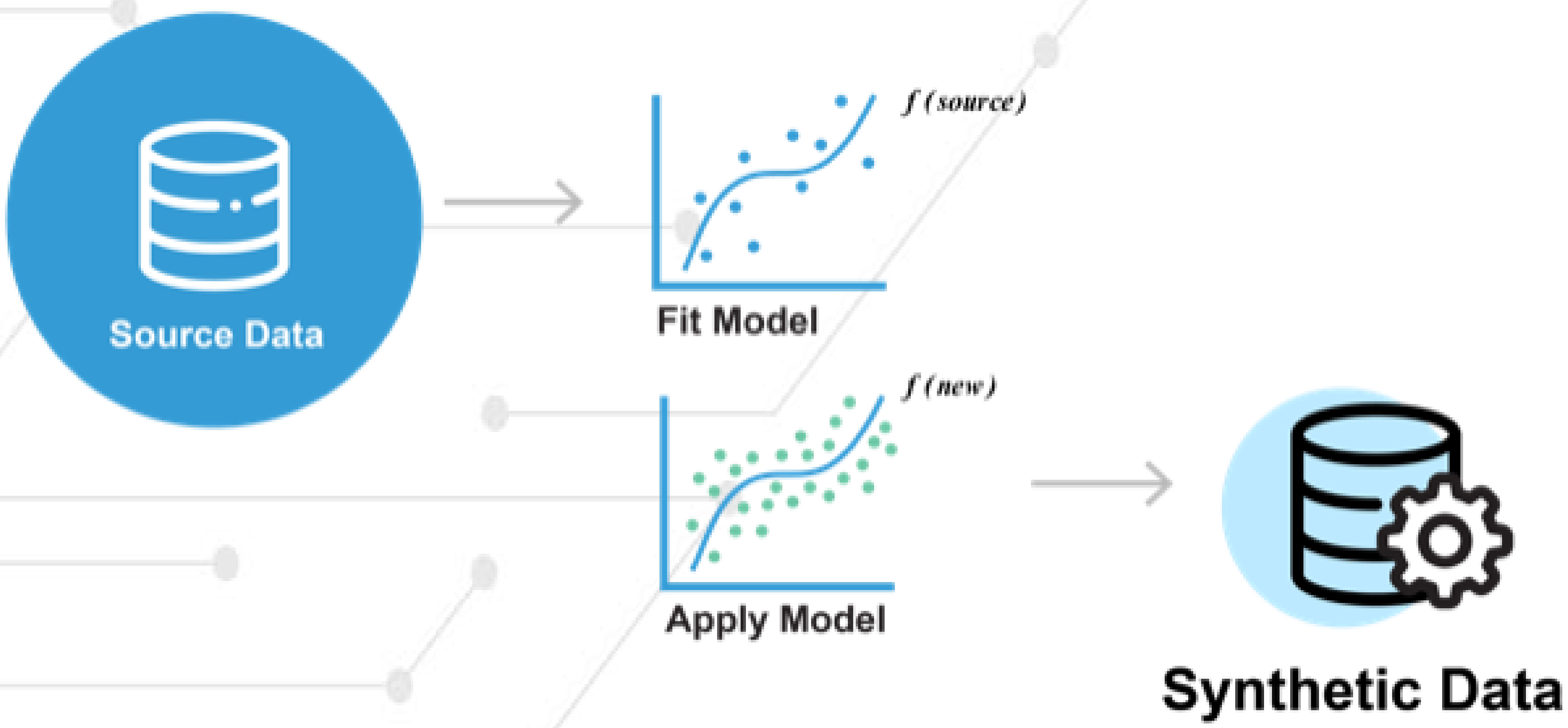
Generating and Applying Synthetic Health Data

Khaled El Emam
kelemam@ehealthinformation.ca

Space Opera Theatre

DEEP FAKES

The synthetic data generation process



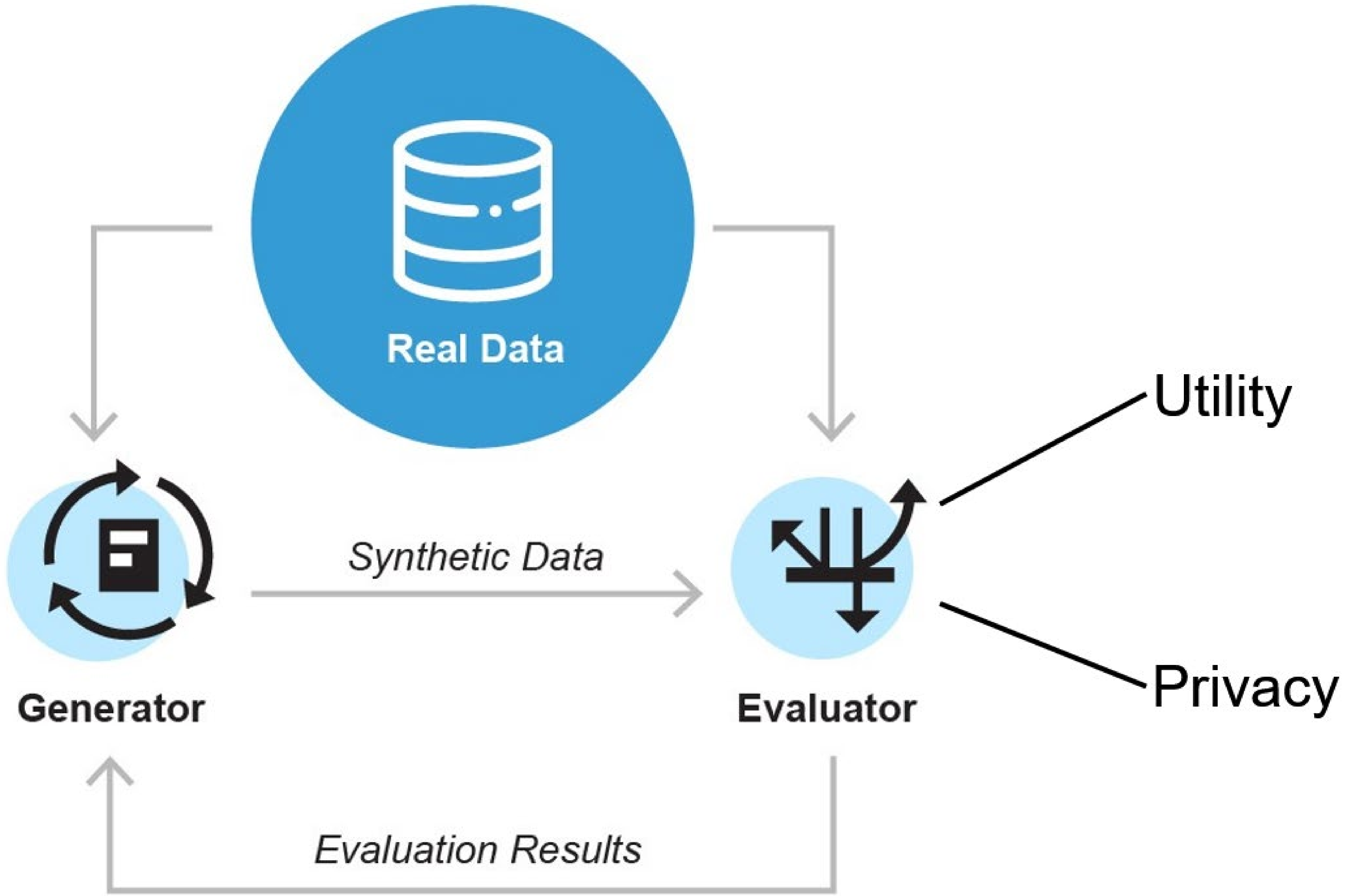
COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

Common Clarifications

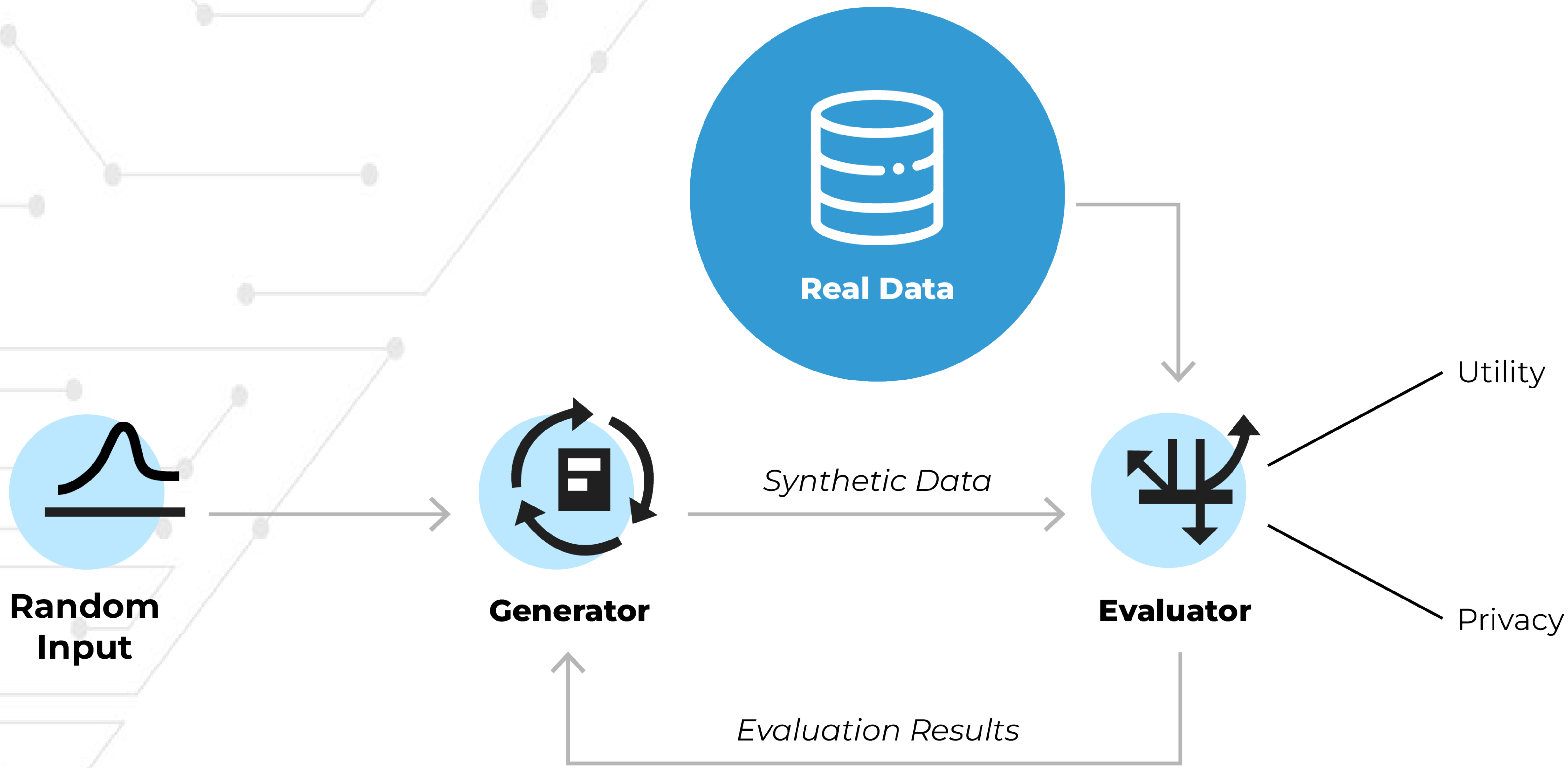
- The source datasets can be as small as 100 or 150 patients. We have developed generative modeling techniques that will work for small datasets.
- The source datasets can be very large – then it becomes a function of compute capacity that is available.
- It is not necessary to know how the synthetic data will be analyzed to build the generative models. The generative models capture many of the patterns in the source data.

