FACULTY OF MEDICINE AND HEALTH SCIENCES

The Real Deal Behind the Artificial Appeal

Inferential Utility of Tabular Synthetic Data

Decruyenaere A*, Dehaene H*, Rabaey P, et al.

poster #406

UAI conference, July 18th, 2024

GHENT UNIVERSITY

UZ GENT

SYNDARA

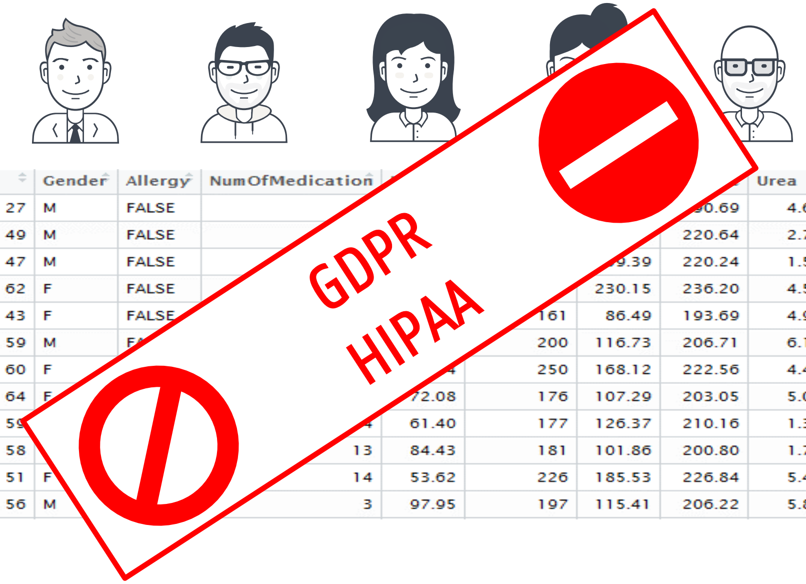
Naive statistical inference from synthetic data cannot be trusted

Original data

Age	Gender	Allergy	NumOfMedication	HDL	Colesterol	LDL	Glucose	Urea	ICD
27	M	FALSE	13	83.61	154	80.71	190.69	4.60	A18
49	M	FALSE	5	68.10	212	160.85	220.64	2.70	A18
47	M	FALSE	1	88.38	228	159.39	220.24	1.50	A02
62	F	FALSE	1	62.77	286	230.15	236.20	4.50	A18
43	F	FALSE	4	83.44	161	86.49	193.69	4.90	A18
59	M	FALSE	1	95.50	200	116.73	206.71	6.15	A18
60	F	FALSE	14	90.84	250	168.12	222.56	4.44	A18
64	F	FALSE	6	72.08	176	107.29	203.05	5.08	A18
59	M	FALSE	4	61.40	177	126.37	210.16	1.37	A02
58	M	FALSE	13	84.43	181	101.86	200.80	1.72	A04
51	F	TRUE	14	53.62	226	185.53	226.84	5.43	A18
56	M	FALSE	3	97.95	197	115.41	206.22	5.81	A18

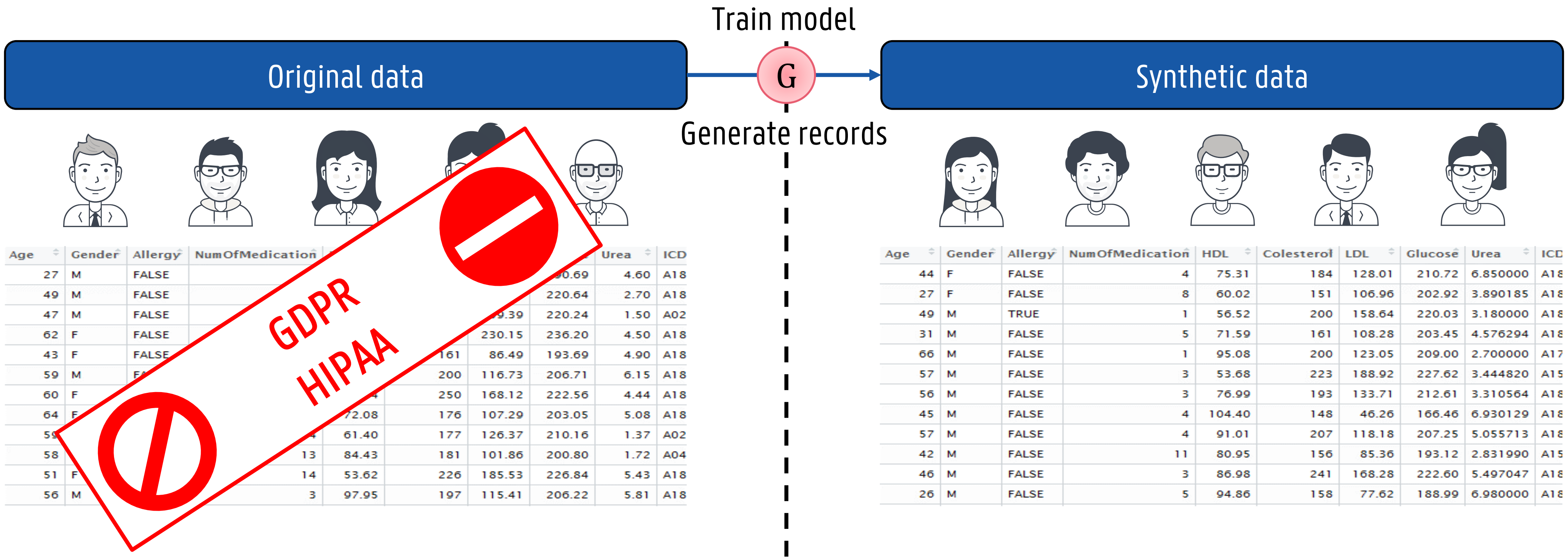
Naive statistical inference from synthetic data cannot be trusted

