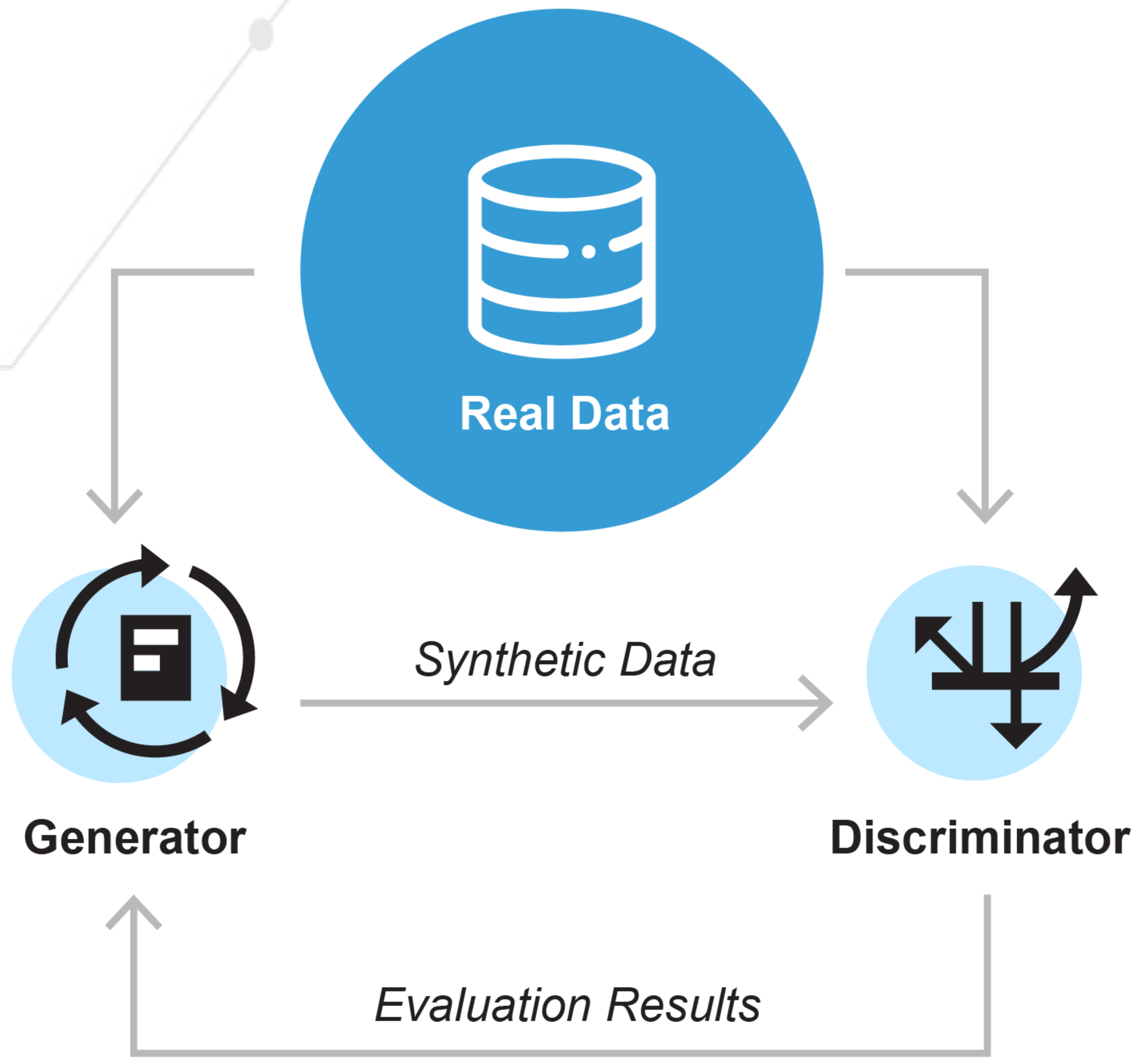
Practical Considerations for Synthetic Data Generation