El Emam K, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. Sebastopol, CA: O'Reilly Media 2020.

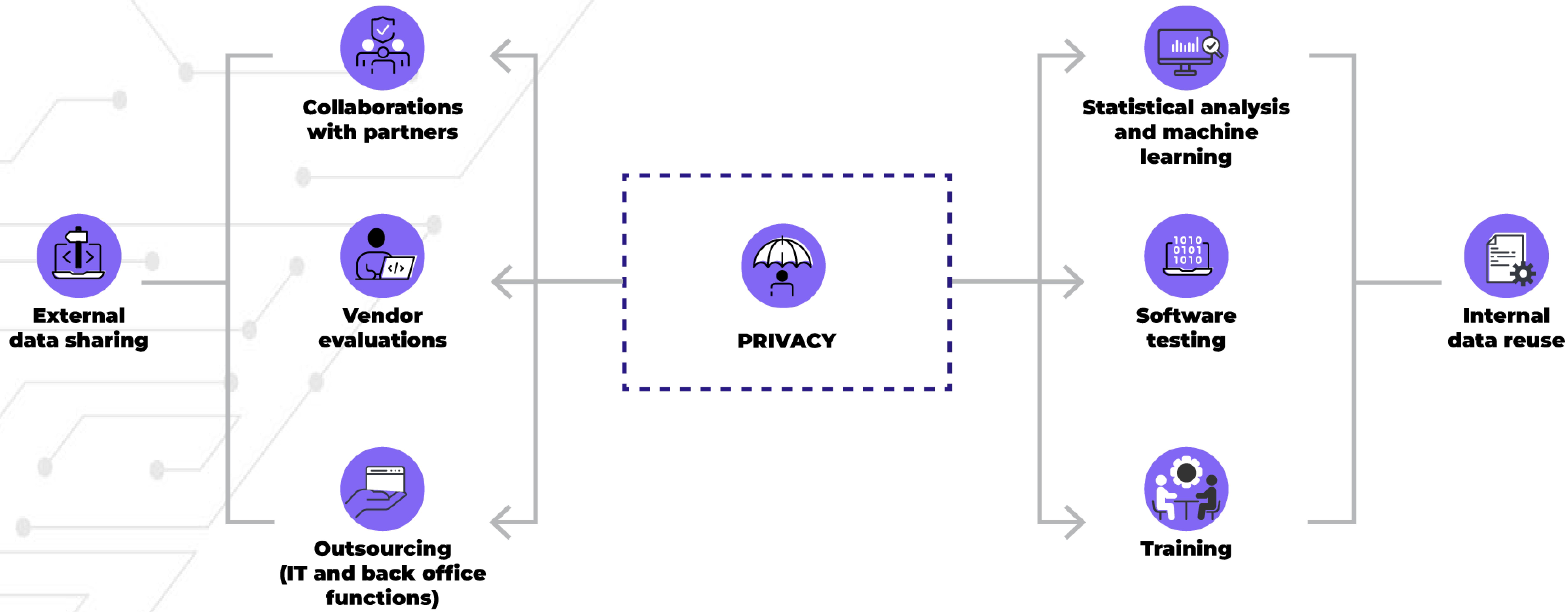
A combined loss of utility and privacy



A combined loss of utility and privacy



Privacy use cases



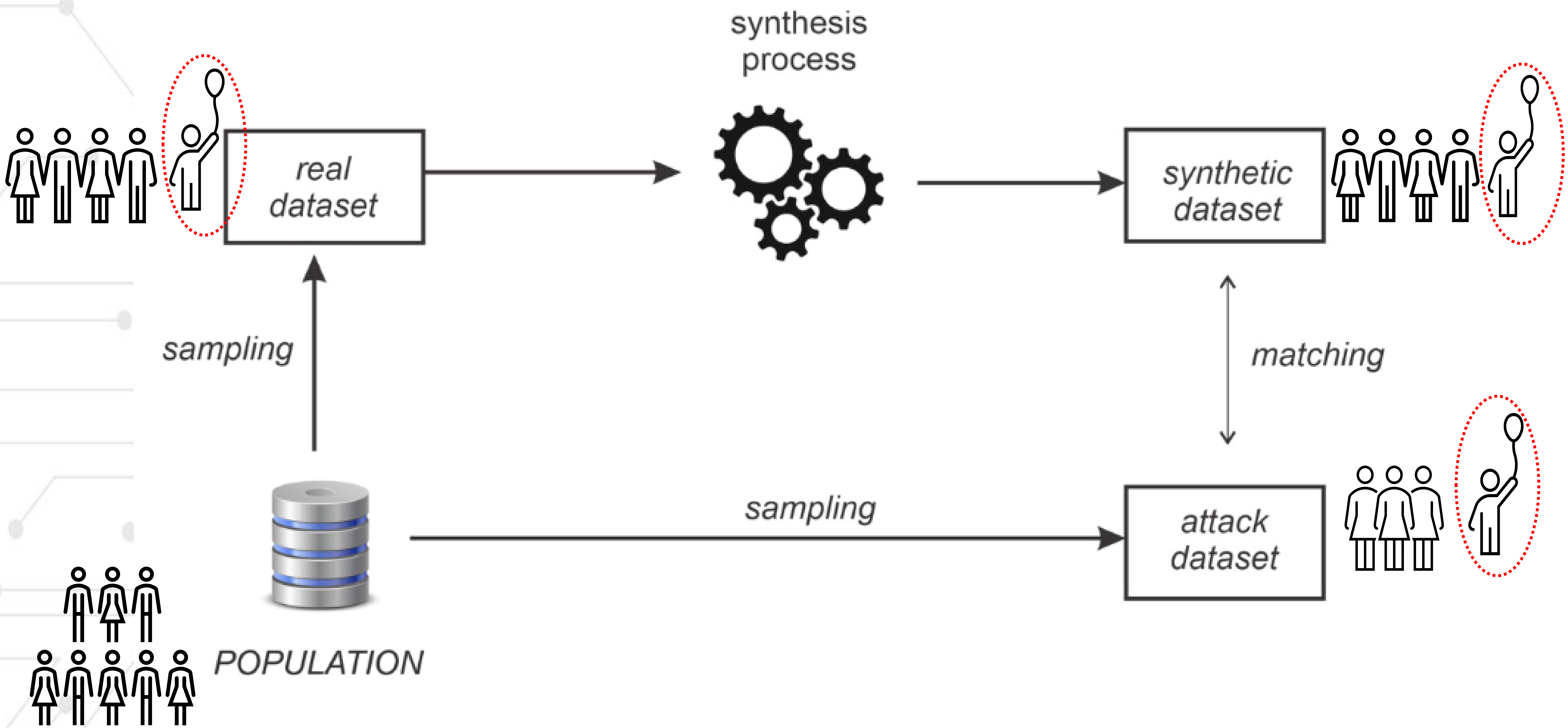
Attribution disclosure: find a record in the synthetic data similar to a high risk real individual <u>and</u> learn something new about that individual

