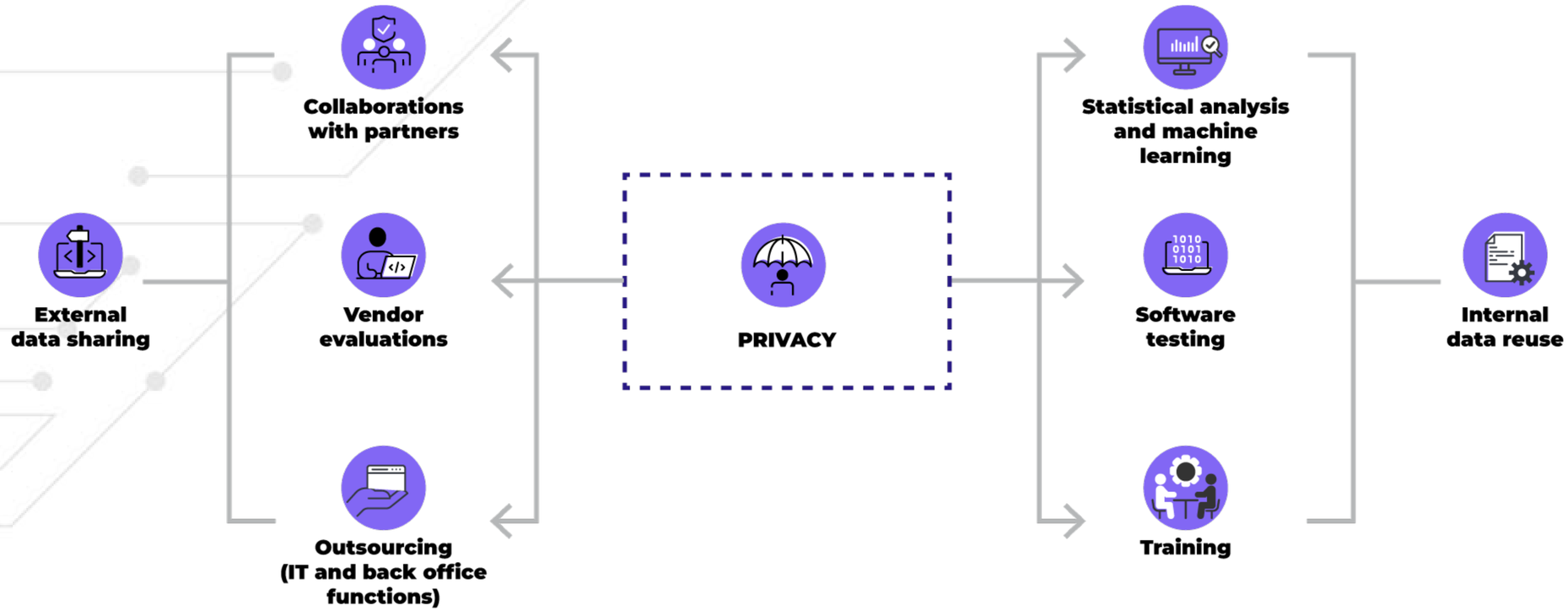
A Brief Overview of Synthetic Data Generation

