Data Synthesis = Future of Data Sharing ?

Khaled El Emam

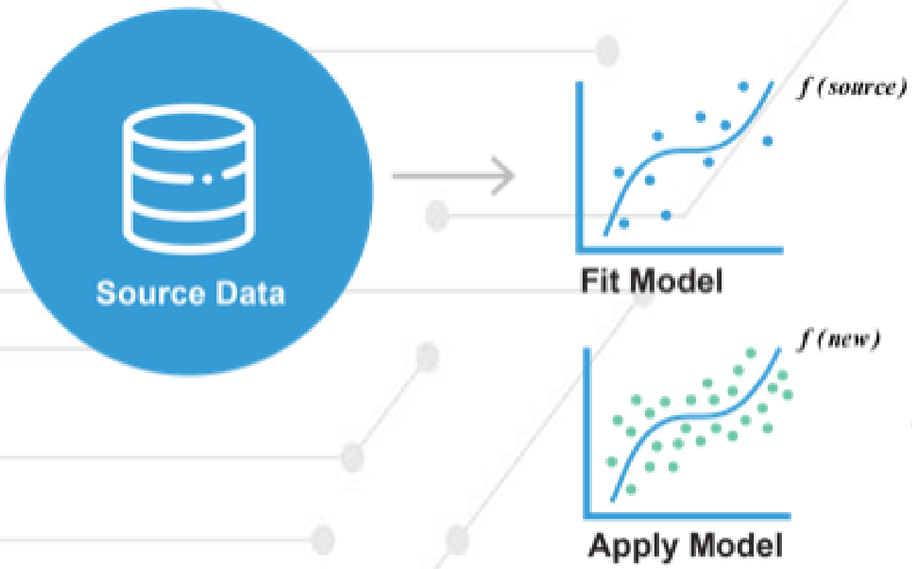
29th June 2021

kelemam@replica-analytics.com

Synthetic Data Uses

- Data Sharing and Data Access
  - AI and data science projects
  - Software testing
  - Proof of concept and technology evaluations
  - Open data/open science
  - Hackathons and data competitions/challenges
- Data Amplification and Data Augmentation
  - Amplifying small datasets
  - Correct bias