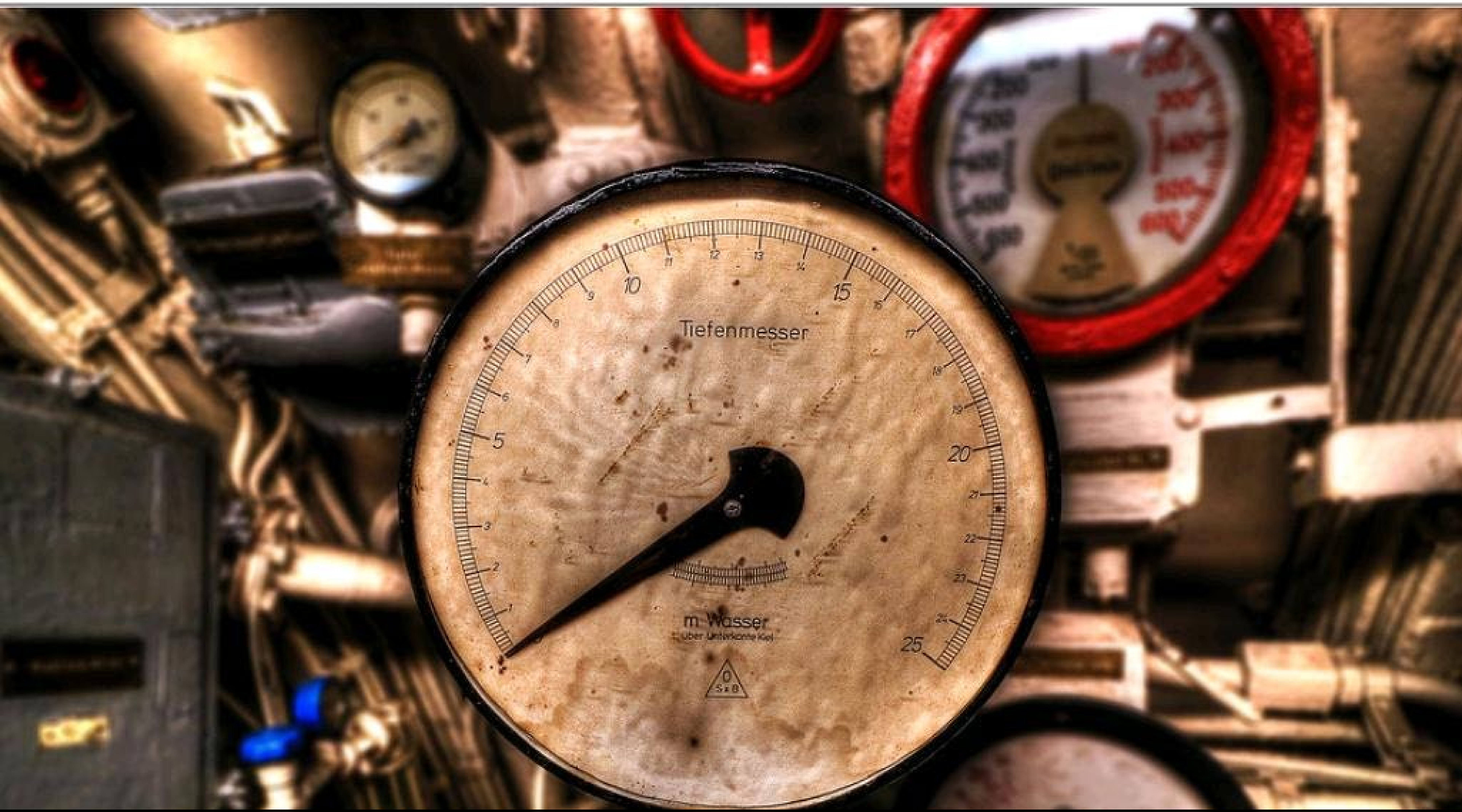
Quasi-identifiers

New Information

Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

The process for a membership disclosure attack



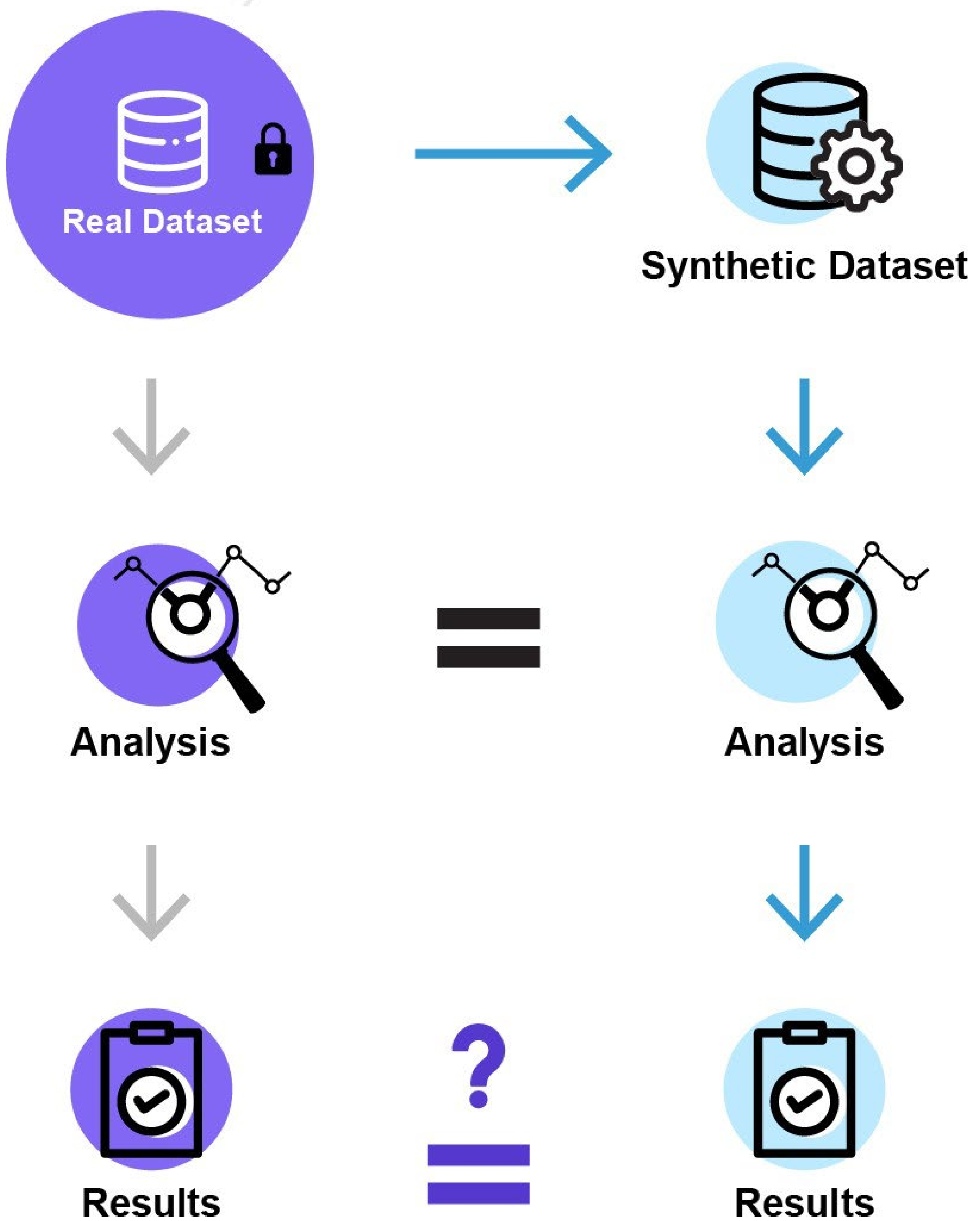


Generative models cannot guarantee always producing data with low privacy risk, but we can measure it every time and validate risk levels

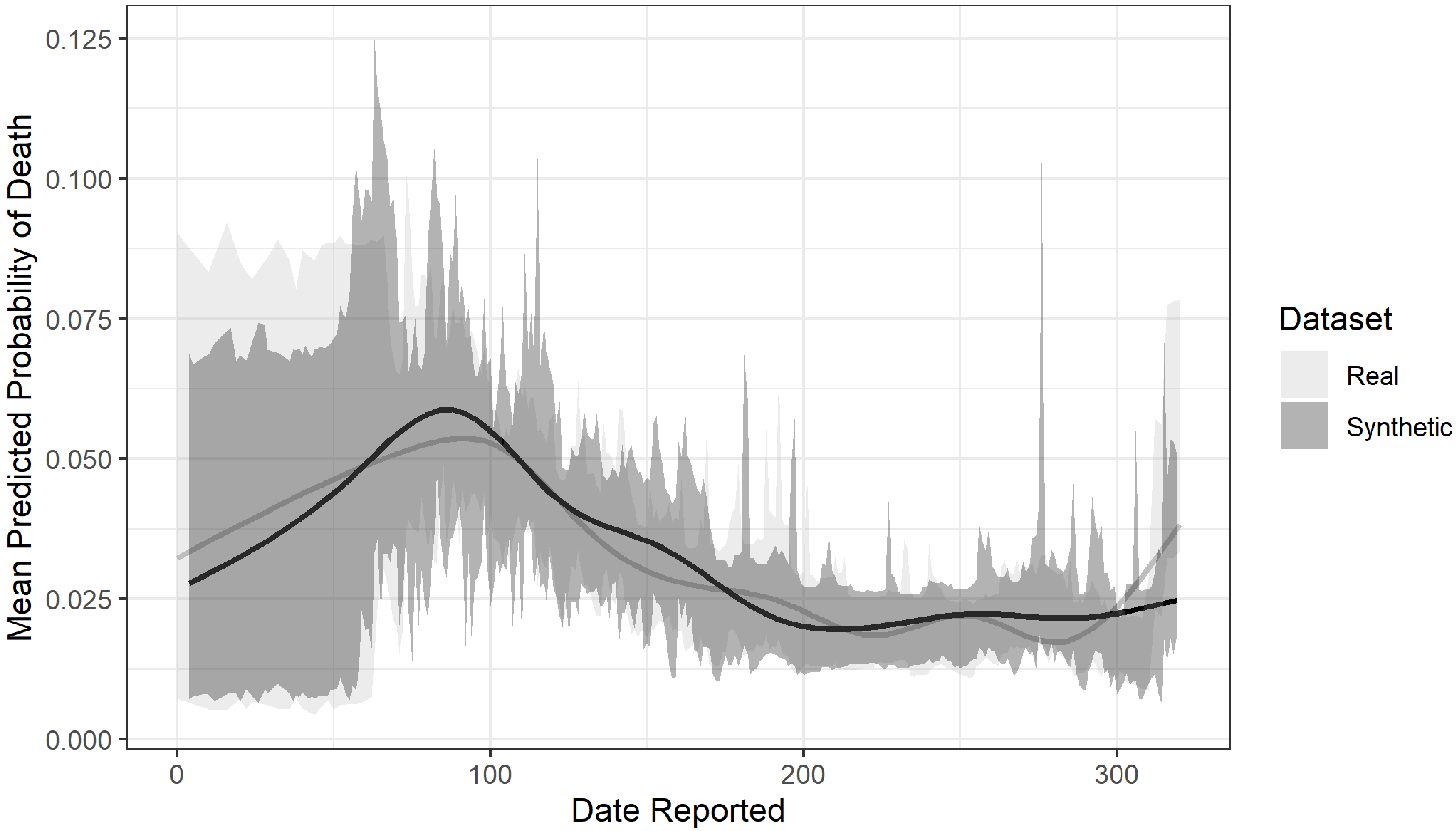
Assessing the Utility of Synthetic Data

- Expert Discrimination
  - Can a clinician to tell the difference between a real and a synthetic record ?
- Fidelity
  - How similar the joint distribution of the synthetic data is to the joint distribution of the real data ?
- Replicability
  - Are the analysis findings from models trained on the synthetic data the same as the findings on the real data, and are the population inferences on the synthetic data valid ?

Replicability of results

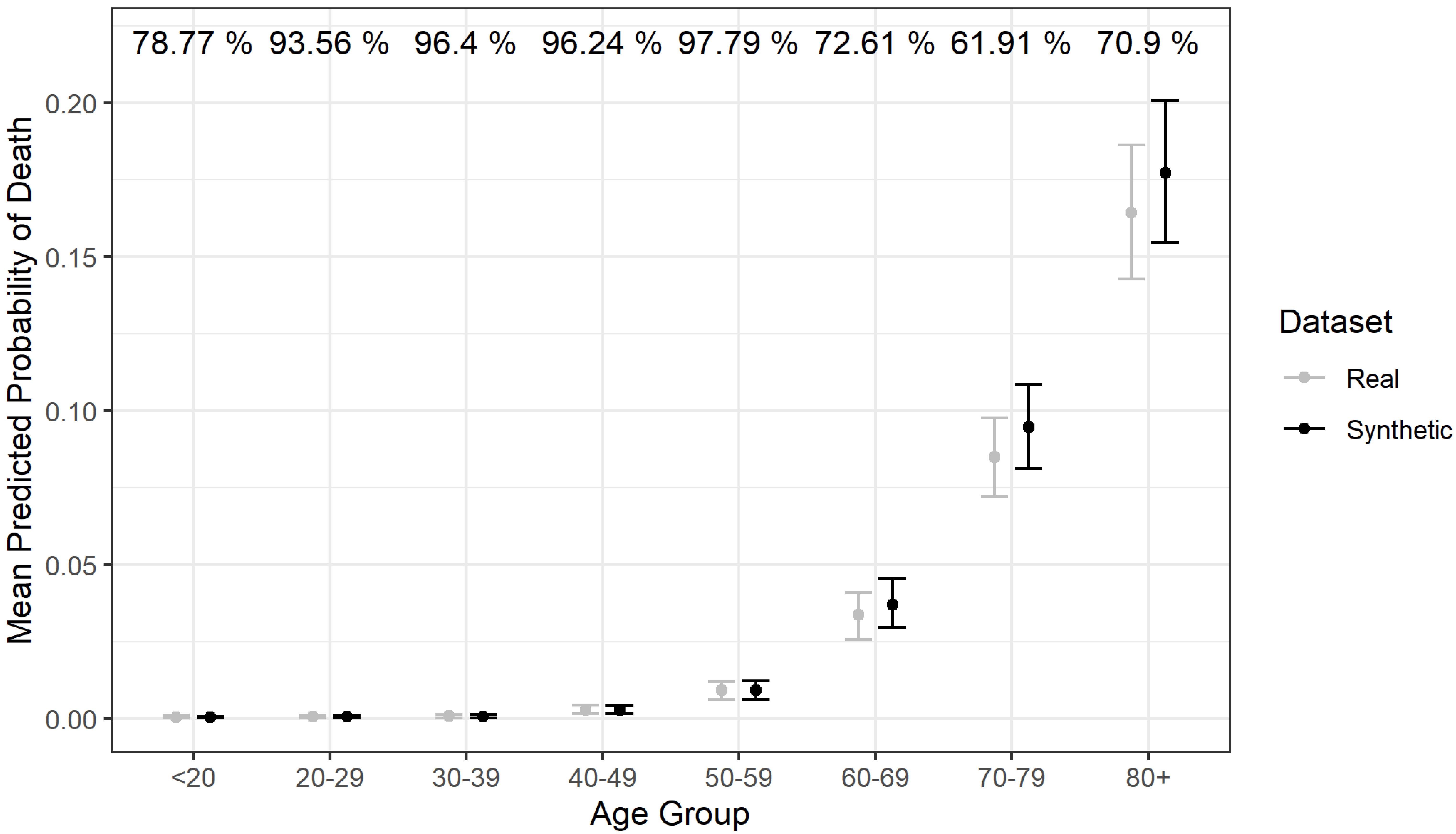


Comparing Real and Synthetic Data: Mortality Over Time



K. El Emam, L. Mosquera, E. Jonker, H. Sood: “Evaluating the Utility of Synthetic COVID-19 Case Data”, JAMIA Open, 14(1):ooab012, 2021.