Original data



Age	Gender	Allergy	NumOfMedication	[illegible]	[illegible]	[illegible]	[illegible]	Urea	ICD
27	M	FALSE		[illegible]	[illegible]	[illegible]	[illegible]	4.60	A18
49	M	FALSE		[illegible]	[illegible]	[illegible]	220.64	2.70	A18
47	M	FALSE		[illegible]	[illegible]	[illegible]	220.24	1.50	A02
62	F	FALSE	[illegible]	[illegible]	[illegible]	230.15	236.20	4.50	A18
43	F	FALSE	[illegible]	[illegible]	161	86.49	193.69	4.90	A18
59	M	[illegible]	[illegible]	[illegible]	200	116.73	206.71	6.15	A18
60	F	[illegible]	[illegible]	[illegible]	250	168.12	222.56	4.44	A18
64	F	[illegible]	[illegible]	72.08	176	107.29	203.05	5.08	A18
[illegible]	[illegible]	[illegible]	[illegible]	61.40	177	126.37	210.16	1.37	A02
58	[illegible]	[illegible]	13	84.43	181	101.86	200.80	1.72	A04
51	F	[illegible]	14	53.62	226	185.53	226.84	5.43	A18
56	M	[illegible]	3	97.95	197	115.41	206.22	5.81	A18

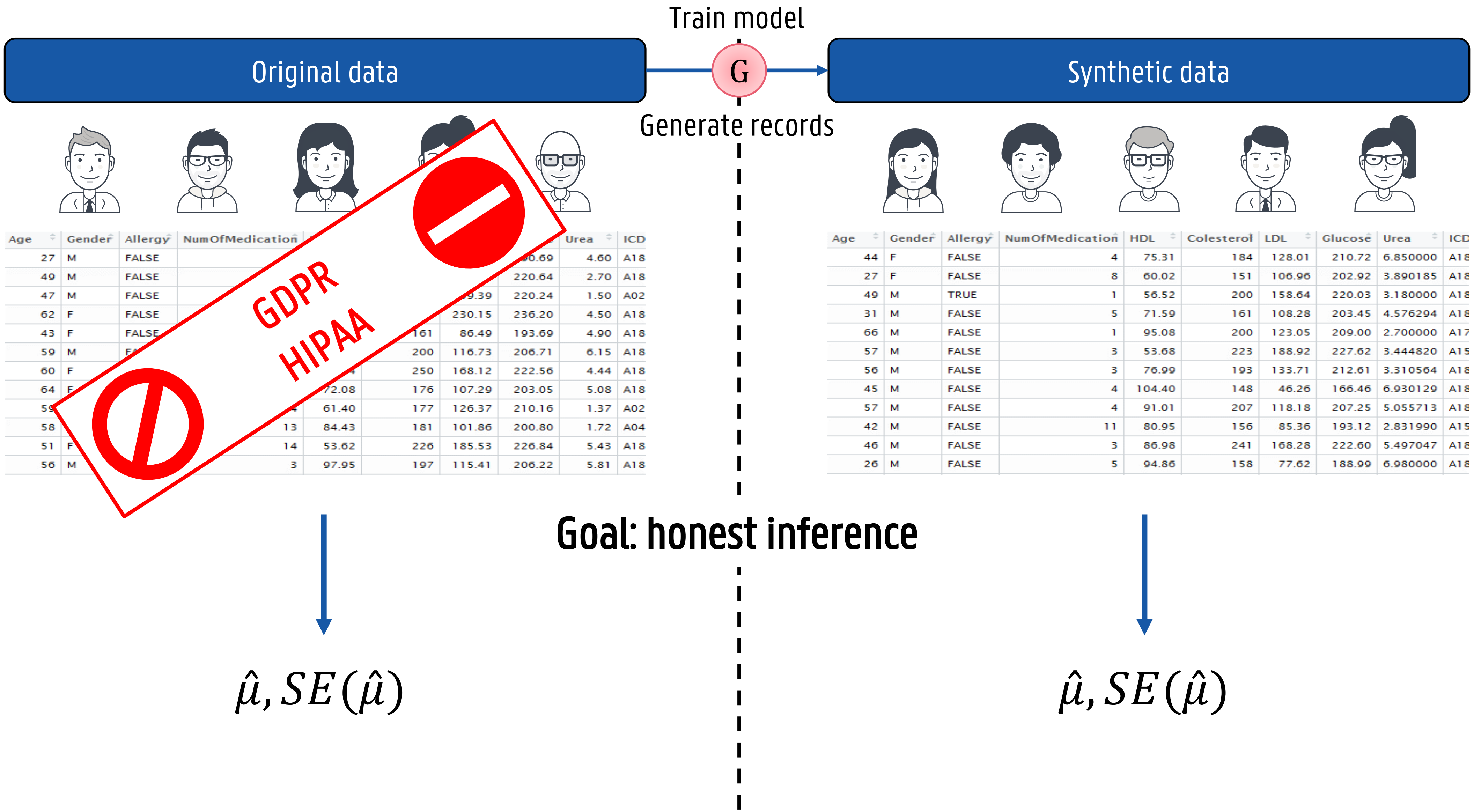
Naive statistical inference from synthetic data cannot be trusted