Khaled El Emam
kelemam@ehealthinformation.ca

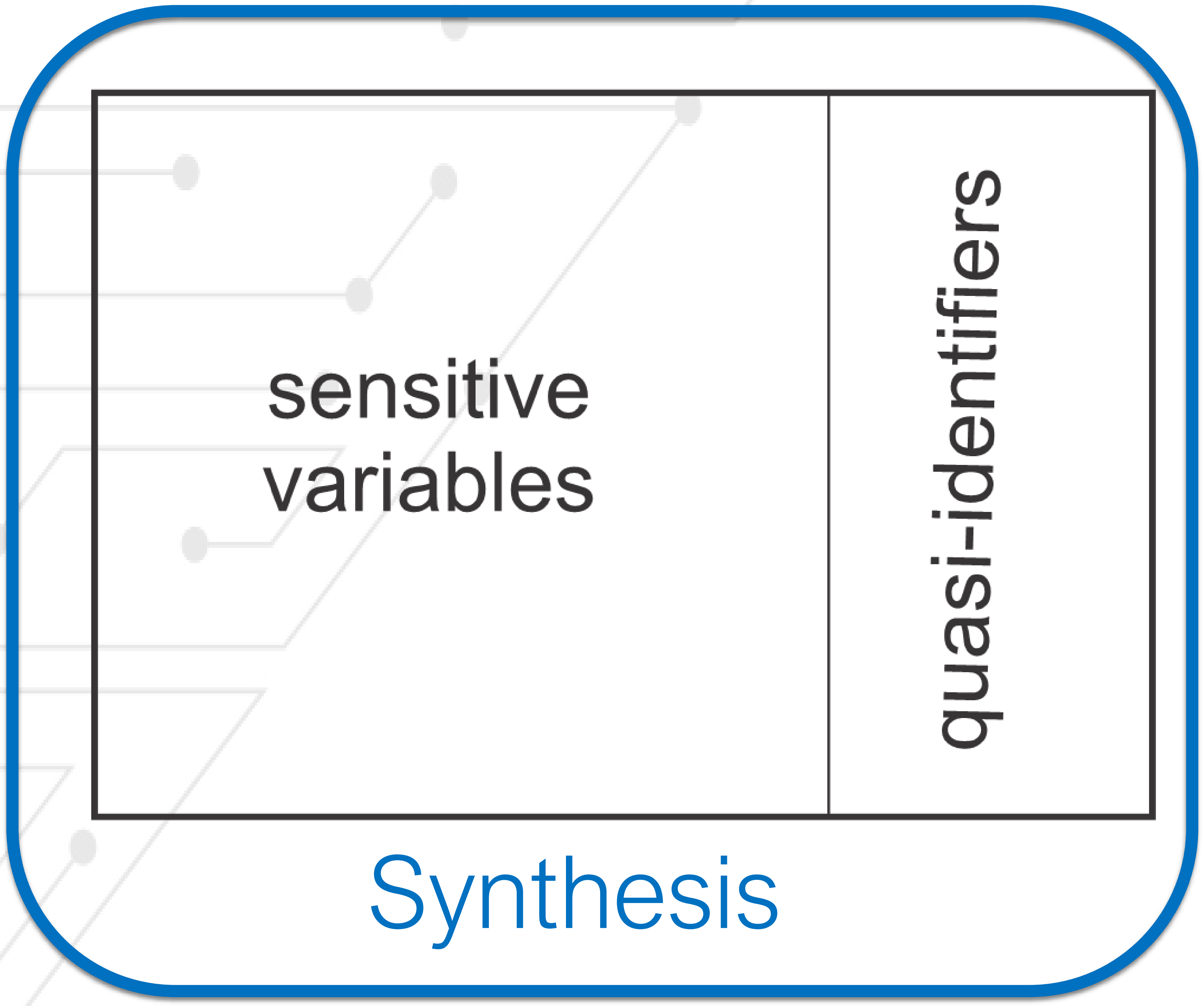
10th November 2021

Training a generative model