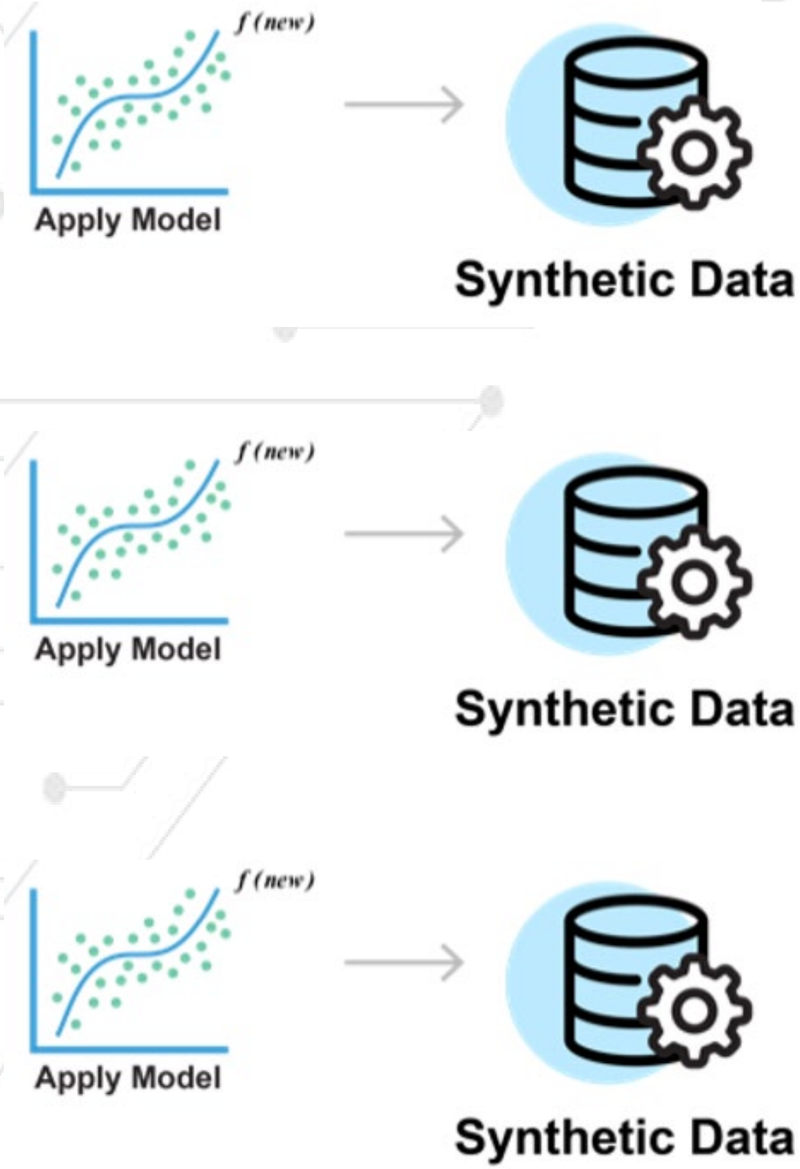
The Synthesis Process

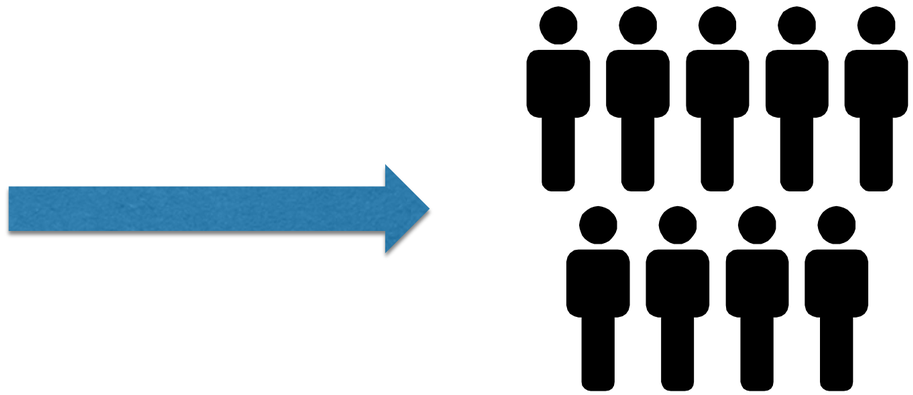


COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

Replica Analytics

Simulator Exchange





Data Consumers