Khaled El Emam
kelemam@ehealthinformation.ca

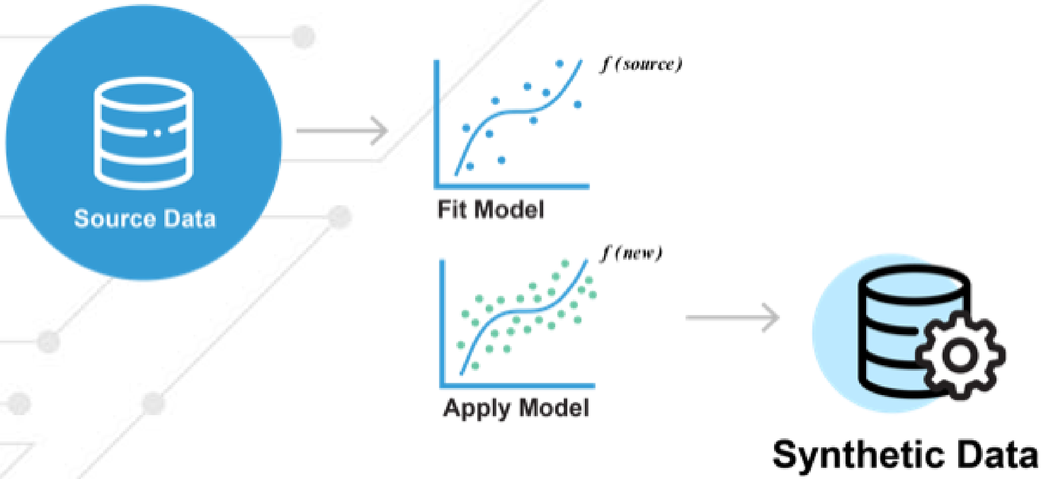
Space Opera Theatre

Privacy use cases



uOttawa

The Synthesis Process



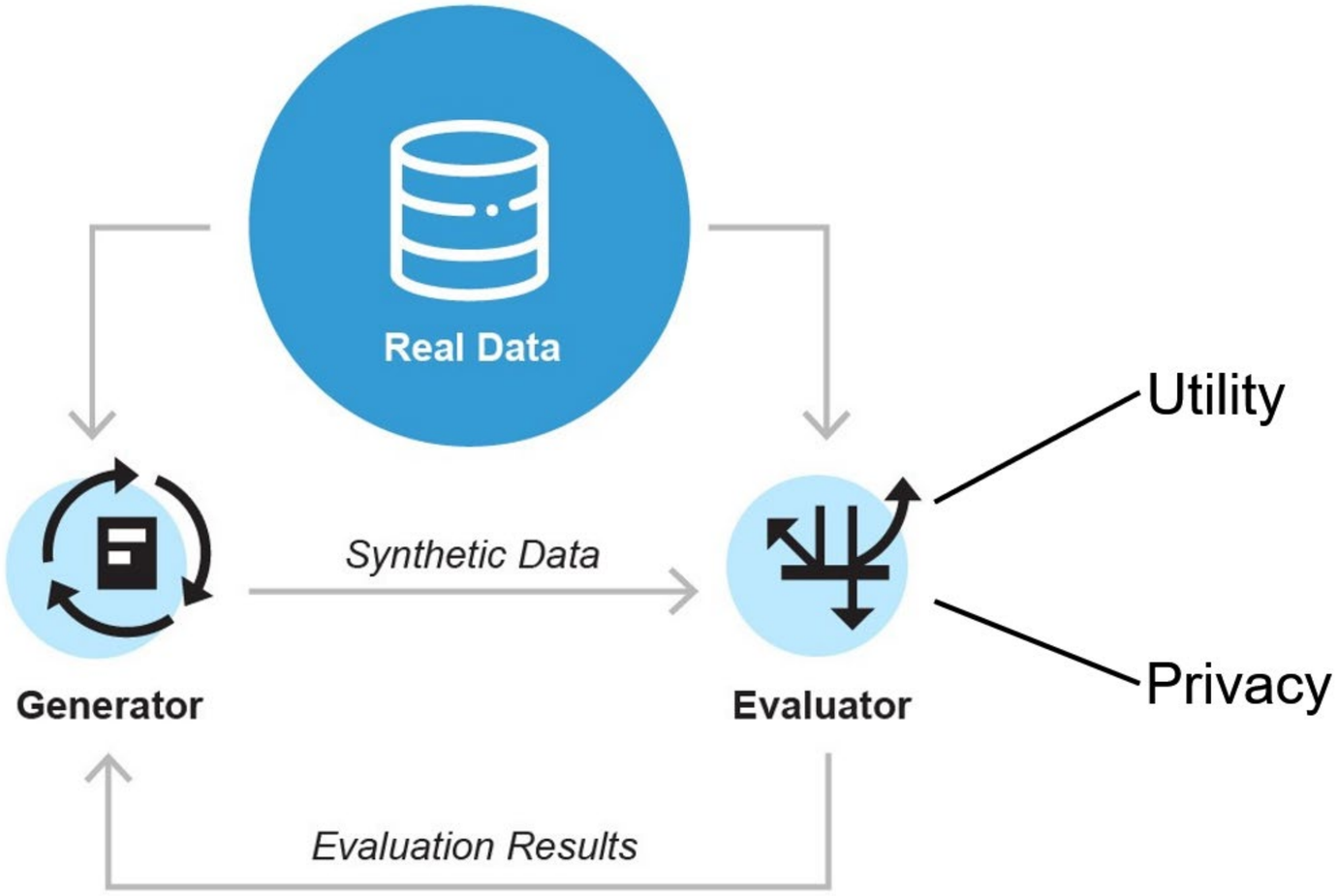
COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

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.