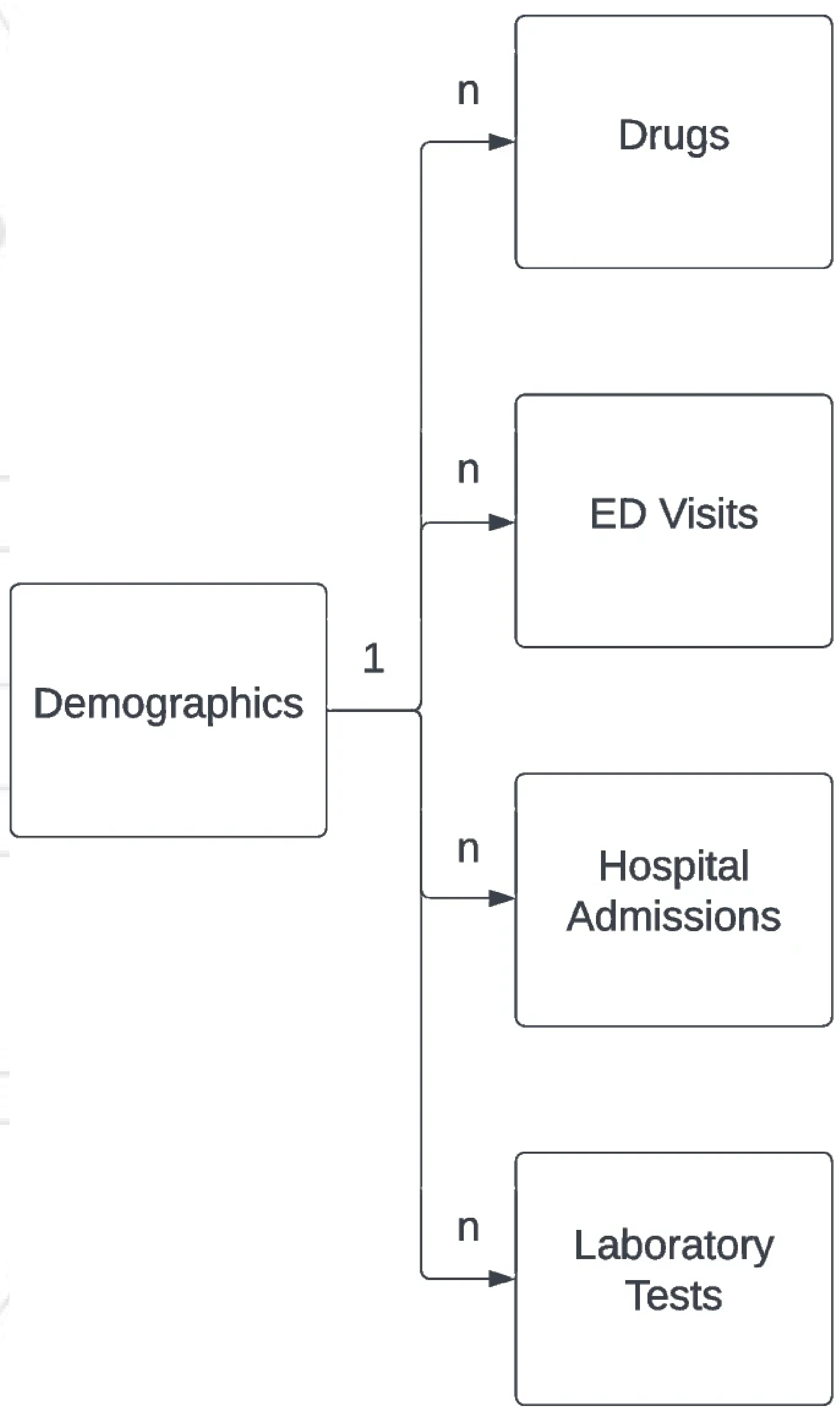
Comparing Real and Synthetic Data: Mortality By Age



K. El Emam, L. Mosquera, E. Jonker, H. Sood: "Evaluating the Utility of Synthetic COVID-19 Case Data", JAMIA Open, 14(1):ooab012, 2021.

uOttawa

Longitudinal Health System Dataset



Demographics
Age Sex Elixhauser comorbidity index Time to last follow-up

ED Visits
Diagnostic code Resource intensity weight Relative date of visit

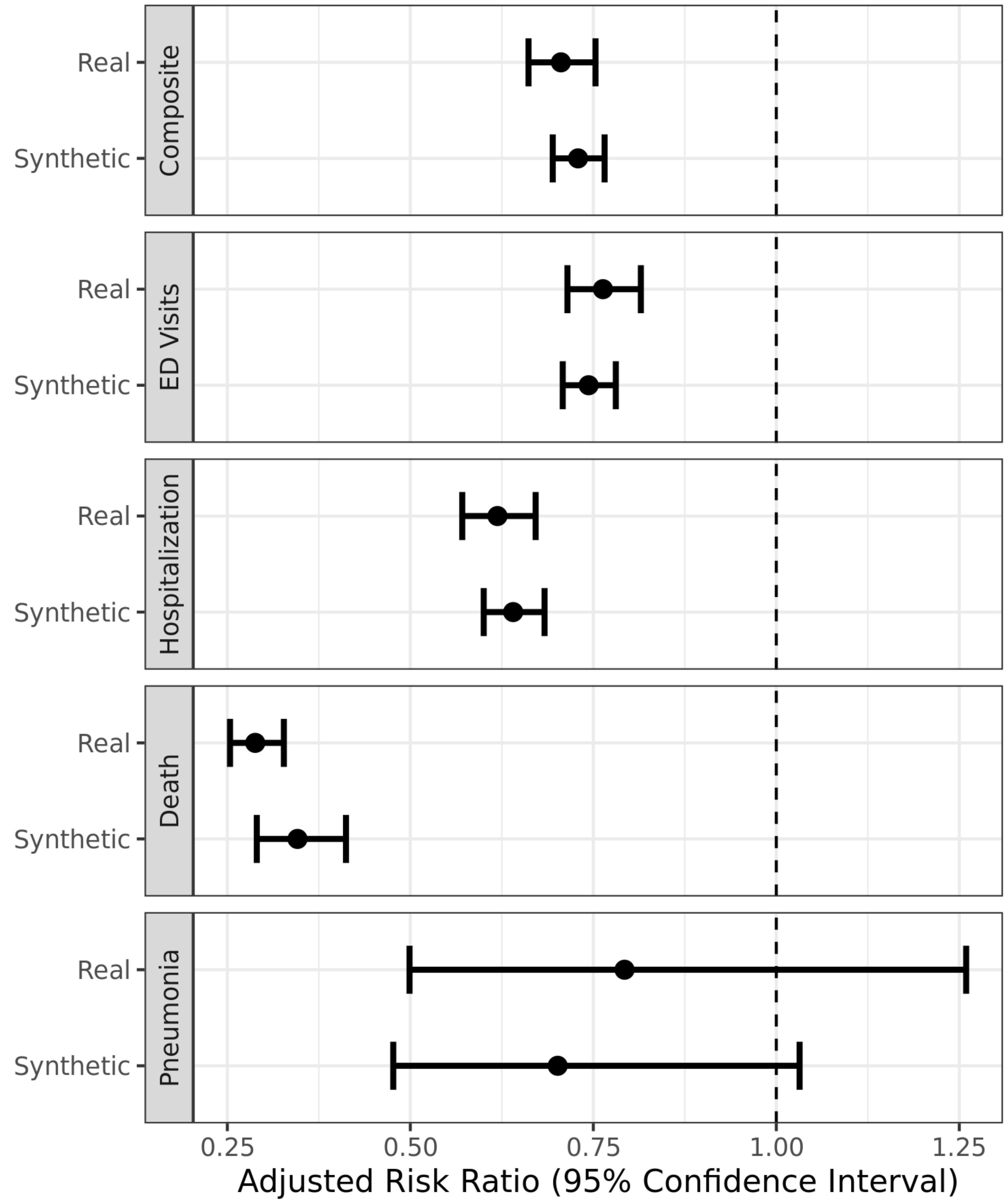
Hospital Admissions
Diagnostic code Resource intensity weight Length of stay Relative date of admission

Drugs
Amount dispensed Dispensed days supply Drug type Relative date of dispensation