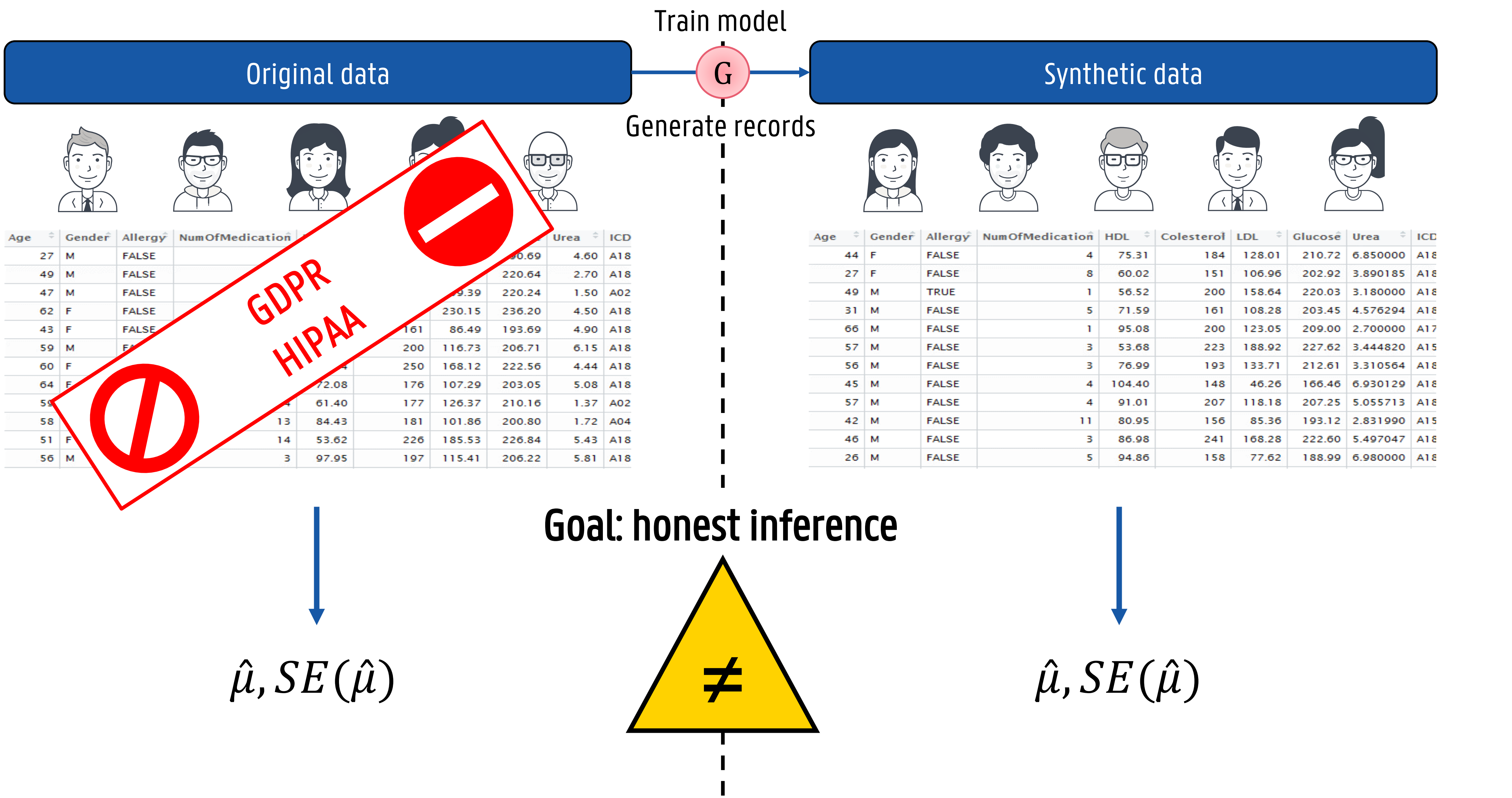
Age	Gender	Allergy	NumOfMedication	[illegible]	[illegible]	[illegible]	[illegible]	Urea	ICD
27	M	FALSE	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.60	A18
49	M	FALSE	[illegible]	[illegible]	[illegible]	[illegible]	220.64	2.70	A18
47	M	FALSE	[illegible]	[illegible]	[illegible]	[illegible]	220.24	1.50	A02
62	F	FALSE	[illegible]	[illegible]	[illegible]	230.15	236.20	4.50	A18
43	F	FALSE	[illegible]	[illegible]	161	86.49	193.69	4.90	A18
59	M	[illegible]	[illegible]	[illegible]	200	116.73	206.71	6.15	A18
60	F	[illegible]	[illegible]	[illegible]	250	168.12	222.56	4.44	A18
64	F	[illegible]	[illegible]	72.08	176	107.29	203.05	5.08	A18
59	[illegible]	[illegible]	4	61.40	177	126.37	210.16	1.37	A02
58	[illegible]	[illegible]	13	84.43	181	101.86	200.80	1.72	A04
51	F	[illegible]	14	53.62	226	185.53	226.84	5.43	A18
56	M	[illegible]	3	97.95	197	115.41	206.22	5.81	A18

Age	Gender	Allergy	NumOfMedication	HDL	Colesterol	LDL	Glucose	Urea	ICD
44	F	FALSE	4	75.31	184	128.01	210.72	6.850000	A18
27	F	FALSE	8	60.02	151	106.96	202.92	3.890185	A18
49	M	TRUE	1	56.52	200	158.64	220.03	3.180000	A18
31	M	FALSE	5	71.59	161	108.28	203.45	4.576294	A18
66	M	FALSE	1	95.08	200	123.05	209.00	2.700000	A17
57	M	FALSE	3	53.68	223	188.92	227.62	3.444820	A15
56	M	FALSE	3	76.99	193	133.71	212.61	3.310564	A18
45	M	FALSE	4	104.40	148	46.26	166.46	6.930129	A18
57	M	FALSE	4	91.01	207	118.18	207.25	5.055713	A18
42	M	FALSE	11	80.95	156	85.36	193.12	2.831990	A15
46	M	FALSE	3	86.98	241	168.28	222.60	5.497047	A18
26	M	FALSE	5	94.86	158	77.62	188.99	6.980000	A18

Naive statistical inference from synthetic data cannot be trusted



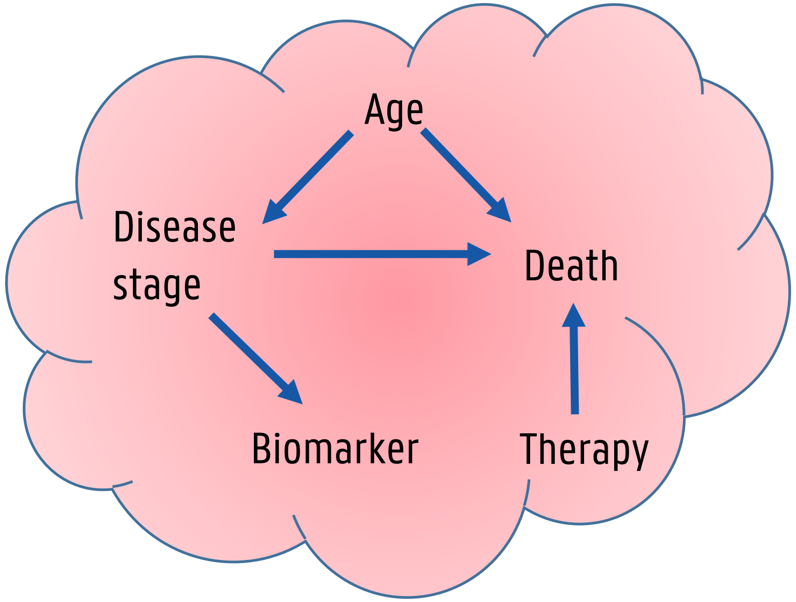
Naive statistical inference from synthetic data cannot be trusted