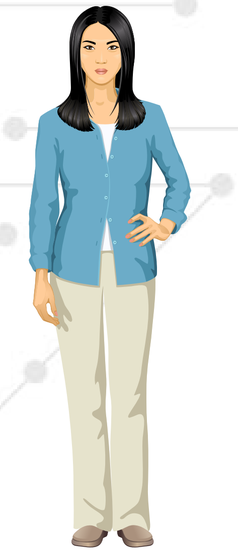
uOttawa

A combined loss of utility and privacy



uOttawa

Identity disclosure is when a person's identity is assigned to a record



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

Attribution disclosure: find a record in the synthetic data similar to a high risk real individual and 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

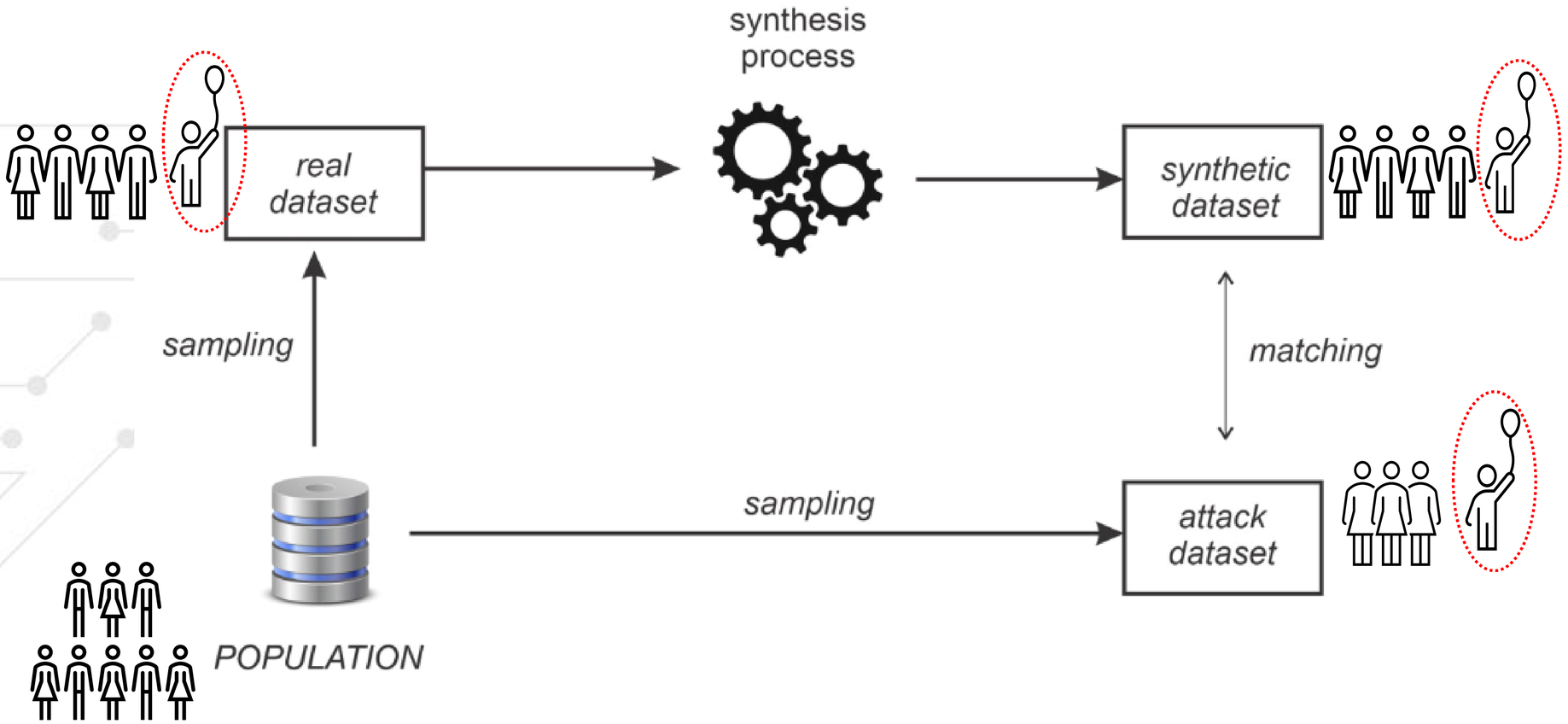
Example of evaluating attribution disclosure

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

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, p. e23139, Nov. 2020.