Replica Analytics

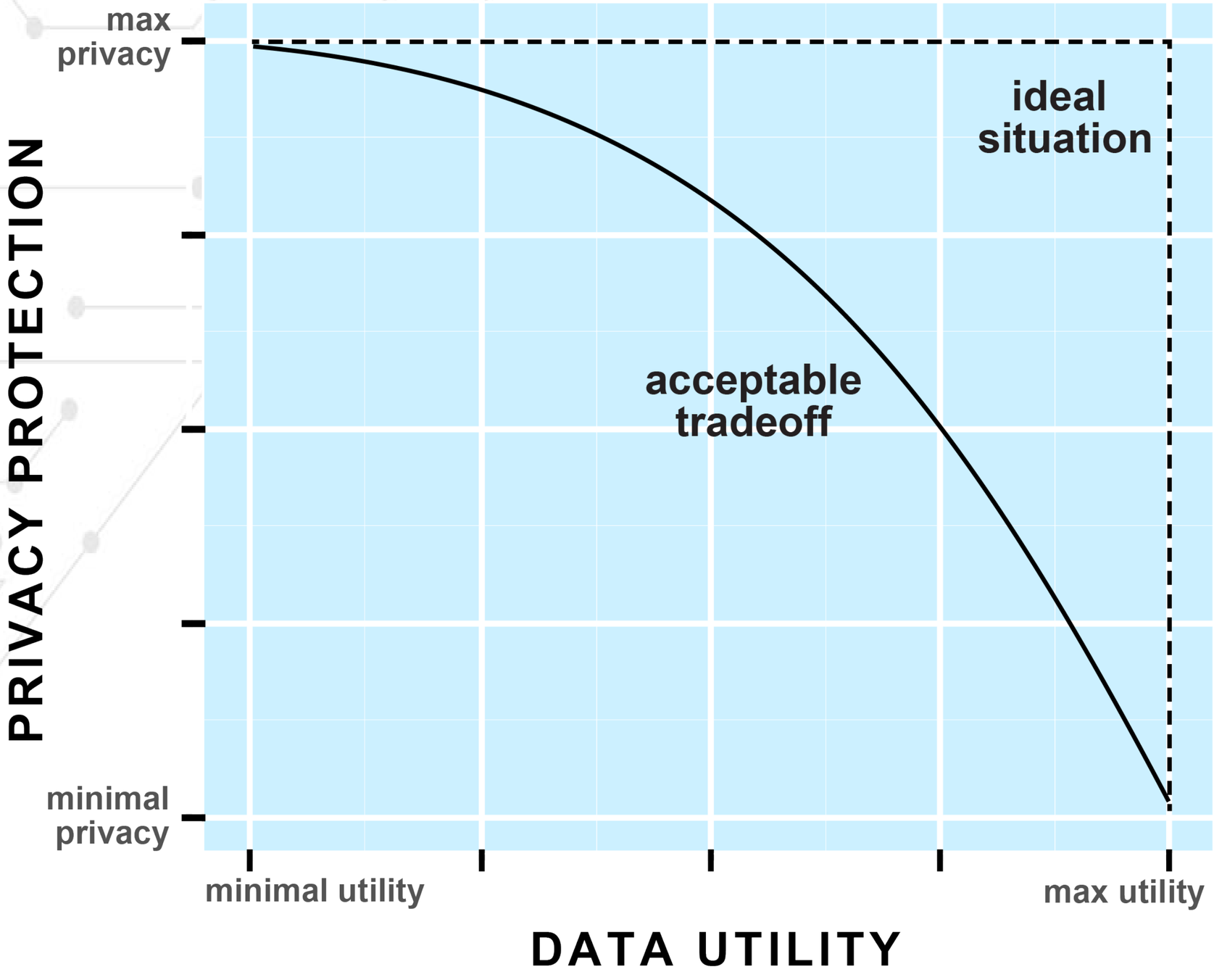
Evaluating (re-)Identification Risks

Dataset	Fully Synthetic Data	Original Data
Washington Hospital Data	0.0197	0.098
Canadian COVID-19 Data	0.0086	0.034