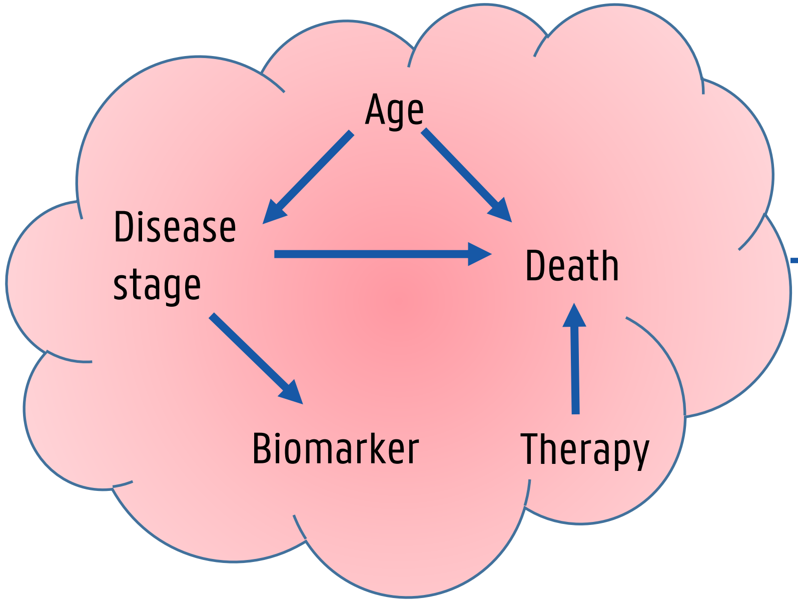
Age	Gender	Allergy	NumOfMedication					Urea	ICD
27	M	FALSE						4.60	A18
49	M	FALSE					220.64	2.70	A18
47	M	FALSE					220.24	1.50	A02
62	F	FALSE				230.15	236.20	4.50	A18
43	F	FALSE			161	86.49	193.69	4.90	A18
59	M				200	116.73	206.71	6.15	A18
60	F				250	168.12	222.56	4.44	A18
64					176	107.29	203.05	5.08	A18
59				61.40	177	126.37	210.16	1.37	A02
58			13	84.43	181	101.86	200.80	1.72	A04
51	F		14	53.62	226	185.53	226.84	5.43	A18
56	M		3	97.95	197	115.41	206.22	5.81	A18

Age	Gender	Allergy	NumOfMedication	HDL	Colesterol	LDL	Glucose	Urea	ICD
44	F	FALSE	4	75.31	184	128.01	210.72	6.850000	A18
27	F	FALSE	8	60.02	151	106.96	202.92	3.890185	A18
49	M	TRUE	1	56.52	200	158.64	220.03	3.180000	A18
31	M	FALSE	5	71.59	161	108.28	203.45	4.576294	A18
66	M	FALSE	1	95.08	200	123.05	209.00	2.700000	A17
57	M	FALSE	3	53.68	223	188.92	227.62	3.444820	A15
56	M	FALSE	3	76.99	193	133.71	212.61	3.310564	A18
45	M	FALSE	4	104.40	148	46.26	166.46	6.930129	A18
57	M	FALSE	4	91.01	207	118.18	207.25	5.055713	A18
42	M	FALSE	11	80.95	156	85.36	193.12	2.831990	A15
46	M	FALSE	3	86.98	241	168.28	222.60	5.497047	A18
26	M	FALSE	5	94.86	158	77.62	188.99	6.980000	A18