The process for a membership disclosure attack



uOttawa

Example of evaluating membership disclosure

Dataset	Dataset size	Risk
Trial#1 (NCT00041197): National Cancer Institute	773	-1.42
Trial#2 (NCT01124786): Clovis Oncology	367	-0.0137
Trial#3 (NCT00688740): Sanofi	746	-0.034
Trial#4 (NCT00113763): Amgen	370	-0.0137
Trial#5 (NCT00460265): Amgen	520	-0.0947
Trial#6 (NCT00119613): Amgen	479	-0.0322
Trial#7 (N0147)	1543	0.052

A commonly used risk threshold = 0.2

K. El Emam, L. Mosquera, and X. Fang, "Validating A Membership Disclosure Metric For Synthetic Health Data," JAMIA Open, vol. 5, no. 4, p. ooac083, Dec. 2022.

Assessing the utility of synthetic data

Generic utility

Show how similar synthetic data is to the real data it was generated from without referencing a specific analysis

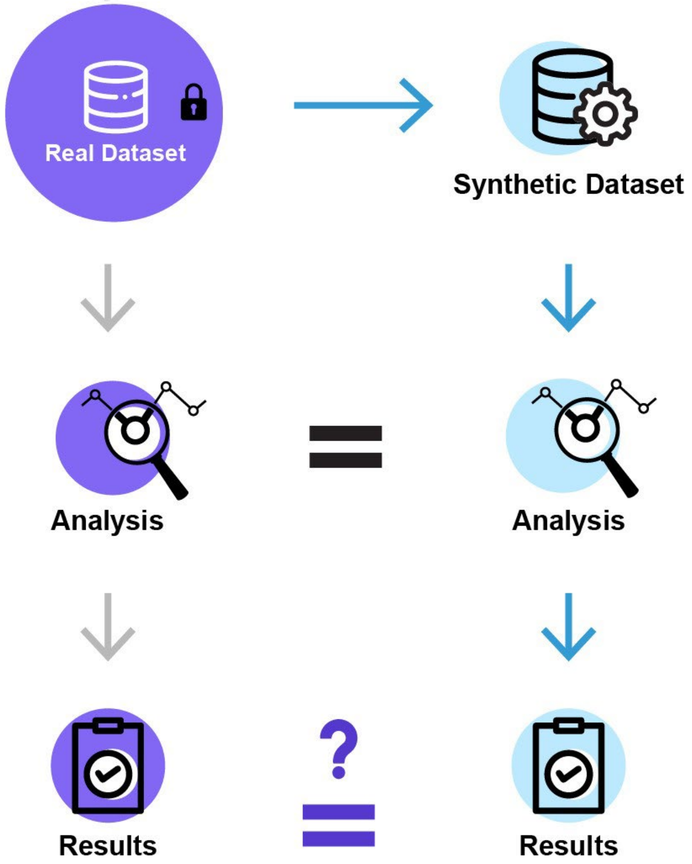
Workload aware utility

Illustrate how well synthetic data can be used as a drop-in replacement or proxy for real data for a specific analysis

