30.3	formerly smoked	0
15093	Male	31.19	0	0	Yes	Private	Urban	77.12	36.9	never smoked	0
3000	Female	13.25	1	0	No	Private	Rural	78.54	20.5	never smoked	0



*Original dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Benchmarking Synthetic Data

Data summary statistics:

- Assessing correlation stability between fields in real data and synthetic data
- Distribution stability for each variable, and deep structure stability comparing data holistically using methods such as PCA.

Synthetic quality score: recommending the appropriate level and application for the synthetic data

- To simulate a test environment to sharing synthetic data more freely
- Using the synthetic data to improve previous model development results
- Synthetic data standalone for analysis

Privacy level recommendation: to assess privacy risk associated with the synthetic data

- Measure the re-identification risk using metrics like record linkage, attribute linkage, and k-anonymity.
- Add privacy-inducing algorithms such as outlier filtering, similarity filter (where the synthetic record is removed if too similar to any synthetic real records)
- Differential Privacy using DG-SGD [Dwork et al, 2006] with noise addition and clipping.

Current and Near Future Work

- Synthetic version of our datasets
- Partners with data willing to collaborate
- Researchers to validate and use synthetic data for use cases such as:
  - Infection Disease i.e. COVID-19
  - Mental Health i.e. Adolescent mental health in US
  - Climate change i.e. CA wildfires and CVD/asthma
  - ...

Thank you!

Stanford University | Center for Population Health Sciences

Isabella Chu
itaylor@stanford.edu

Ayin Vala
ayinv@stanford.edu

The PHS Data Core: phsdata.stanford.edu