Laboratory Tests
Test name Test result Relative date of test

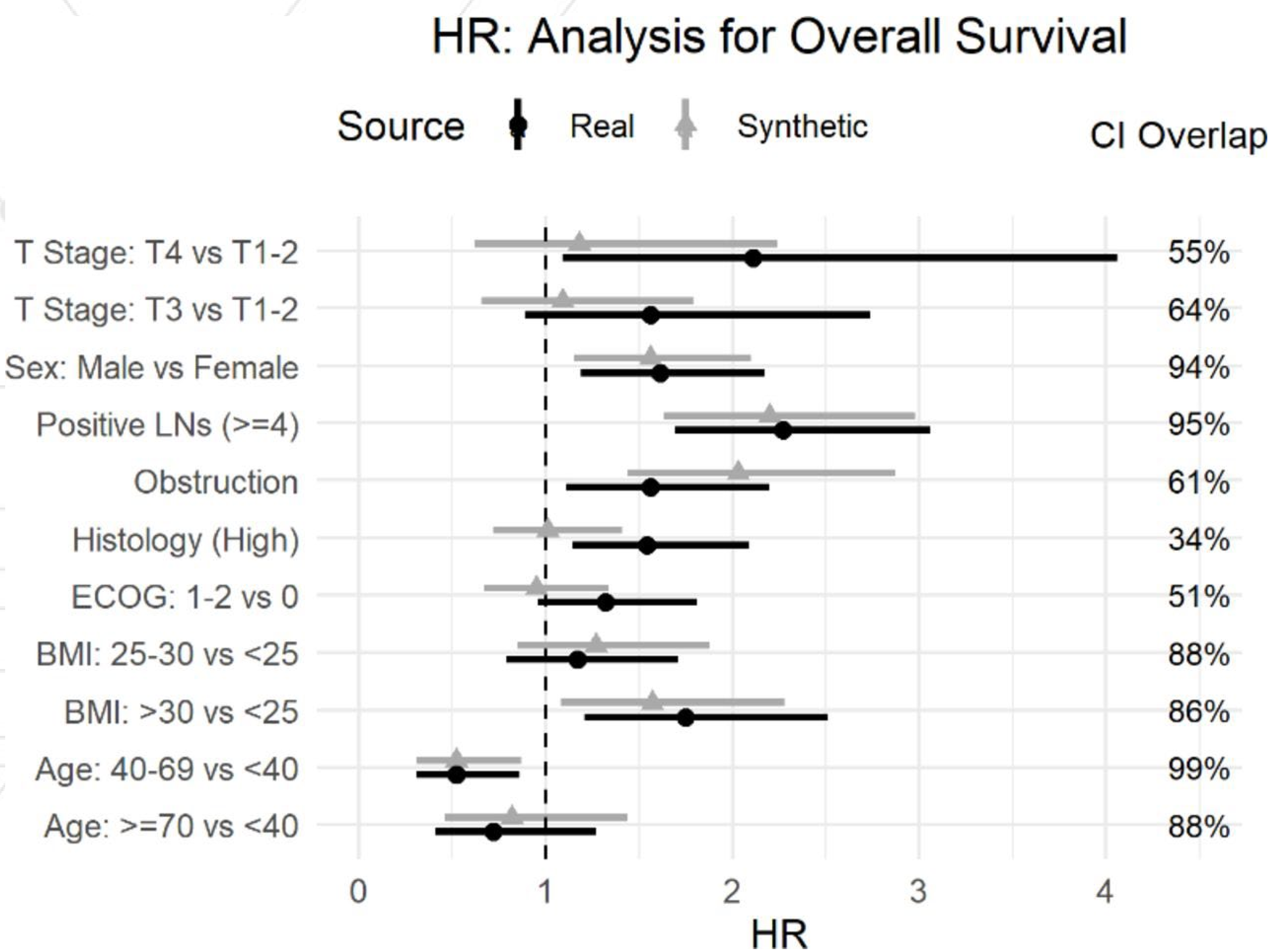
L. Mosquera, K. El Emam, L. Ding, V. Sharma, XH Zhang, S. Kababji, C. Carvalho, B. Hamilton, D. Palfrey, L. Kong, B. Jiang, D.T. Eurich: “A Method for Generating Synthetic Longitudinal Health Data”, BMC Medical Research Methodology, 23(1): 67, 2023.

Cox Regression Results



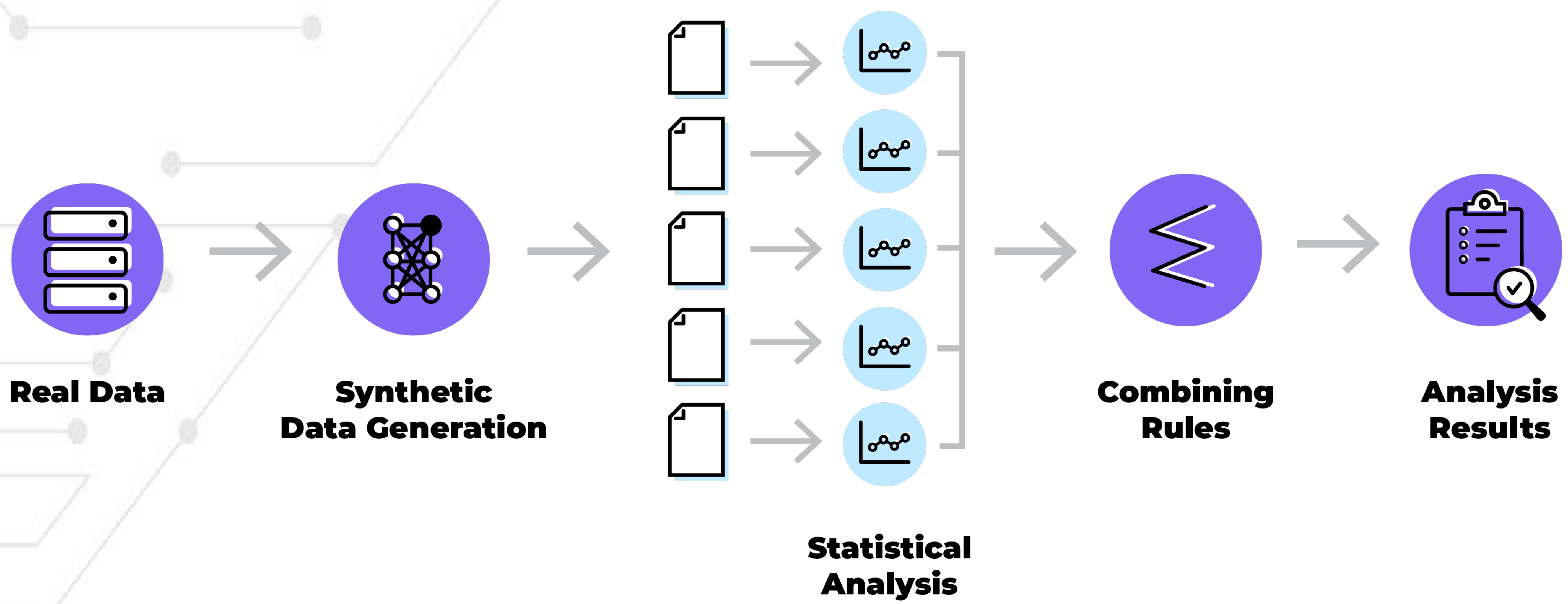
L. Mosquera, K. El Emam, L. Ding, V. Sharma, XH Zhang, S. Kababji, C. Carvalho, B. Hamilton, D. Palfrey, L. Kong, B. Jiang, D.T. Eurich: “A Method for Generating Synthetic Longitudinal Health Data”, BMC Medical Research Methodology, 23(1): 67, 2023.

Colon Cancer Clinical Trial



Azizi Z, Zheng M, Mosquera L, et al. Can synthetic data be a proxy for real clinical trial data ? A validation study. *BMJ Open*. 2021;11:e043497.

Because synthesis introduces additional variation, this needs to be accounted for in models to get valid estimates



El Emam K, Mosquera L, Fang X, et al. An evaluation of the replicability of analyses using synthetic health data. Sci Rep. 2024;14:6978.

Replication utility on eight breast cancer clinical trials

		SEQ			GAN			VAE		
Data Set	Sample Size	Estimate Agreement	Decision Agreement	CI Overlap	Estimate Agreement	Decision Agreement	CI Overlap	Estimate Agreement	Decision Agreement	CI Overlap
REaCT-HER2+	48	1	1	0.77	1	1	0.88	1	1	0.94
REaCT-G/G2	401	1	1	0.91	[a]	[a]	[a]	1	1	0.67
REaCT-ILIAD	218	1	1	0.99	1	1	0.85	1	0	0.74
REaCT-ZOL	211	1	[b]	0.98	1	[b]	0.88	0	[b]	0.61
REaCT-BTA	230	1	1	0.85	1	0	0.68	1	0	0.72
CCTG MA27	7,576	1	1	0.90	1	1	0.62	1	1	0.82
SWOG 0307	6,097	1	1	0.93	1	0	0.50	1	1	0.95
NSABP B34	3,323	1	1	0.93	1	1	0.83	1	1	0.61

Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

[a]Training the generative model failed.

[b]The analysis is descriptive and hence decision agreement does not apply.

S. El Kababji, N. Mitsakakis, X. Fang, A.Beltran-Bless, G. Pond, L. Vandermeer, D. Radhakrishnan, L. Mosquera, A. Paterson, L. Shepherd, B. Chen, W. Barlow, J. Gralow, M-F Savard, M. Clemons, K. El Emam. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. JCO Clin Cancer Inform. 2023;e2300116

Attribution disclosure on eight breast cancer clinical trial datasets

	SEQ		GAN		VAE	
Data Set	Maximum Risk	Risk	Maximum Risk	Risk	Maximum Risk	Risk
REaCT-HER2+	2.56E-04	LO	2.35E-04	LO	2.35E-04	LO
REaCT-G/G2	1.10E-04	LO	1.10E-04	LO	1.10E-04	LO
REaCT-ILIAD	2.90E-05	LO	2.90E-05	LO	2.90E-05	LO
REaCT-ZOL	1.58E-03	LO	1.41E-03	LO	1.10E-03	LO
REaCT-BTA	6.48E-04	LO	6.43E-04	LO	6.43E-04	LO
CCTG MA27	1.37E-03	LO	1.37E-03	LO	1.38E-03	LO
SWOG 0307	2.09E-03	LO	2.17E-03	LO	2.02E-03	LO
NSABP B34	2.25E-02	LO	2.02E-02	LO	1.83E-02	LO

Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; LO, low risk; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

S. El Kababji, N. Mitsakakis, X. Fang, A.Beltran-Bless, G. Pond, L. Vandermeer, D. Radhakrishnan, L. Mosquera, A. Paterson, L. Shepherd, B. Chen, W. Barlow, J. Gralow, M-F Savard, M. Clemons, K. El Emam. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. JCO Clin Cancer Inform. 2023;e2300116