We investigated the behaviour of estimators in synthetic data



We investigated the behaviour of estimators in synthetic data

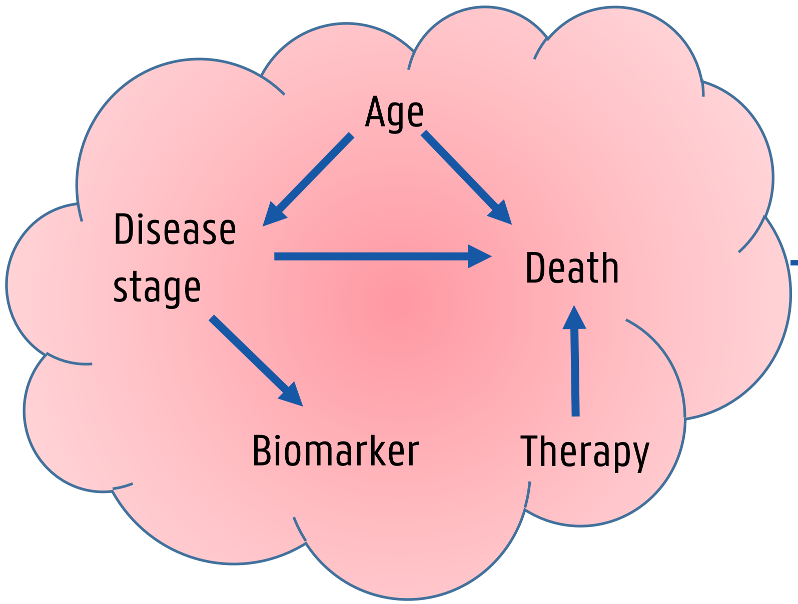


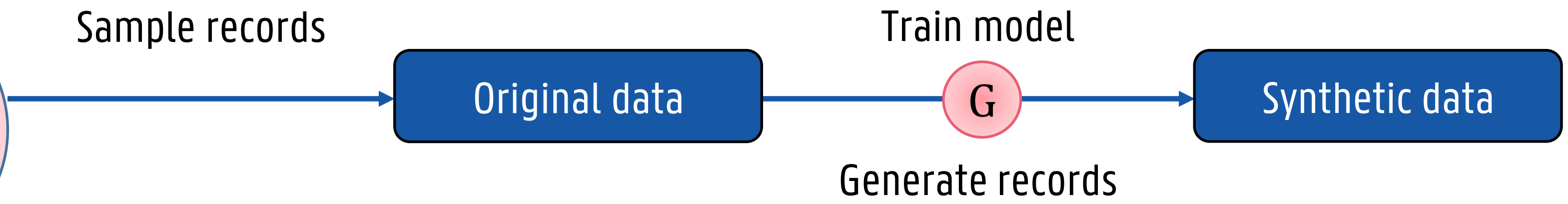
Sample records

Original data

27	M	FALSE	13	83.61
49	M	FALSE	5	68.10
47	M	FALSE	1	88.38
62	F	FALSE	1	62.77
43	F	FALSE	4	83.44
59	M	FALSE	1	95.50
60	F	FALSE	14	90.84
64	F	FALSE	6	72.08
59	M	FALSE	4	61.40
58	M	FALSE	13	84.43
51	F	TRUE	14	53.62
56	M	FALSE	3	97.95

We investigated the behaviour of estimators in synthetic data

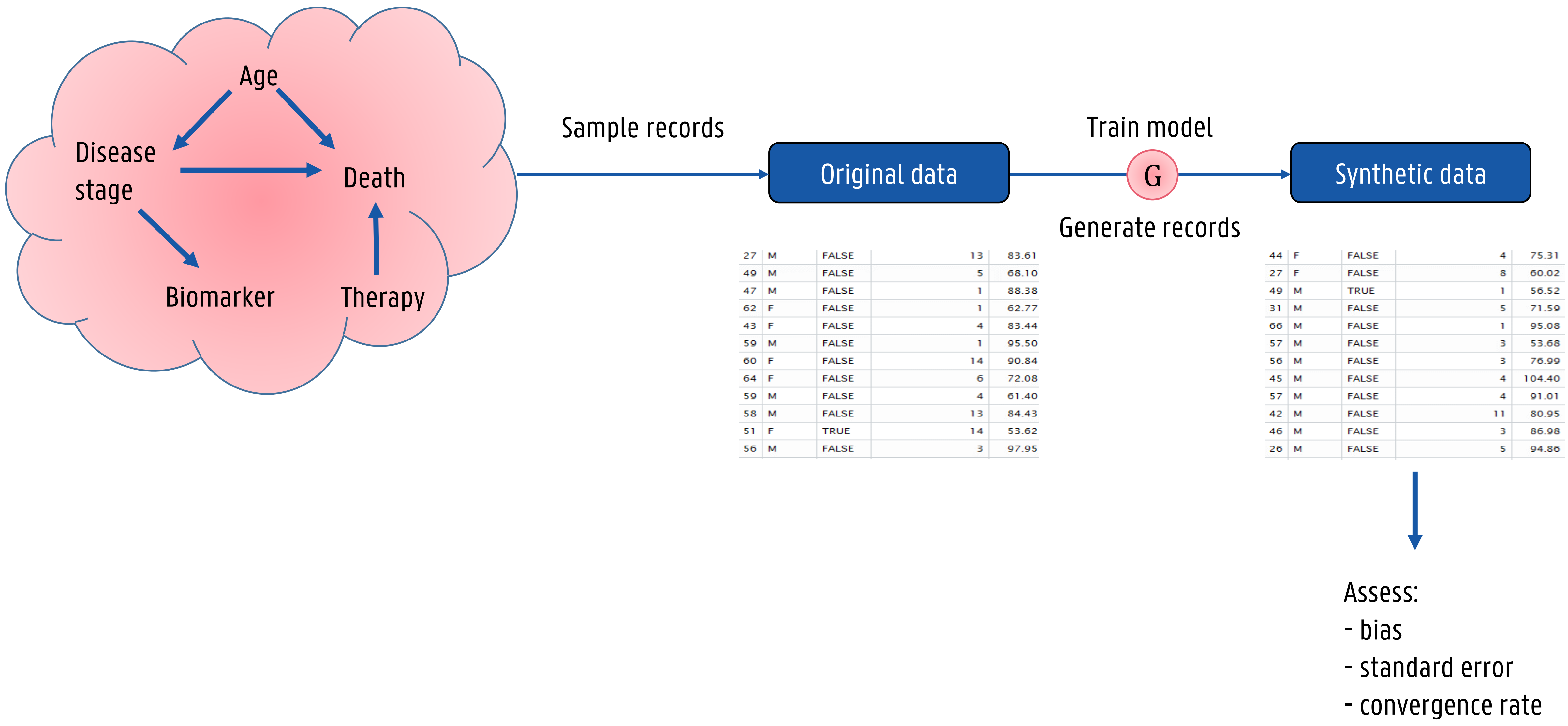




27	M	FALSE	13	83.61
49	M	FALSE	5	68.10
47	M	FALSE	1	88.38
62	F	FALSE	1	62.77
43	F	FALSE	4	83.44
59	M	FALSE	1	95.50
60	F	FALSE	14	90.84
64	F	FALSE	6	72.08
59	M	FALSE	4	61.40
58	M	FALSE	13	84.43
51	F	TRUE	14	53.62
56	M	FALSE	3	97.95

44	F	FALSE	4	75.31
27	F	FALSE	8	60.02
49	M	TRUE	1	56.52
31	M	FALSE	5	71.59
66	M	FALSE	1	95.08
57	M	FALSE	3	53.68
56	M	FALSE	3	76.99
45	M	FALSE	4	104.40
57	M	FALSE	4	91.01
42	M	FALSE	11	80.95
46	M	FALSE	3	86.98
26	M	FALSE	5	94.86