often uses a discriminator

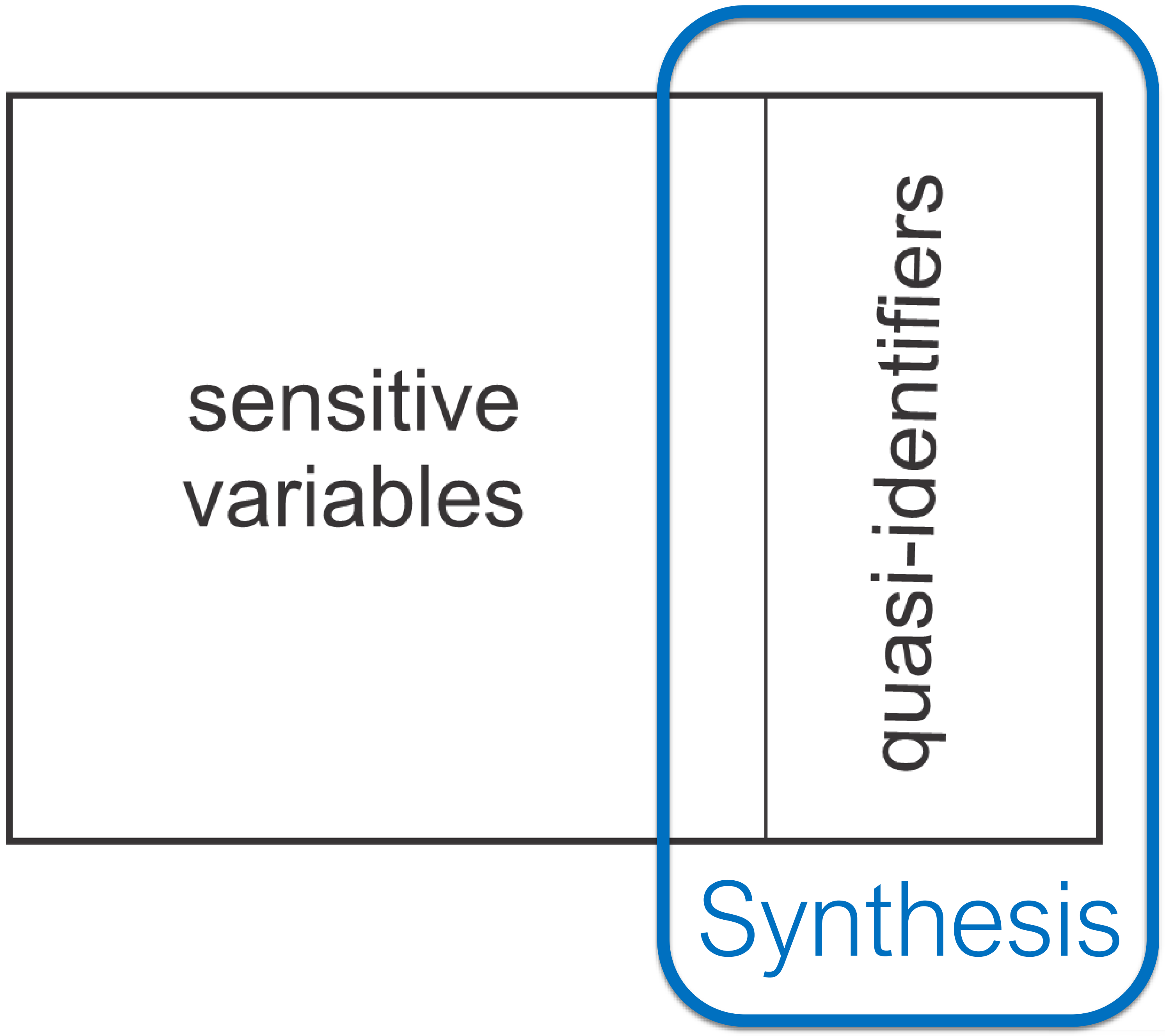


Two Synthesis Strategies

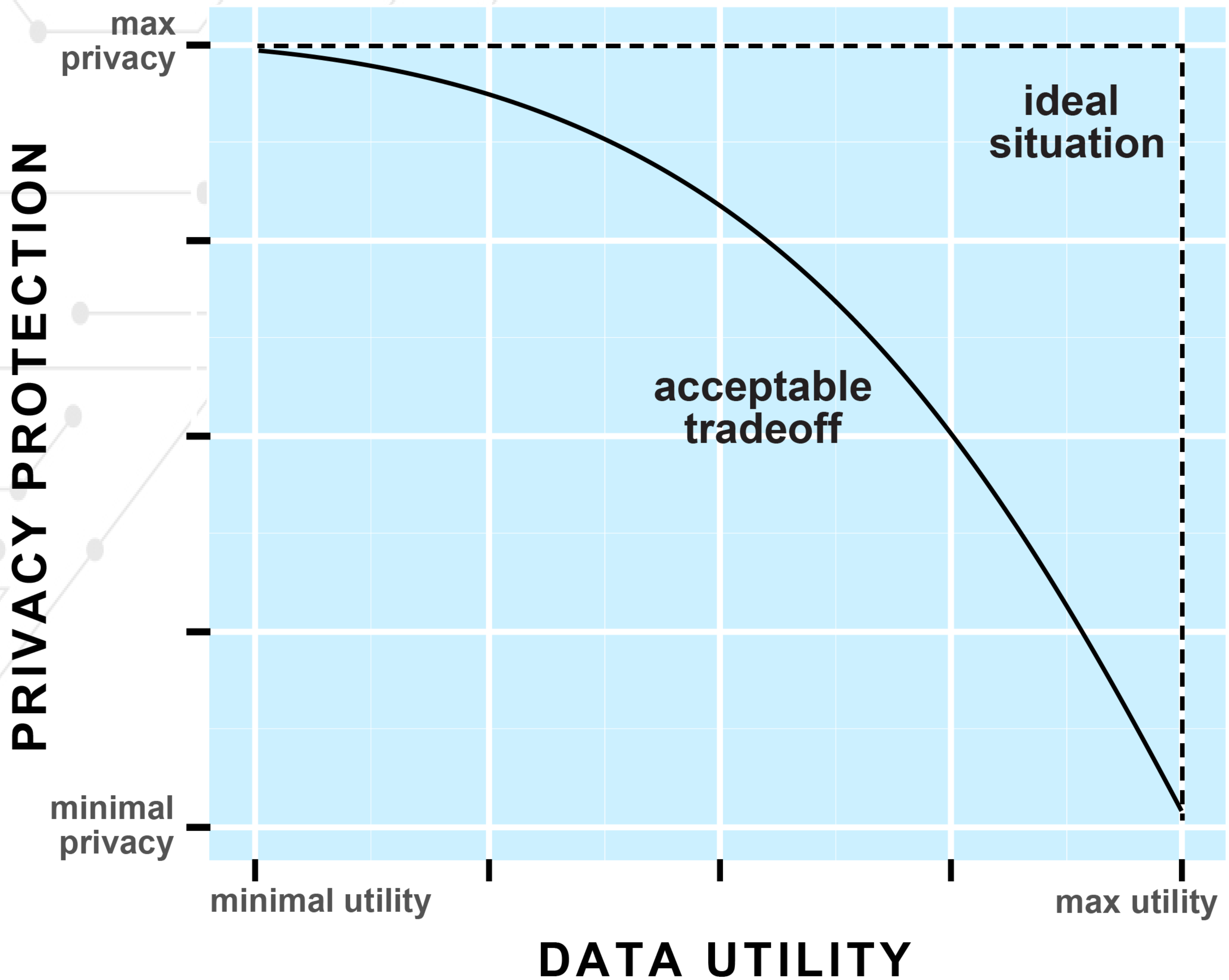