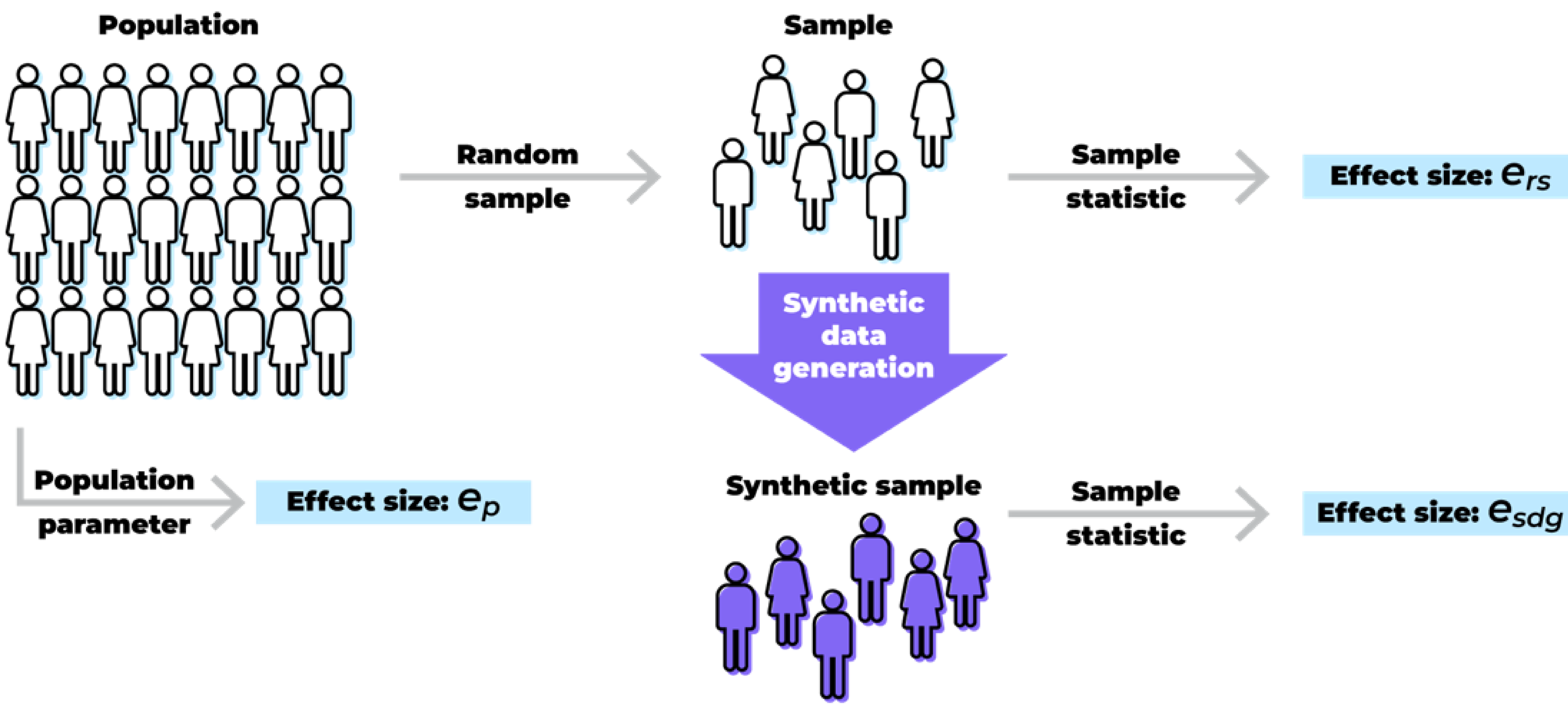
Membership disclosure on eight clinical trial datasets

		SEQ		GAN		VAE	
Data Set	n/N (sampling fraction)	F_rel	Risk	F_rel	Risk	F_rel	Risk
REaCT-HER2+	0.021	0.15	LO	0.07	LO	0.09	LO
REaCT-G/G2	0.062	0.06	LO	0.06	LO	0.06	LO
REaCT-ILIAD	0.004	0.02	LO	0.02	LO	0.02	LO
REaCT-ZOL	0.023	0.02	LO	0.02	LO	0.02	LO
REaCT-BTA	0.207	0.13	LO	0.18	LO	0.18	LO
CCTG MA27	0.573	0.31	HI	0.32	HI	0.34	HI
SWOG 0307	0.147	0.13	LO	0.13	LO	0.13	LO
NSABP B34	0.158	−0.02	LO	−0.15	LO	−0.19	LO

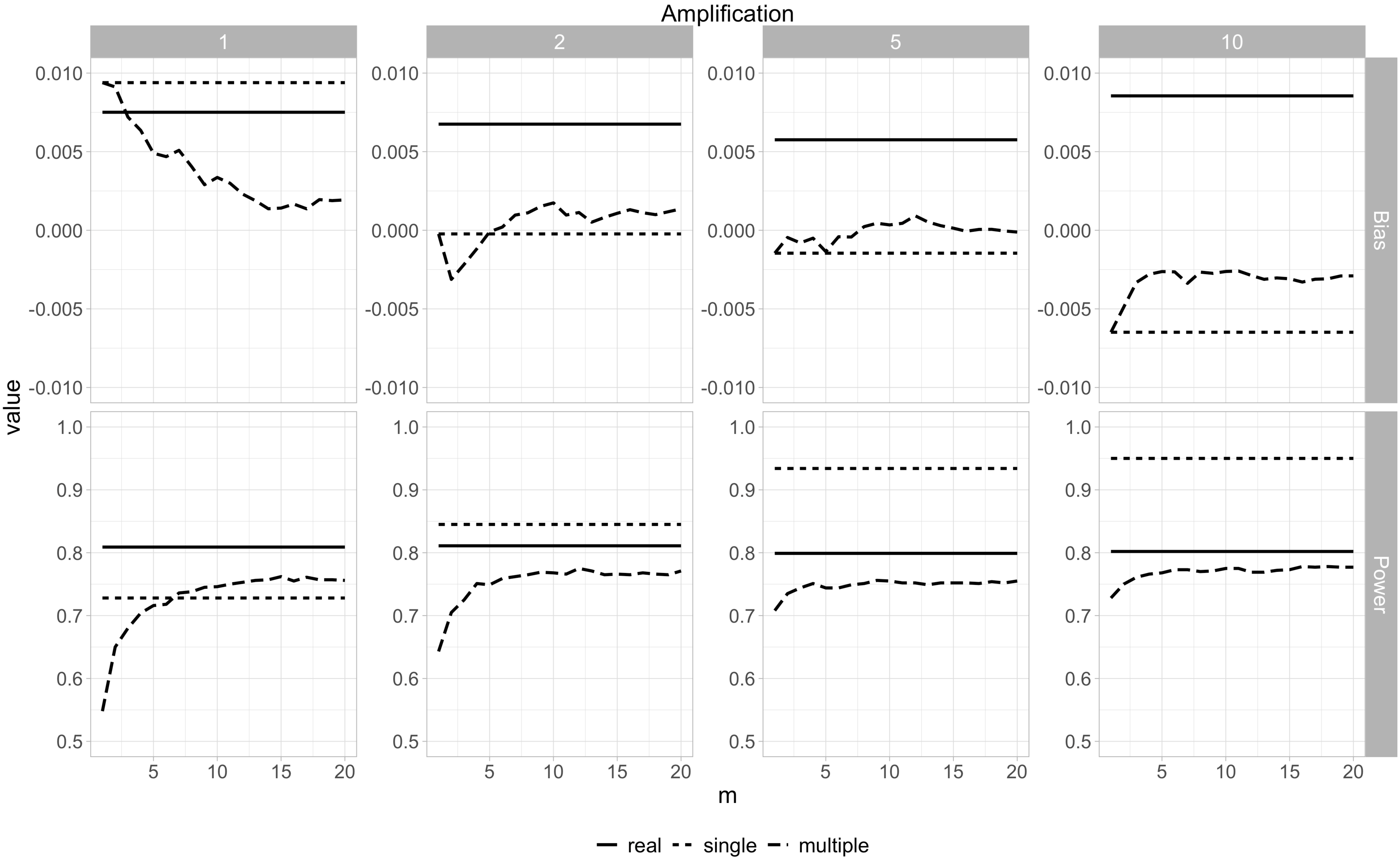
NOTE. The threshold for the sampling fraction is 0.33, and 0.2 for the relative F1 score (F_rel).
Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; HI, high risk; LO, low risk; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

S. El Kababji, N. Mitsakakis, X. Fang, A.Beltran-Bless, G. Pond, L. Vandermeer, D. Radhakrishnan, L. Mosquera, A. Paterson, L. Shepherd, B. Chen, W. Barlow, J. Gralow, M-F Savard, M. Clemons, K. El Emam. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. JCO Clin Cancer Inform. 2023;e2300116

Validity of population inferences

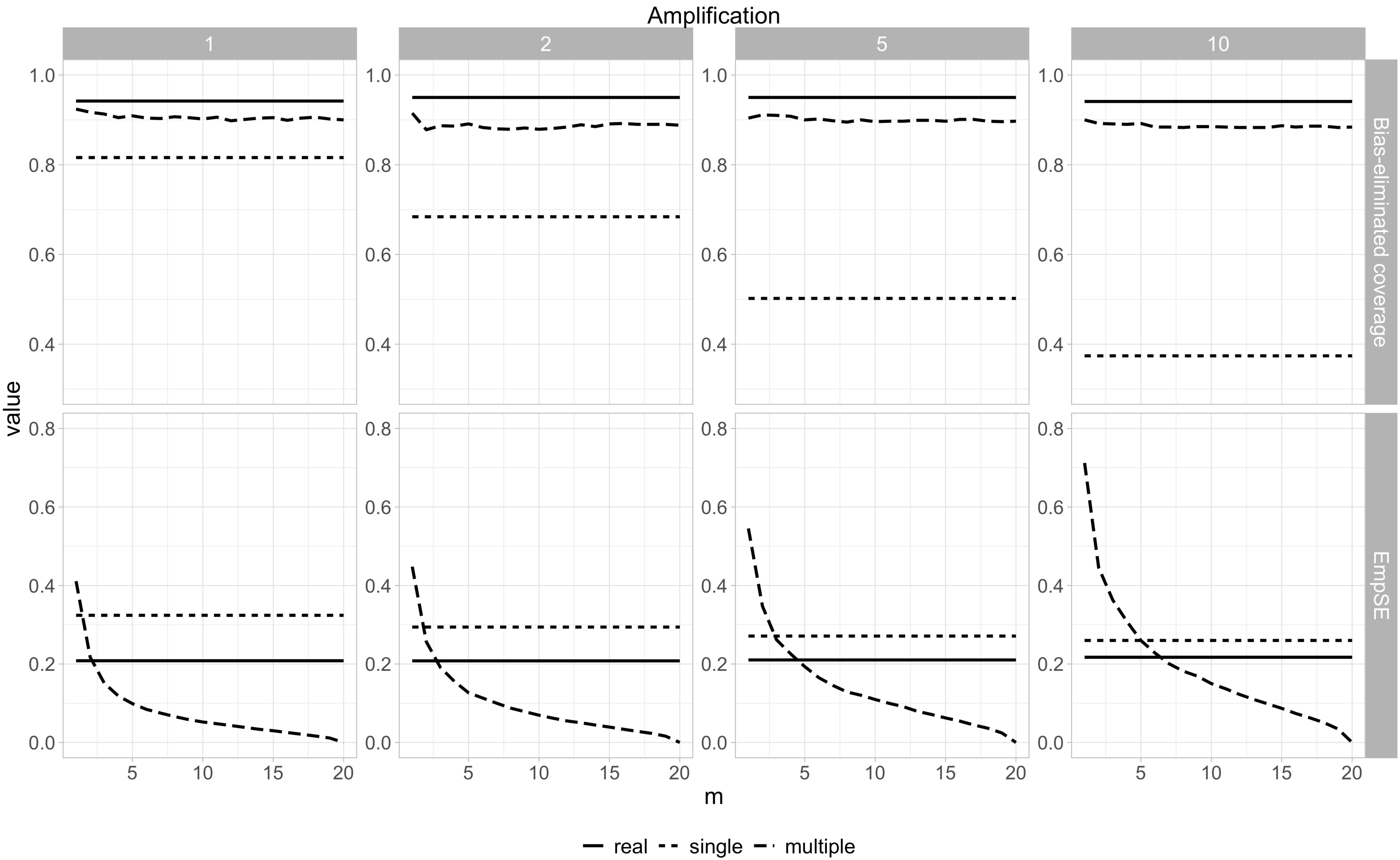


El Emam K, Mosquera L, Fang X, et al. An evaluation of the replicability of analyses using synthetic health data. Sci Rep. 2024;14:6978.