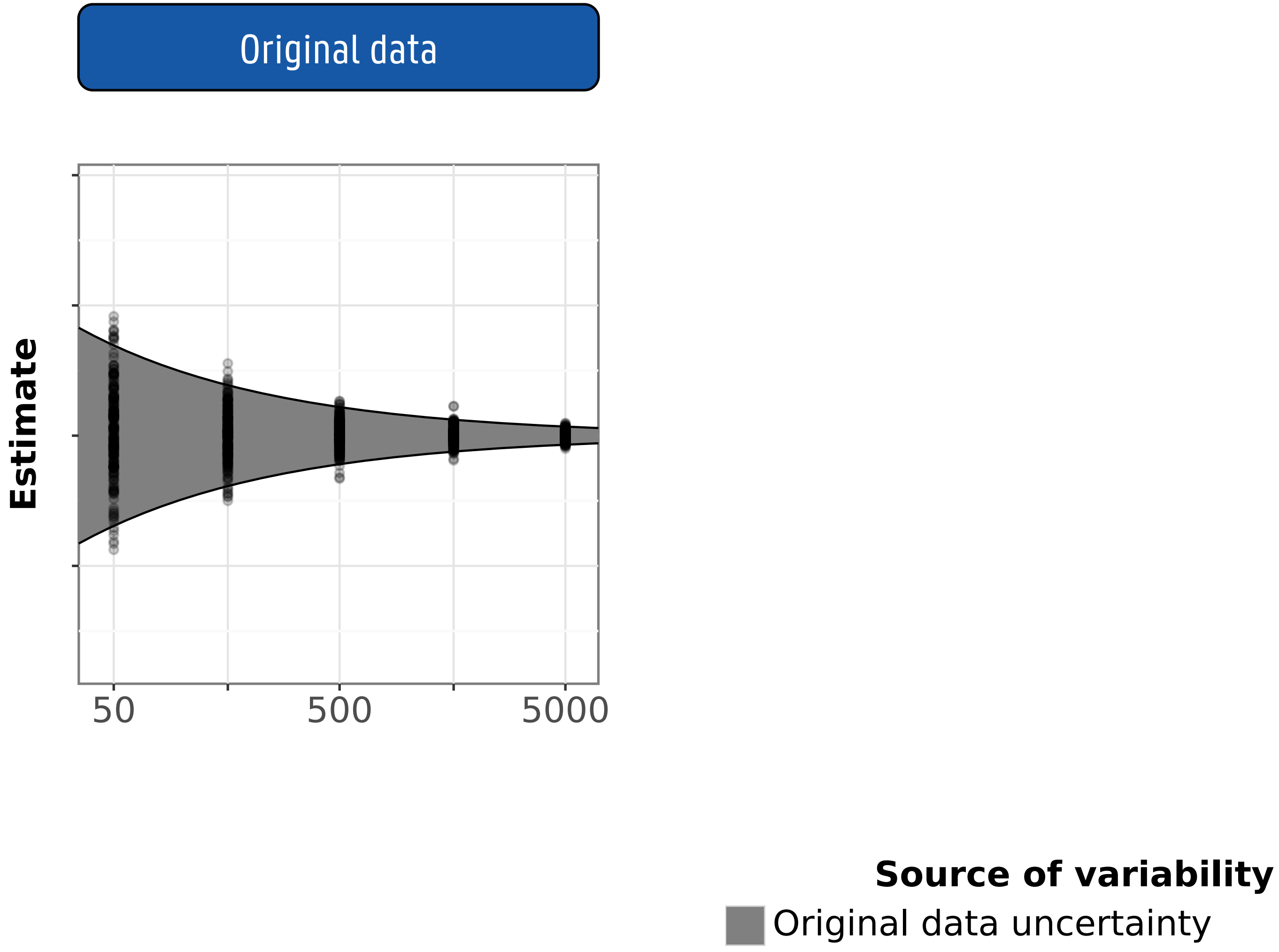
We investigated the behaviour of estimators in synthetic data



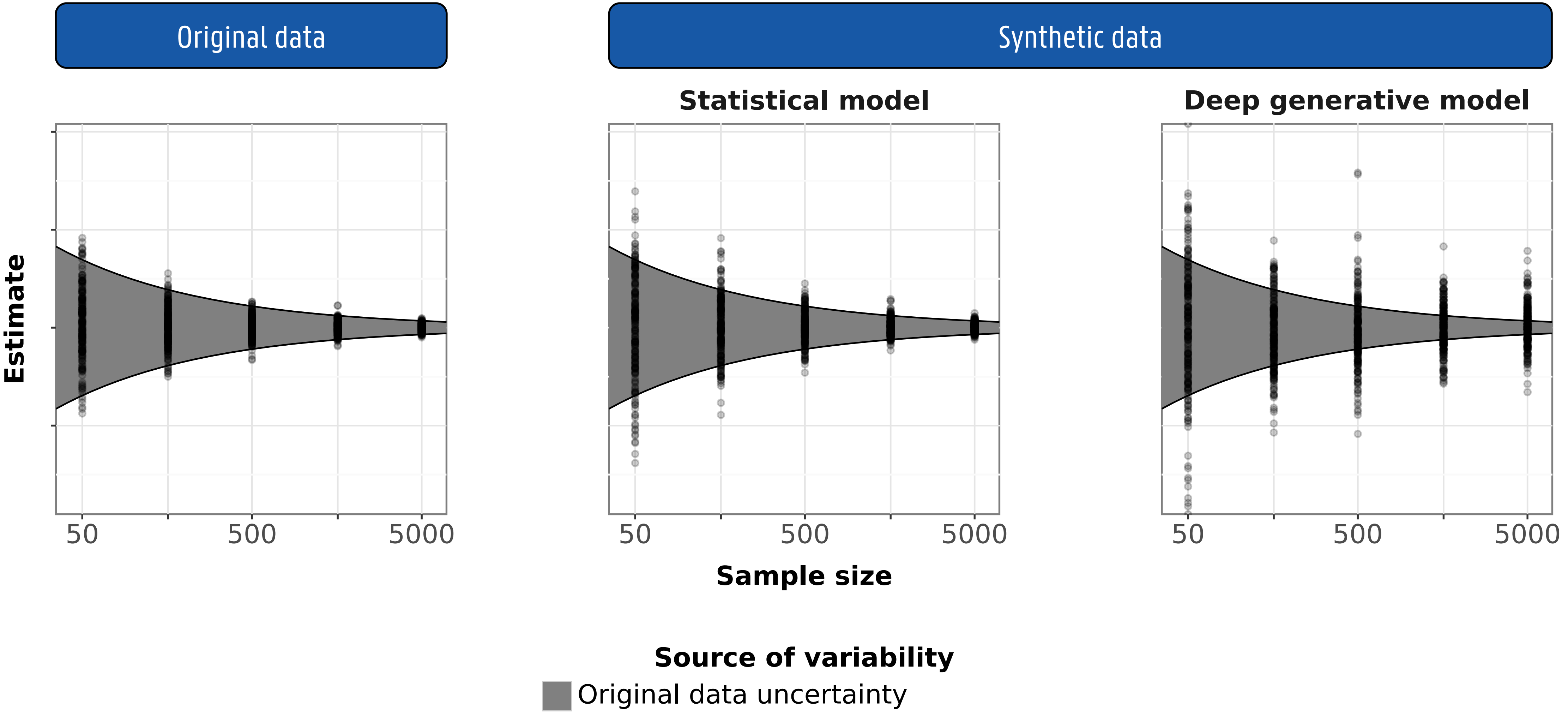
27	M	FALSE	13	83.61
49	M	FALSE	5	68.10
47	M	FALSE	1	88.38
62	F	FALSE	1	62.77
43	F	FALSE	4	83.44
59	M	FALSE	1	95.50
60	F	FALSE	14	90.84
64	F	FALSE	6	72.08
59	M	FALSE	4	61.40
58	M	FALSE	13	84.43
51	F	TRUE	14	53.62
56	M	FALSE	3	97.95