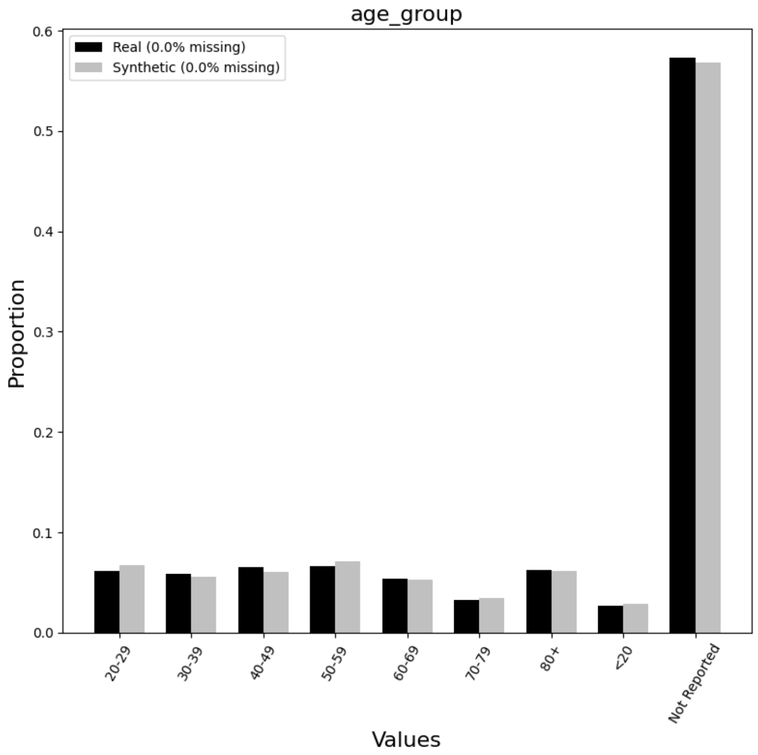
Expert discrimination

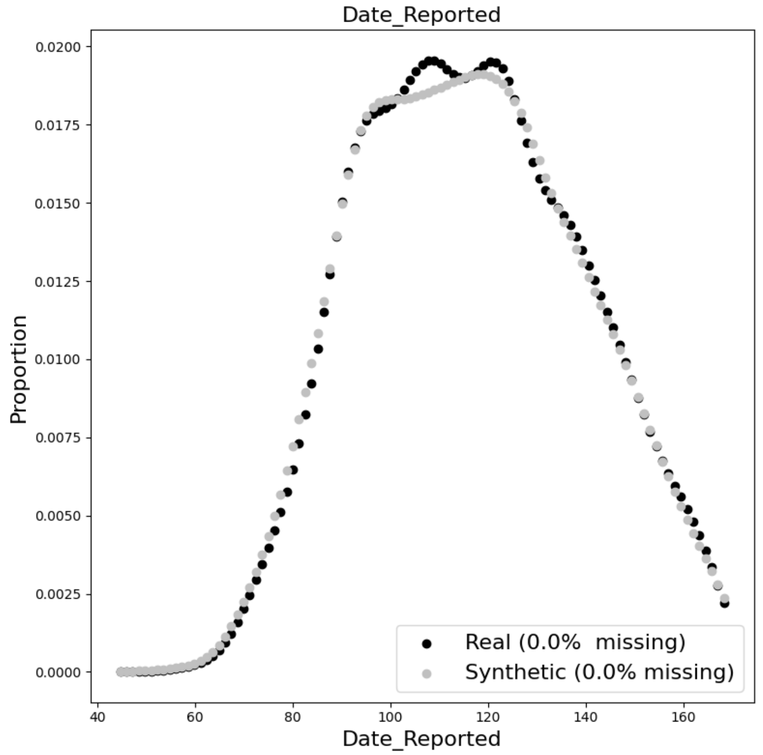
A clinician would manually examine multiple records and classify each one as real or synthetic

Reproducibility of results



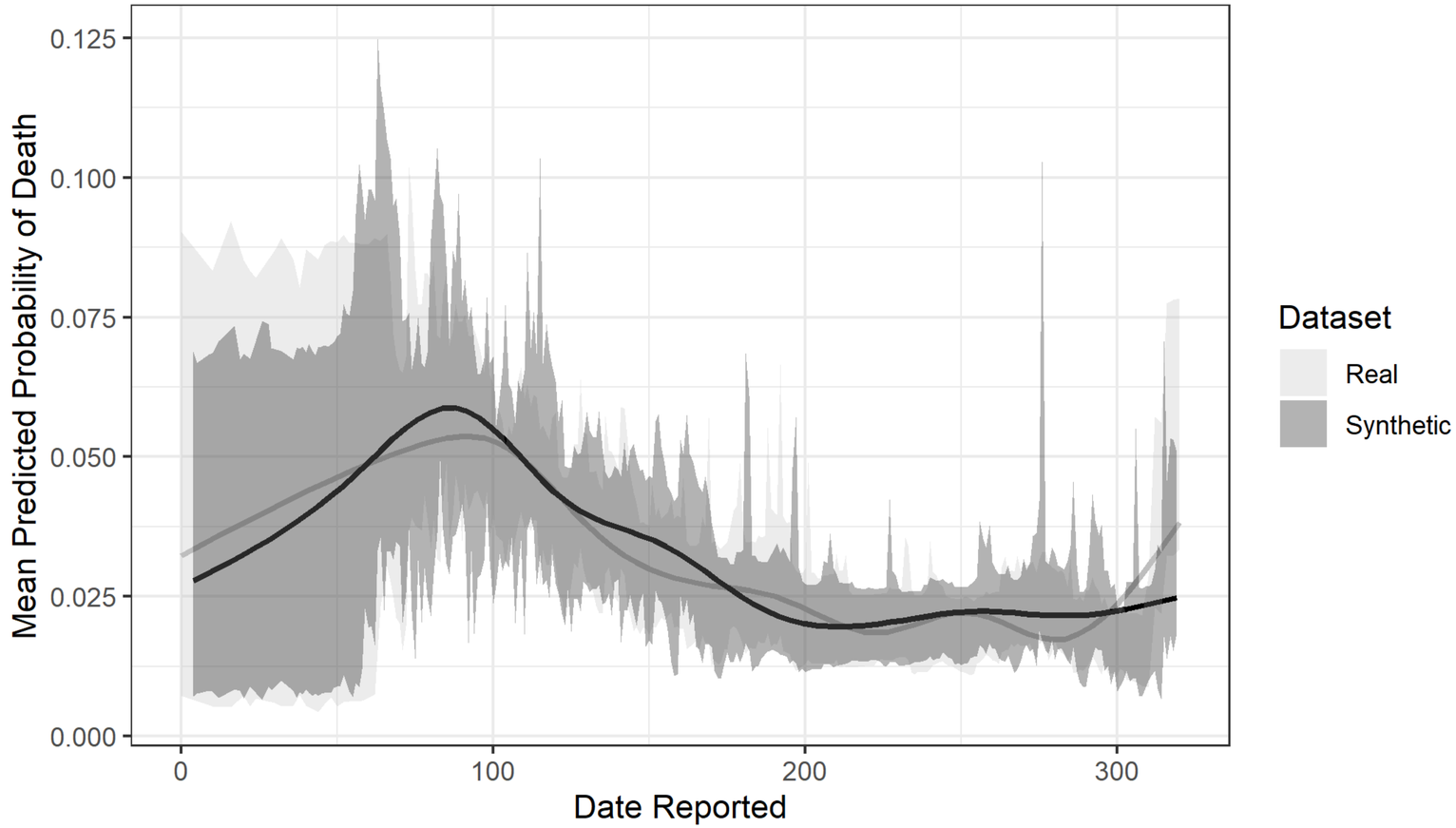
The distributions of real and synthetic datasets look similar