El Emam K, Mosquera L, Fang X, et al. An evaluation of the replicability of analyses using synthetic health data. Sci Rep. 2024;14:6978.

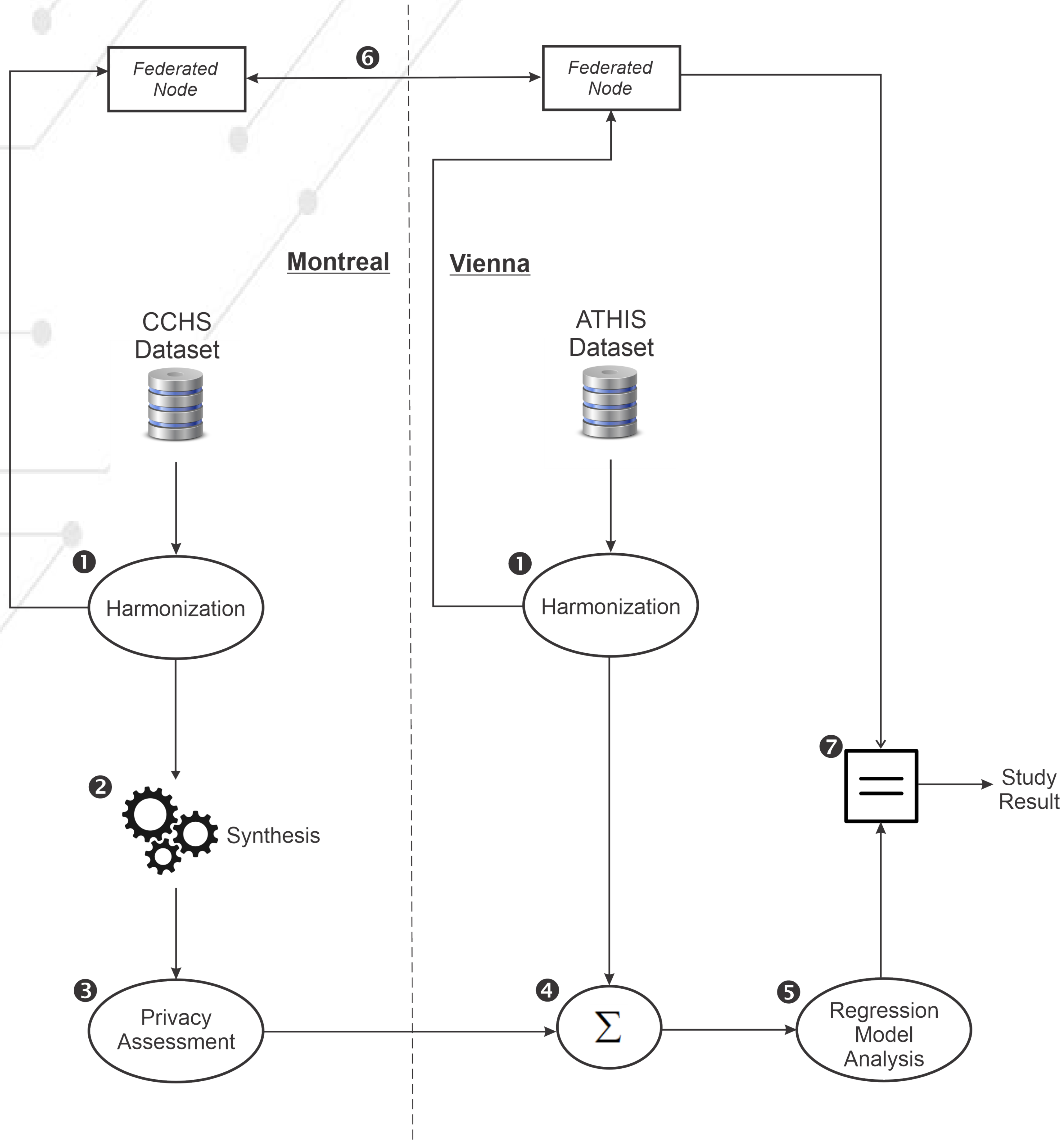
uOttawa



El Emam K, Mosquera L, Fang X, et al. An evaluation of the replicability of analyses using synthetic health data. Sci Rep. 2024;14:6978.

There is accumulating evidence that synthetic data is a good proxy for real data, but there isn't a single generative model that always performs well

Federated analysis using synthetic data - evaluation



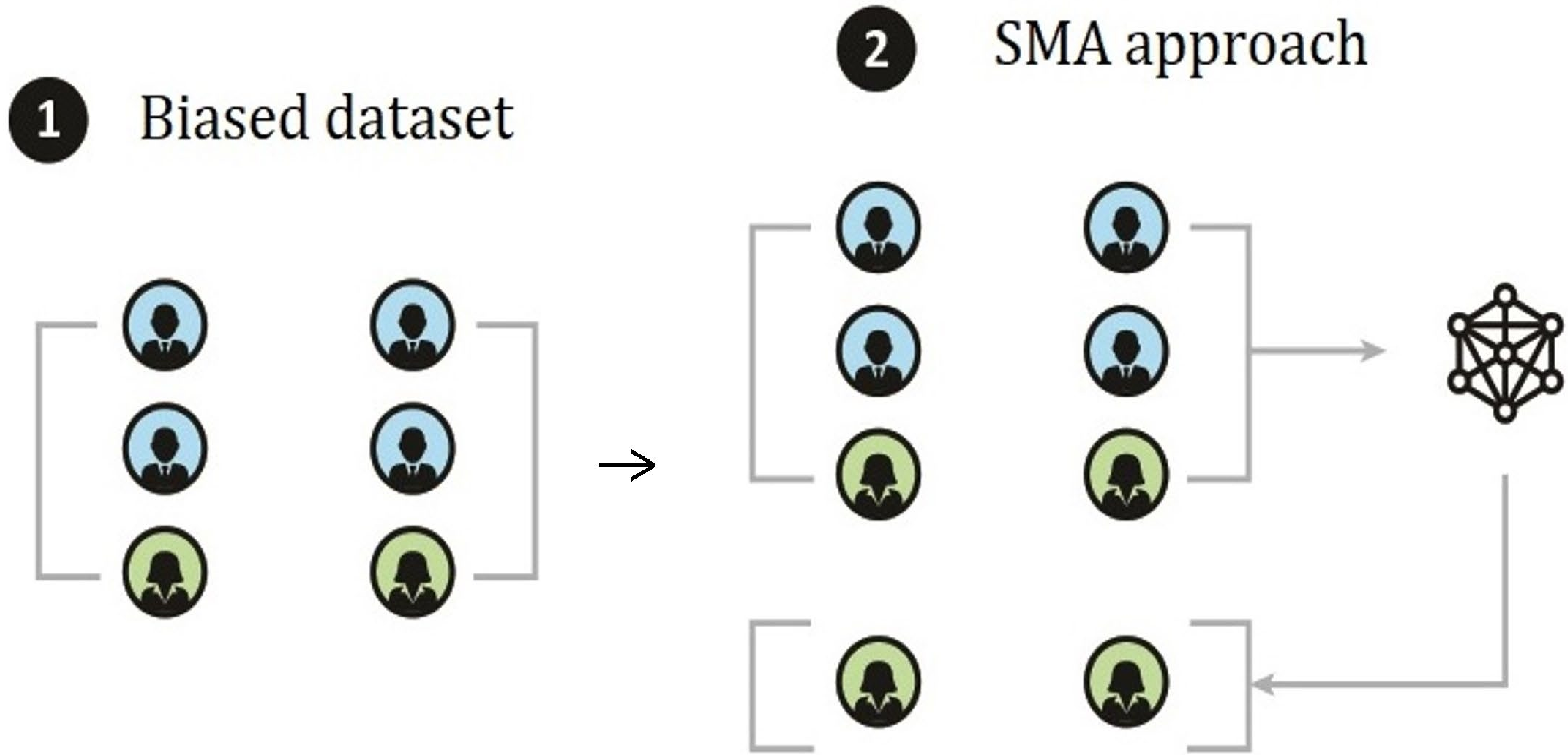
Z. Azizi, S. Lindner, Y. Shiba, V. Raparelli, C.M. Norris, K. Kublickiene, M.T. Herrero, A. Kautzky-Willer, P. Klimek, T. Gisinger, L. Pilote, K. El Emam: "A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health". Sci Rep 13: 11540, 2023.

Federated analysis using synthetic data - results

CANHEART score**	Federated analysis Regression coeff ***	Pooled analysis Regression coeff***
Sex (ref: male)	0.25 (0.23, 0.26)*	0.24 (0.23, 0.25)*
Education	0.04 (0.04, 0.05)*	0.04 (0.04, 0.05)*
Marital status (ref: Single)		
Divorced/widowed	−0.12 (−0.14, −0.09)*	−0.11 (−0.14, −0.09)*
Married	−0.15 (−0.17, −0.13)*	−0.16 (−0.18, −0.14)*
Household size	0.05 (0.04, 0.06)*	0.06 (0.05, 0.06)*
House income (reverse coded)	−0.08 (−0.09, −0.07)*	−0.09 (−0.10, −0.08)*
Immigrant(ref: No)	0.13 (0.12, 0.15)*	0.14 (0.13, 0.16)*
Age	−0.13 (−0.14, −0.13)*	−0.14 (−0.14, −0.13)*
Country (ref: CA)	−0.01 (−0.03, 0.002)	−0.02 (−0.04, 0.00)
R^2	0.163	0.165