Full Synthesis
Synthesize all variables



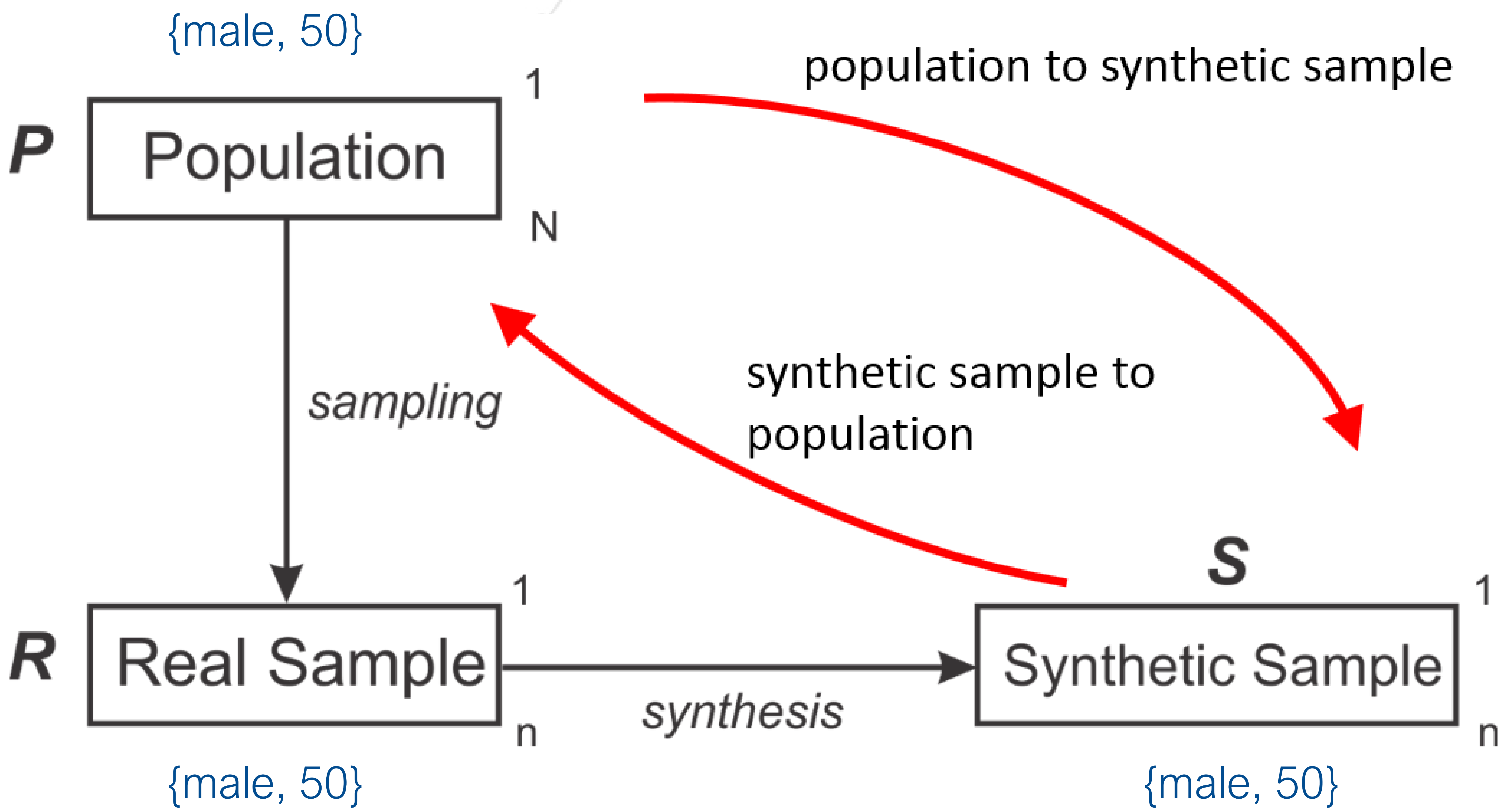
Partial Synthesis
Synthesize quasi-identifiers