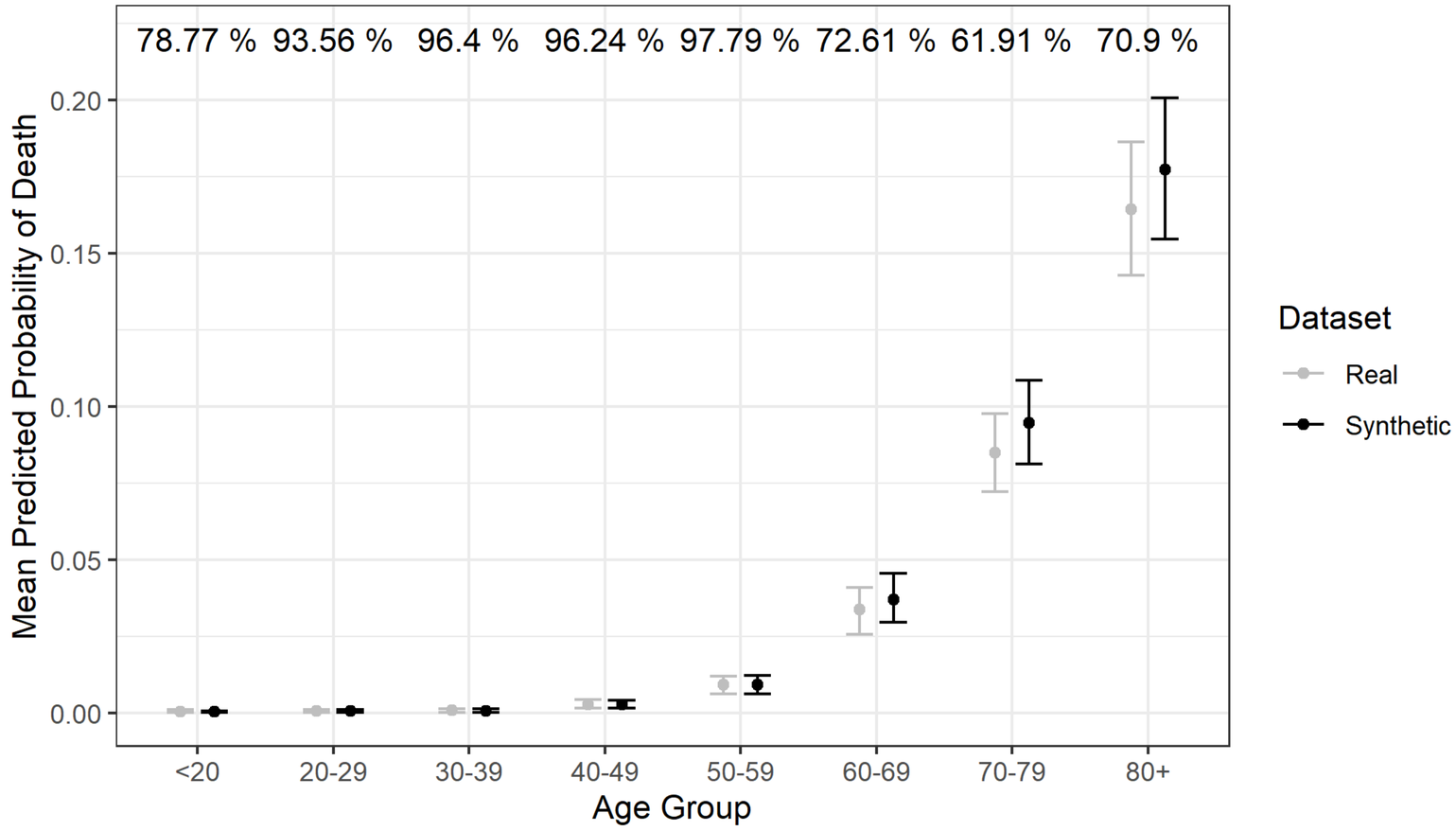
uOttawa

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.