44	F	FALSE	4	75.31
27	F	FALSE	8	60.02
49	M	TRUE	1	56.52
31	M	FALSE	5	71.59
66	M	FALSE	1	95.08
57	M	FALSE	3	53.68
56	M	FALSE	3	76.99
45	M	FALSE	4	104.40
57	M	FALSE	4	91.01
42	M	FALSE	11	80.95
46	M	FALSE	3	86.98
26	M	FALSE	5	94.86

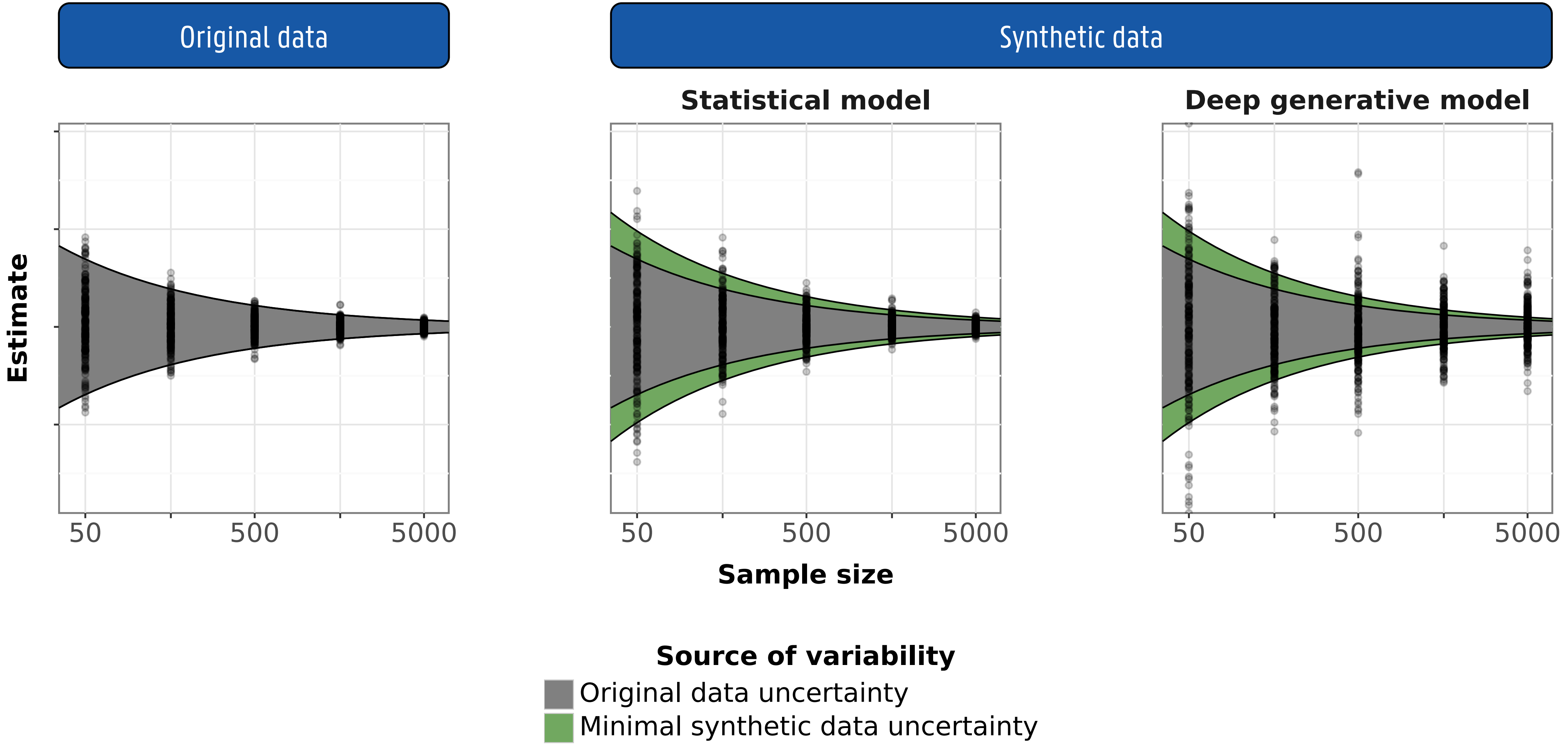
Original standard errors are insufficient for synthetic data