Privacy-Utility Trade-off



Identity Disclosure Model

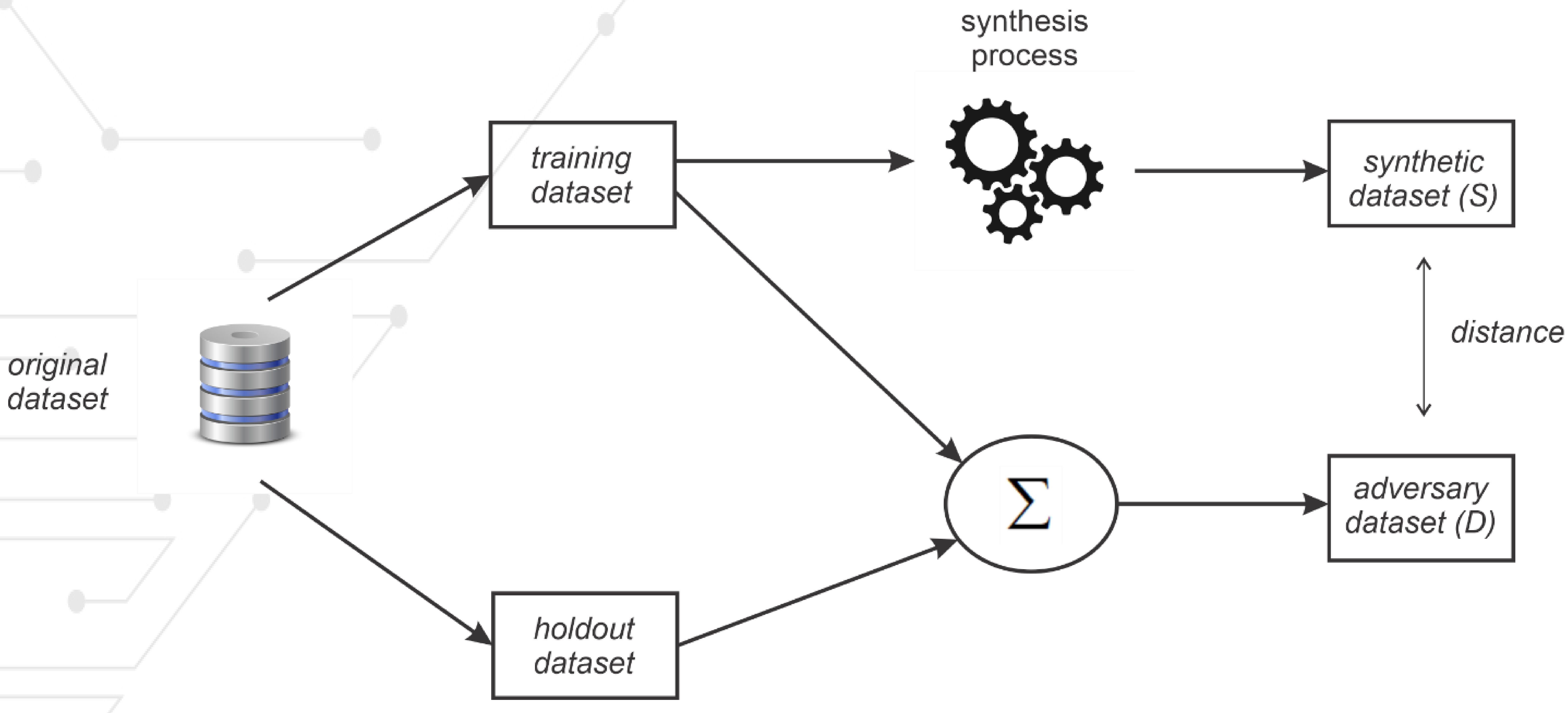


Evaluations of (re-)identification risks show that it is low in multiple studies across multiple datasets