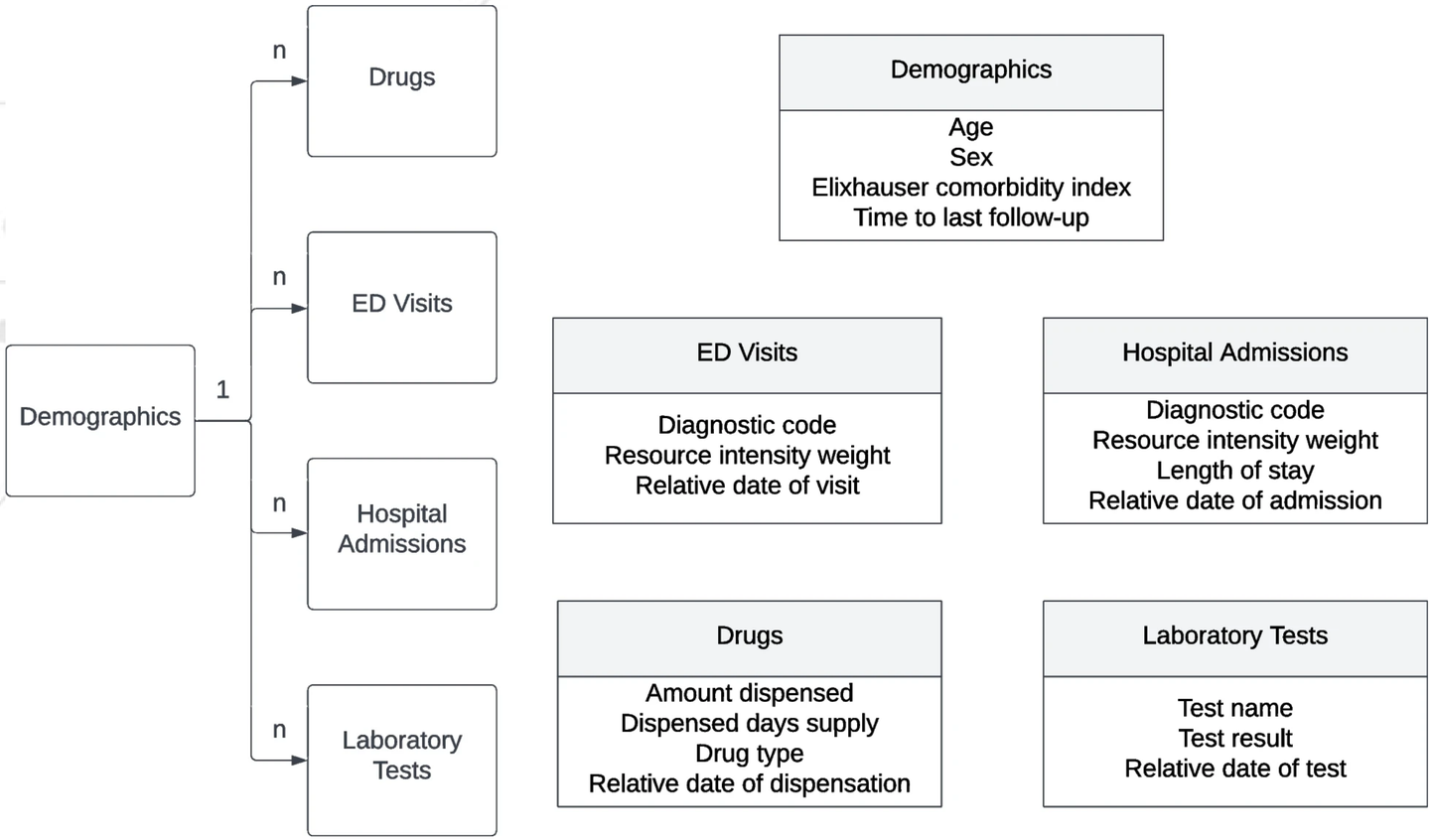
Comparing Real and Synthetic Data: Mortality By Age