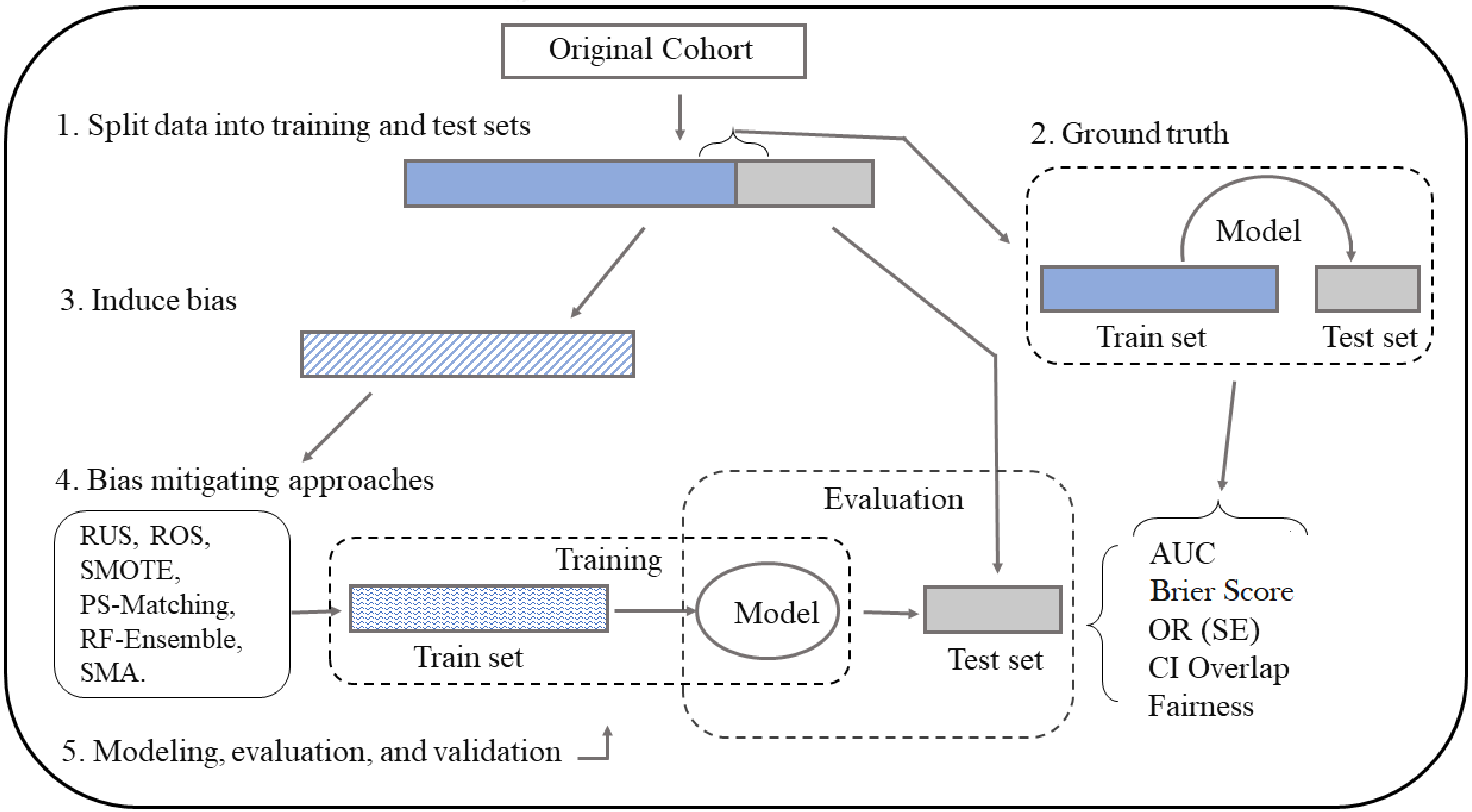
Z. Azizi, S. Lindner, Y. Shiba, V. Raparelli, C.M. Norris, K. Kublickiene, M.T. Herrero, A. Kautzky-Willer, P. Klimek, T. Gisinger, L. Pilote, K. El Emam: "A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health". Sci Rep 13: 11540, 2023.

Mitigating Bias

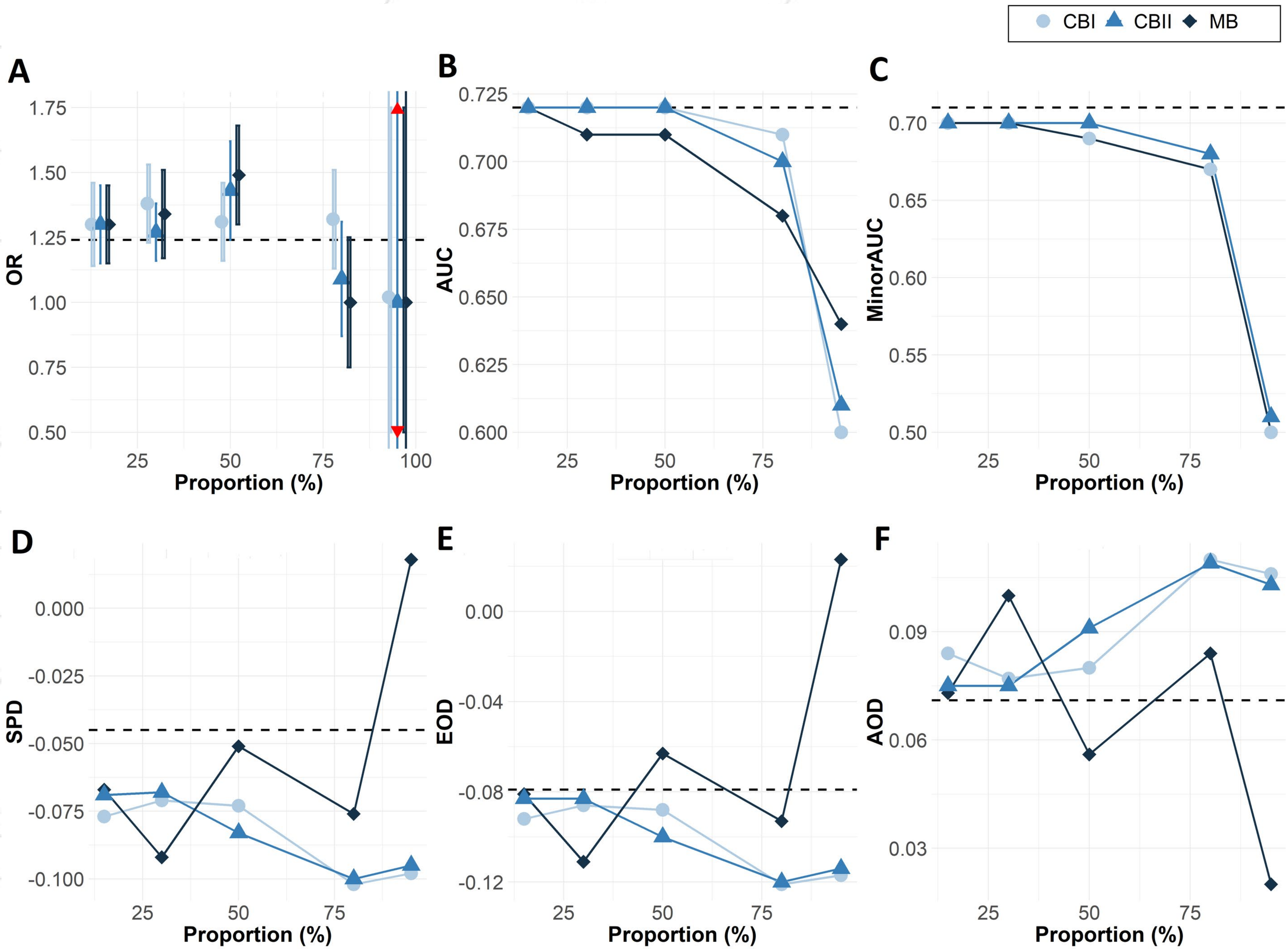


uOttawa

Bias evaluation using simulations

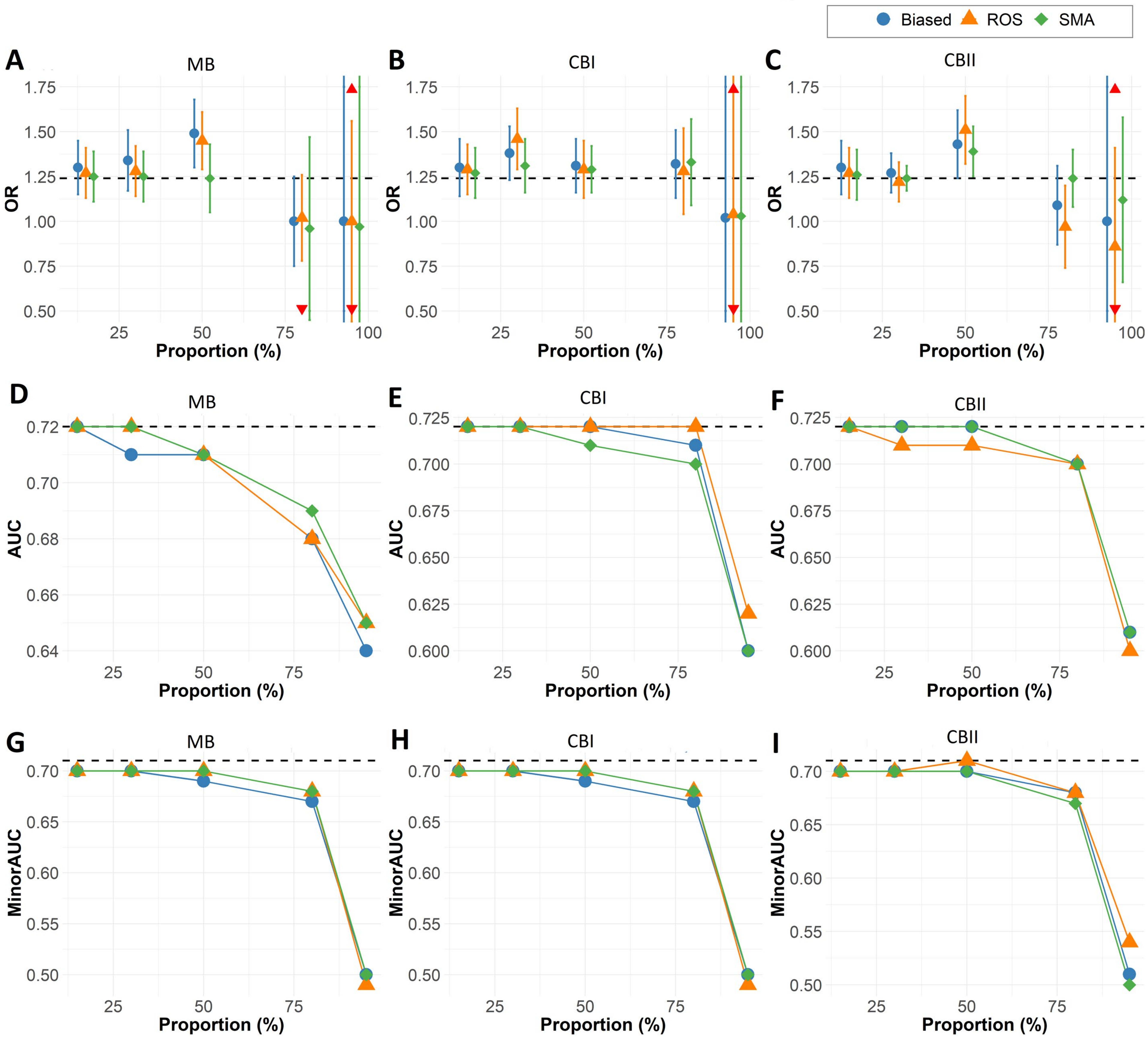


Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns. doi: 10.1016/j.patter.2024.100946



Data bias has an impact on model parameters and fairness

Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns. doi: 10.1016/j.patter.2024.100946



Synthetic data generation can mitigate low to medium bias better than other methods

Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns. doi: 10.1016/j.patter.2024.100946

Beyond data sharing, synthetic data can potentially help with federated analysis, and data bias mitigation

cc: tunnelarmr - https://www.flickr.com/photos/27311060@N00


QUESTIONS
cc: an untrained eye - https://www.flickr.com/photos/