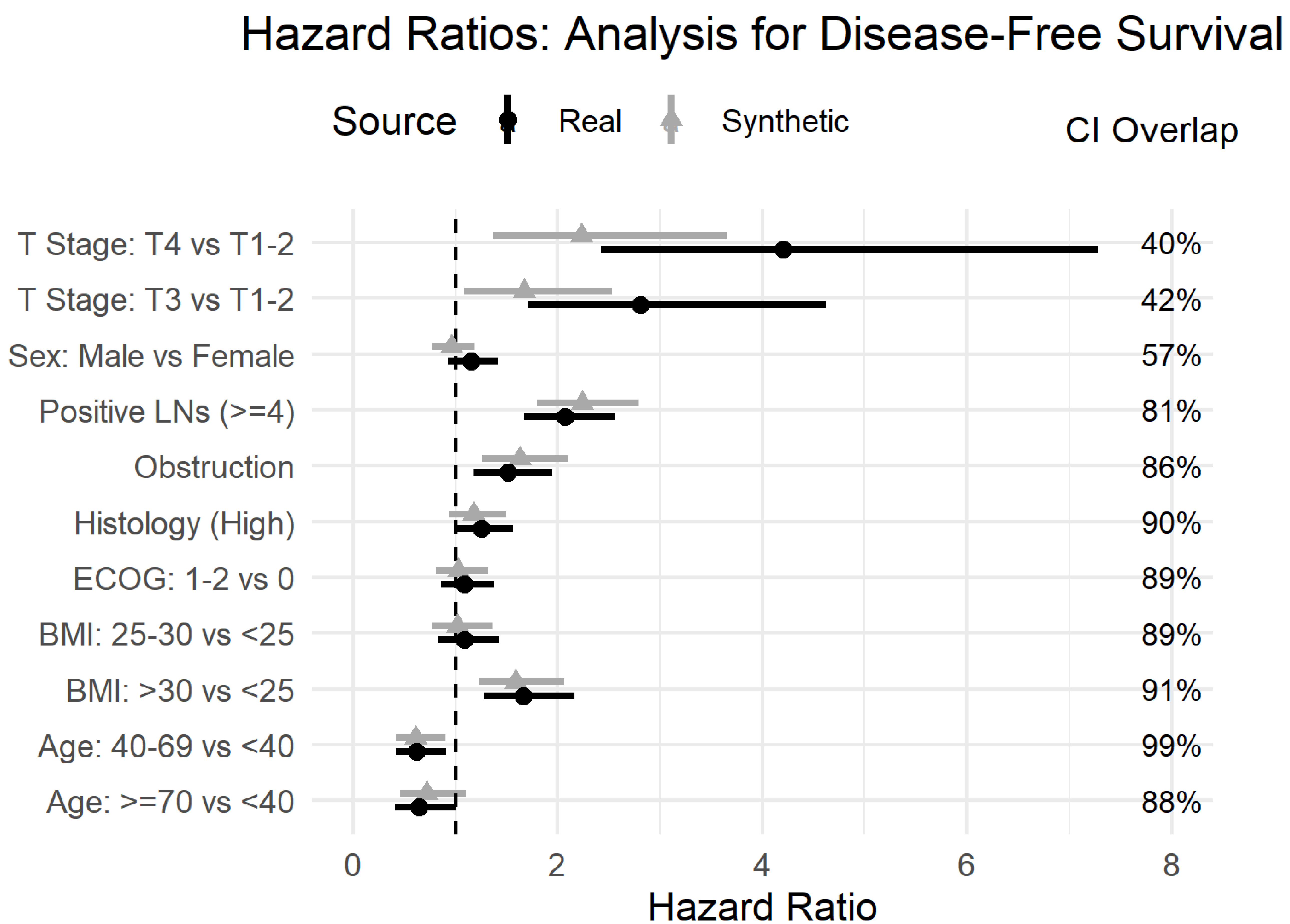
Dataset	Fully Synthetic Data	Original Data
Washington Hospital Data (Discharge)	0.0197	0.098
Canadian COVID-19 Data (Public Health)	0.0086	0.034

A commonly used risk threshold = 0.09

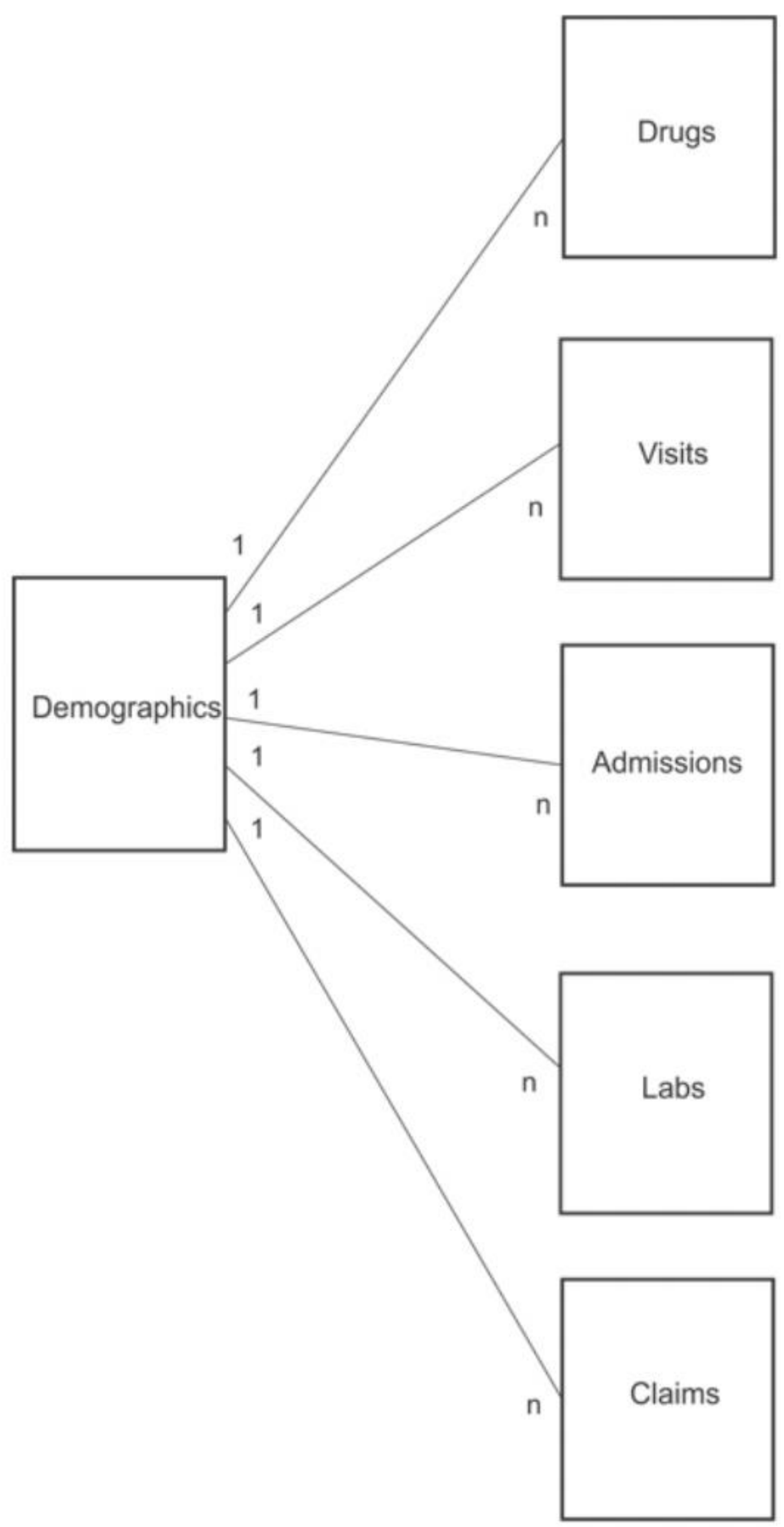
Membership disclosure: is the distance between S and D predictive of which records are in the training dataset