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

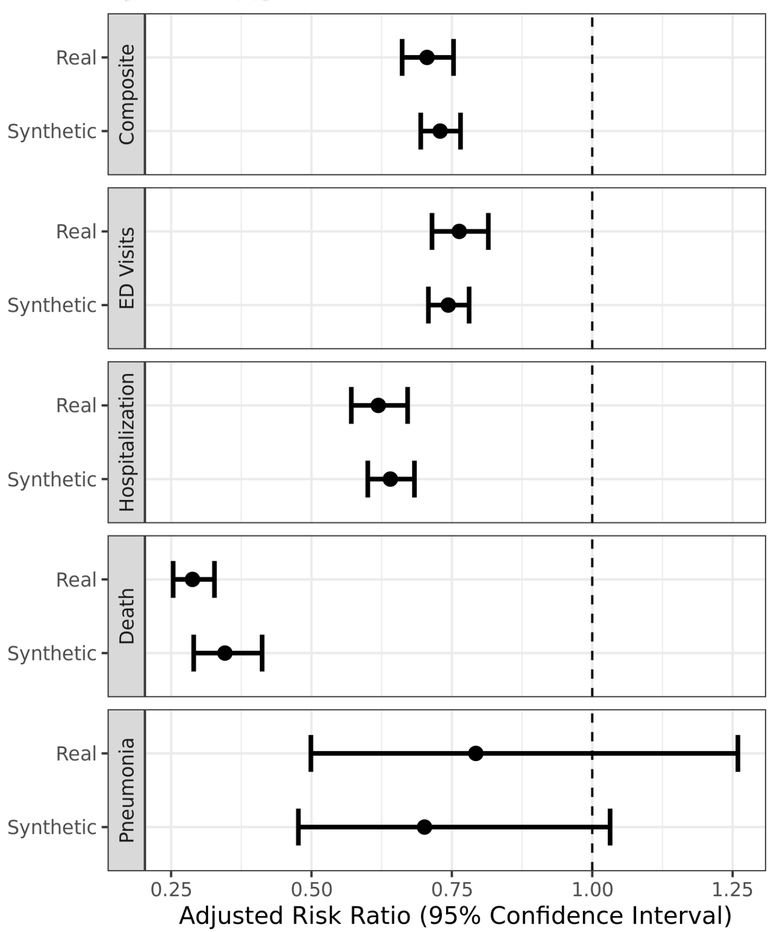
uOttawa

Longitudinal Health System Dataset



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.

Utility on eight breast cancer clinical trials

		SEQ			GAN			VAE		
Dataset	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	*	*	*	1	1	0.67
REaCT-ILIAD	218	1	1	0.99	1	1	0.85	1	0	0.74
REaCT-ZOL	211	1	**	0.98	1	**	0.88	0	**	0.61
REaCT-BTA	230	1	1	0.85	1	0	0.68	1	0	0.72
CCTG MA27	7576	1	1	0.90	1	1	0.62	1	1	0.82
SWOG 0307	6097	1	1	0.93	1	0	0.50	1	1	0.95
NSABP B34	3323	1	1	0.93	1	1	0.83	1	1	0.61

S. El Kababji et al., “Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Datasets,” J. Clin. Oncol. Clin. Cancer Inform., (accepted).

Attribution disclosure on eight breast cancer clinical trials

	SEQ		GAN		VAE	
Dataset	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

S. El Kababji et al., "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Datasets," J. Clin. Oncol. Clin. Cancer Inform., (accepted).

Membership disclosure on eight breast cancer clinical trials

		SEQ		GAN		VAE	
Dataset	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

S. El Kababji et al., "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Datasets," J. Clin. Oncol. Clin. Cancer Inform., (accepted).


QUESTIONS