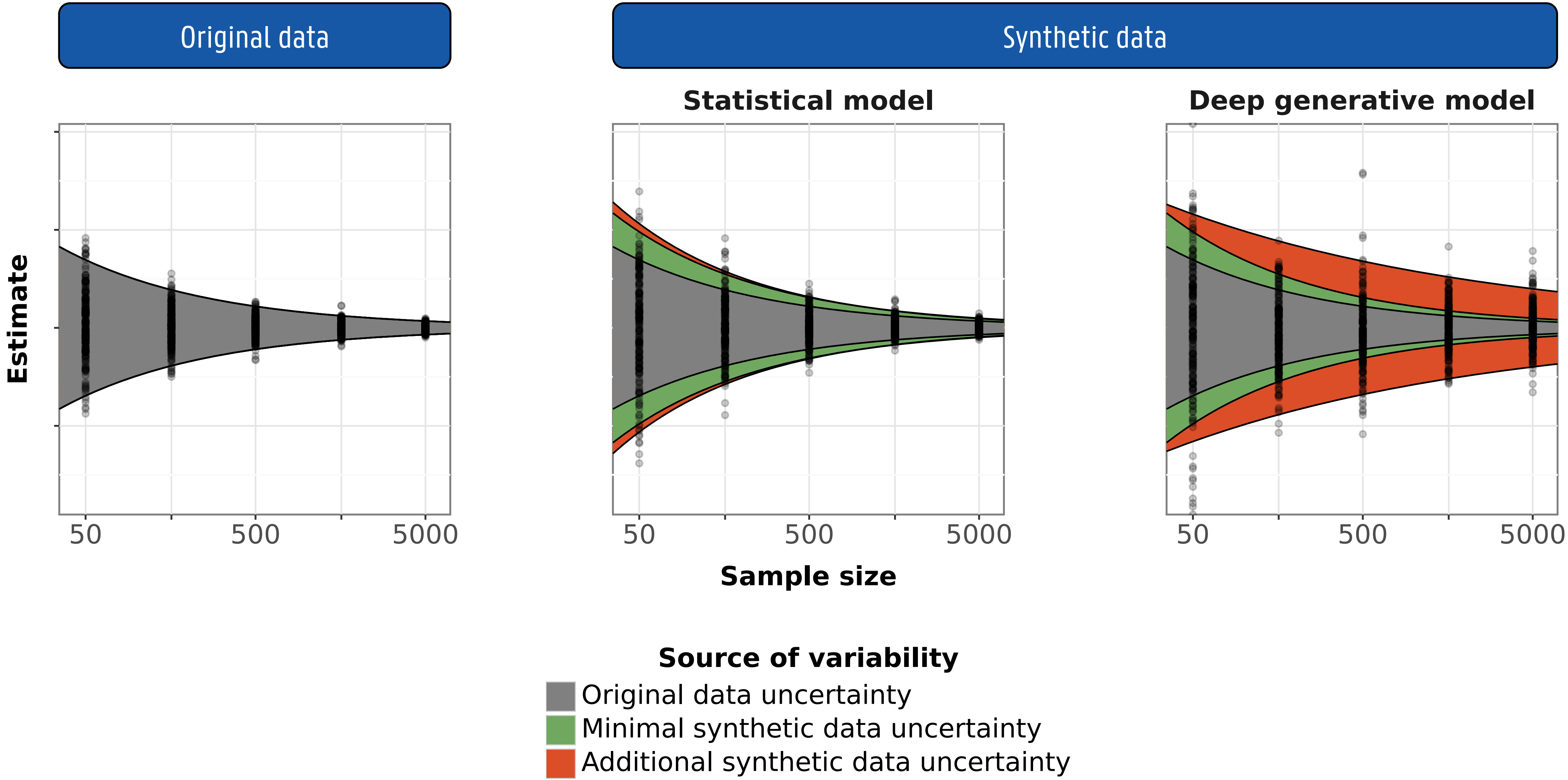
Original standard errors are insufficient for synthetic data



Corrected standard errors are sufficient only for statistical models



Corrected standard errors are insufficient for deep generative models



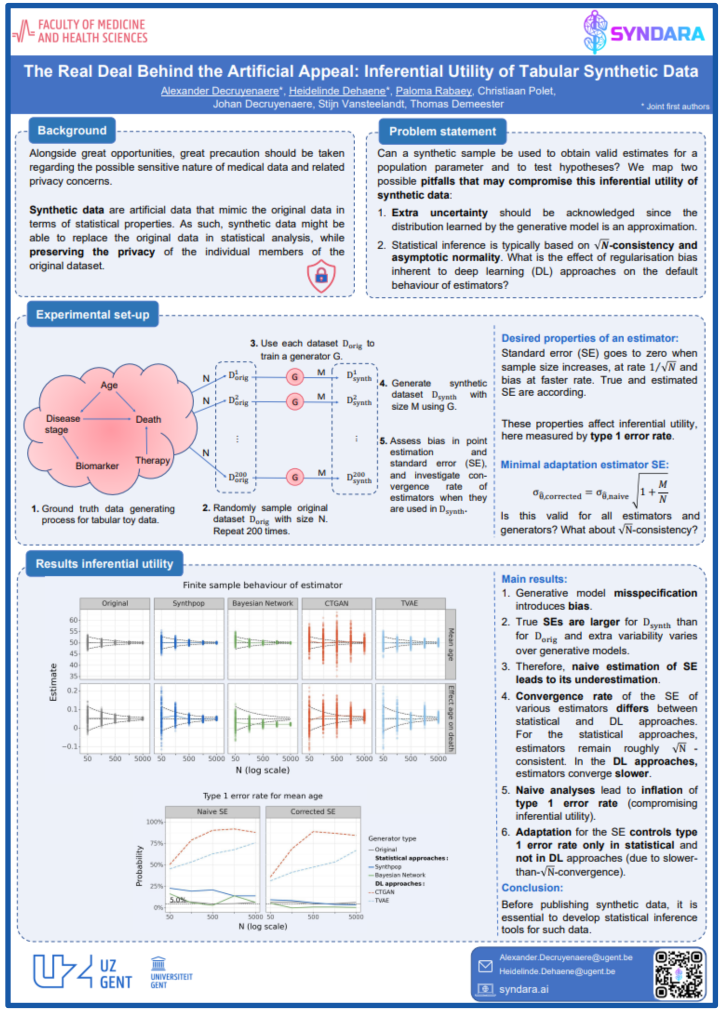
Find out more!

Heidelinde

Paloma

Alexander

Visit our poster (#406)



SYNthetic DAta for Research Acceleration

SYNDARA

GHENT UNIVERSITY

UZ GENT