Comparing real and synthetic data: Adjusted model of impact of bowel obstruction on DFS



Longitudinal Data Model



Demographics
Age
Sex
Time to last day of follow-up available
Comorbidity score (elixhauser)

Drugs
Dispensed amount quantity
Relative dispensed time in days
Dispensed day supply quantity
Morphine use (binary)
Oxycodone use (binary)
Antidepressant use (binary)

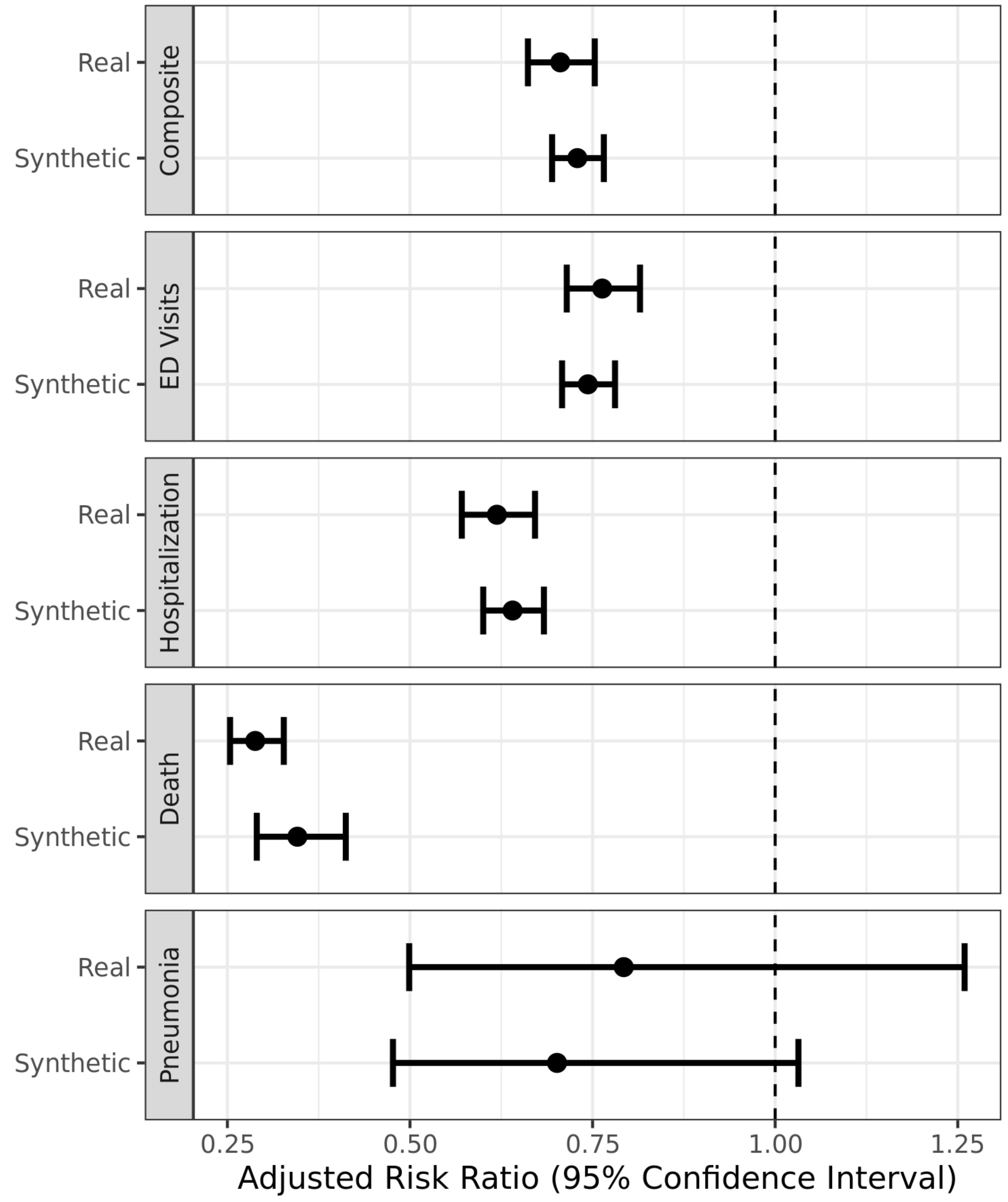
Visits (ED)
Relative admission time in days
Problem code 1
Problem code 2
Resource intensity weights