A commonly used risk threshold = 0.09

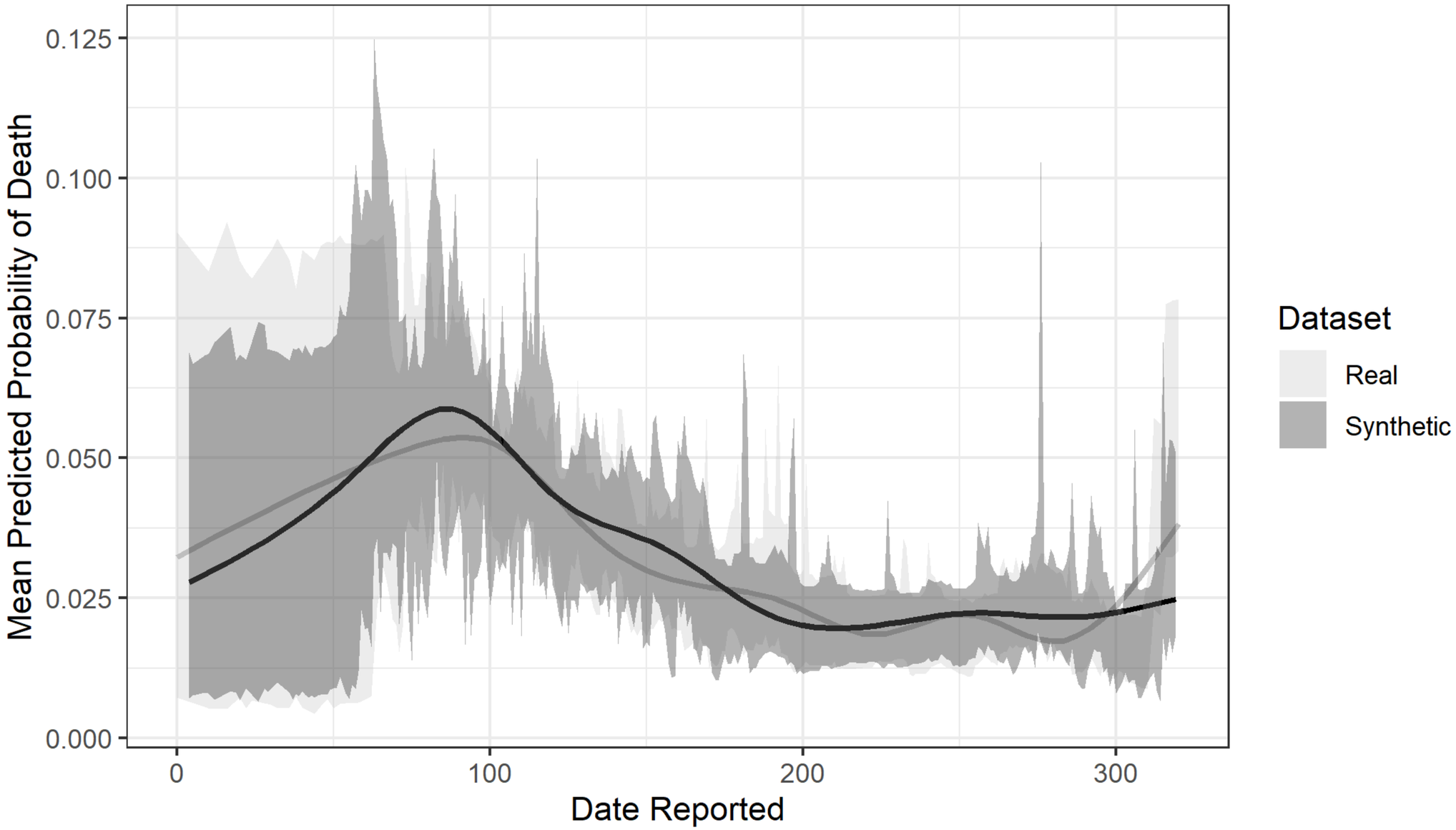
Replica Analytics

Privacy-Utility Trade-off



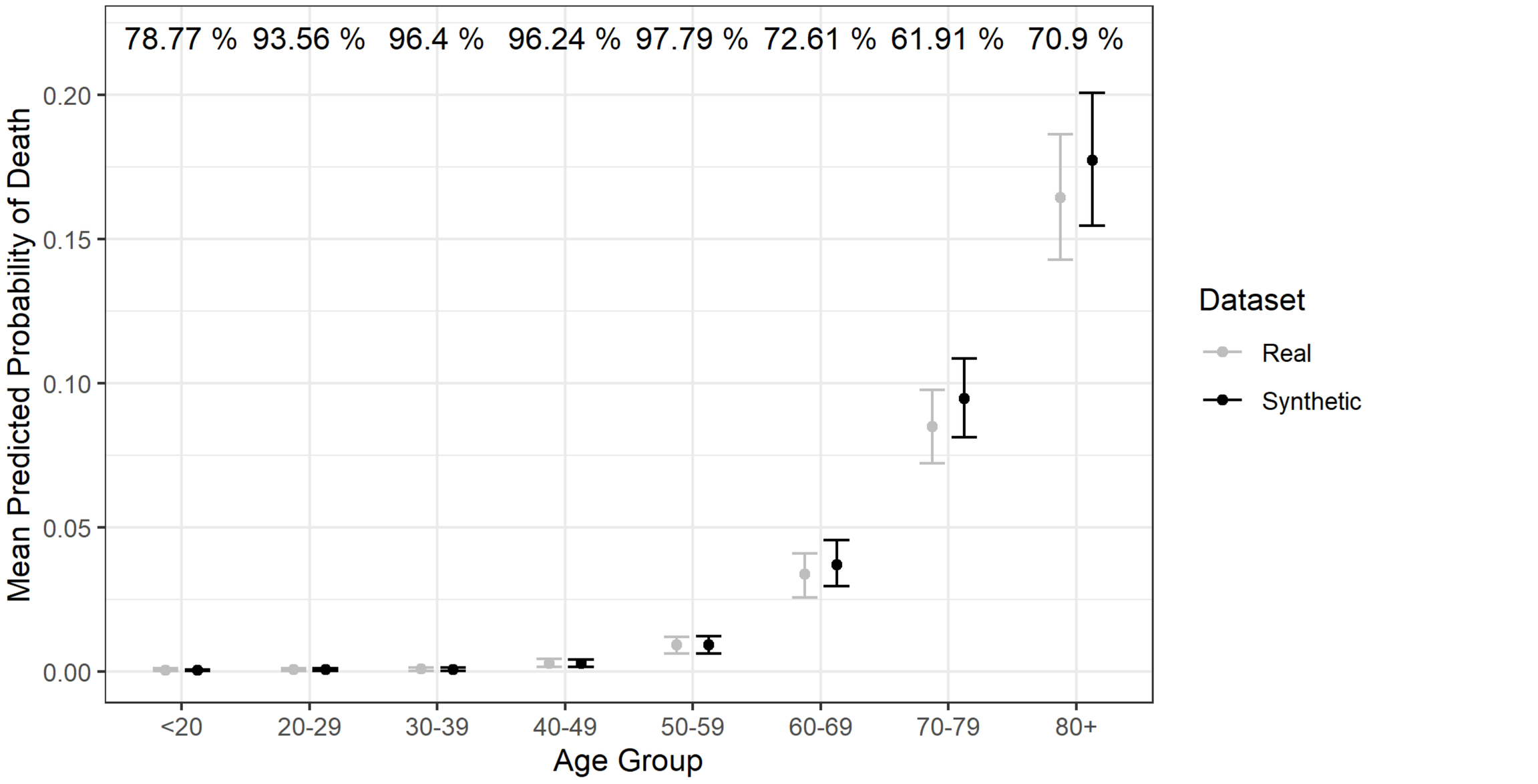
Replica Analytics

Comparing Real and Synthetic Data: Mortality Over Time



Replica Analytics

Comparing Real and Synthetic Data: Mortality By Age



Replica Analytics

References

- Z. Azizi, C. Zheng, L. Mosquera, L. Pilote, K. El Emam: “Replicating Secondary Studies Using Synthetic Clinical Trial Data”, BMJ Open, 11:e043497, 2021.
- K. El Emam, L. Mosquera, E. Jonker, H. Sood: “Evaluating the Utility of Synthetic COVID-19 Case Data”, JAMIA Open, 14(1):ooab012, January 2021.
- K. El Emam, L. Mosquera, and C. Zheng, “Optimizing the synthesis of clinical trial data using sequential trees,” J Am Med Inform Assoc, 28(1): 3-13, 2021.
- K. El Emam, L. Mosquera, and J. Bass, “Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation,” JMIR, vol. 22, no. 11, Nov. 2020. [Online]. Available: https://www.jmir.org/2020/11/e23139.
- K. El Emam, L. Mosquera, and R. Hoptroff, Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. O’Reilly, 2020.
- K. El Emam, “Seven Ways to Evaluate the Utility of Synthetic Data,” IEEE Security and Privacy, no. July/August, 2020.

Replica Analytics

Thank you

- Replica Analytics develops the Replica Synthesis software – generator of privacy protective synthetic health data and simulator exchange
- For more information on our synthetic data solutions:
  - Visit our website www.replica-analytics.com
  - Message us via the website contact page

Privacy Law & Synthetic Data

Mike Hintze

Partner, Hintze Law PLLC
Affiliate Instructor, University of Washington School of Law
Senior Fellow, Future of Privacy Forum

3 Key Legal Questions for Synthetic Data

1. Is the use of the original (real) data set to generate and/or evaluate a synthetic data set restricted or regulated under the law?
2. Is sharing the original data set with a third-party service provider to generate the synthetic data set restricted or regulated under the law?
3. Does the law regulate or otherwise affect the resulting synthetic data set?

Key Takeaways

Because the creation of synthetic data starts with processing a set of personal data (i.e. data relating to real people), laws that affect the processing of personal data will impact the initial creation and testing of synthetic data

Privacy laws generally permit such processing

The processing must adhere to other requirements of the law that apply in any case– such as transparency and data security

Sharing the original (real) data set with service providers that create synthetic data is also permitted, subject to certain obligations such as no secondary uses by the service providers, ensuring adequate protection of the personal data, and other contractual assurances

The resulting synthetic data set(s) should not be considered personal data, even under privacy laws with very broad definitions

Thus, synthetic data can be freely used and shared– even publicly released – for any purposes

Hintze Law

Privacy + Security

Mike Hintze

mike@hintzelaw.com

Twitter: @mhintze