Admissions (Hospital)
Relative time admitted in days
LOS
Diagnosis code 1
Diagnosis code 2
Resource intensity weight

Lab
Test name
Test result (integer)
Relative time in days lab taken

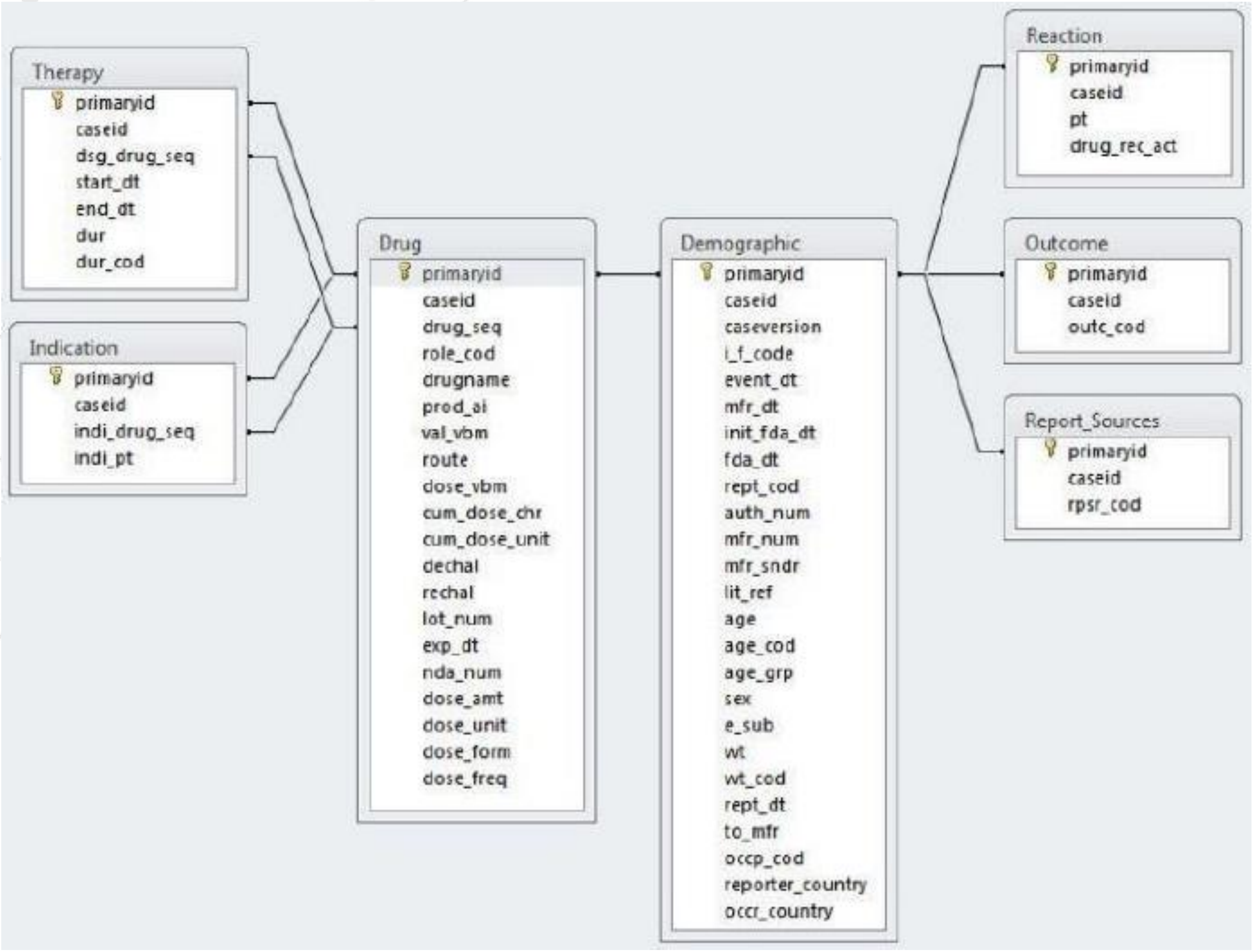
Claims
Primary diagnosis code
Provide specialty
Relative service event start date

Adjusted Cox Regression



Note: Adjusted estimates include the following co-variates: age, sex, antidepressant use, Elixhauser score, ALT, eGFR, HCT; Opioid 1 served as the reference group

Hierarchical datasets require a different approach




QUESTIONS

SDG 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,” *JAMIA*, 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.
- 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*, July/August, 2020.