Balancing the privacy-utility trade-off for synthetic time-to-event data

Generated with sequential regressions in Stata

Sigrid Leithe, Bjørn Møller, Bjarte Aagnes, Yngvar Nilssen, Tor Åge Myklebust

Cancer Registry of Norway

NIPH
Norwegian Institute of Public Health

Synthetic data

Artificially generated data from a model that is trained to reproduce characteristics of the original data.

European Data Protection Supervisor (EDPS)

Cancer Registry of Norway

Public release of example data.

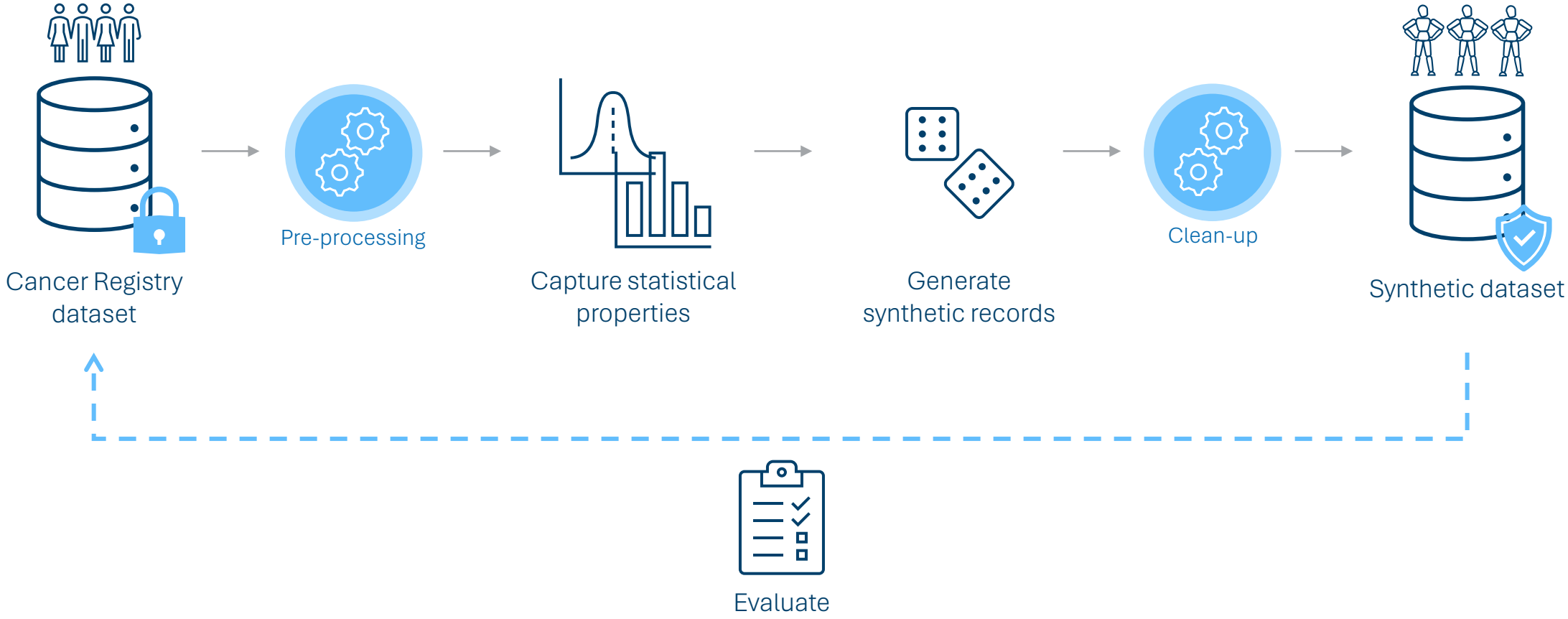
Publish data alongside journal articles to enable reproducibility.

IT development and testing without exposing sensitive information.

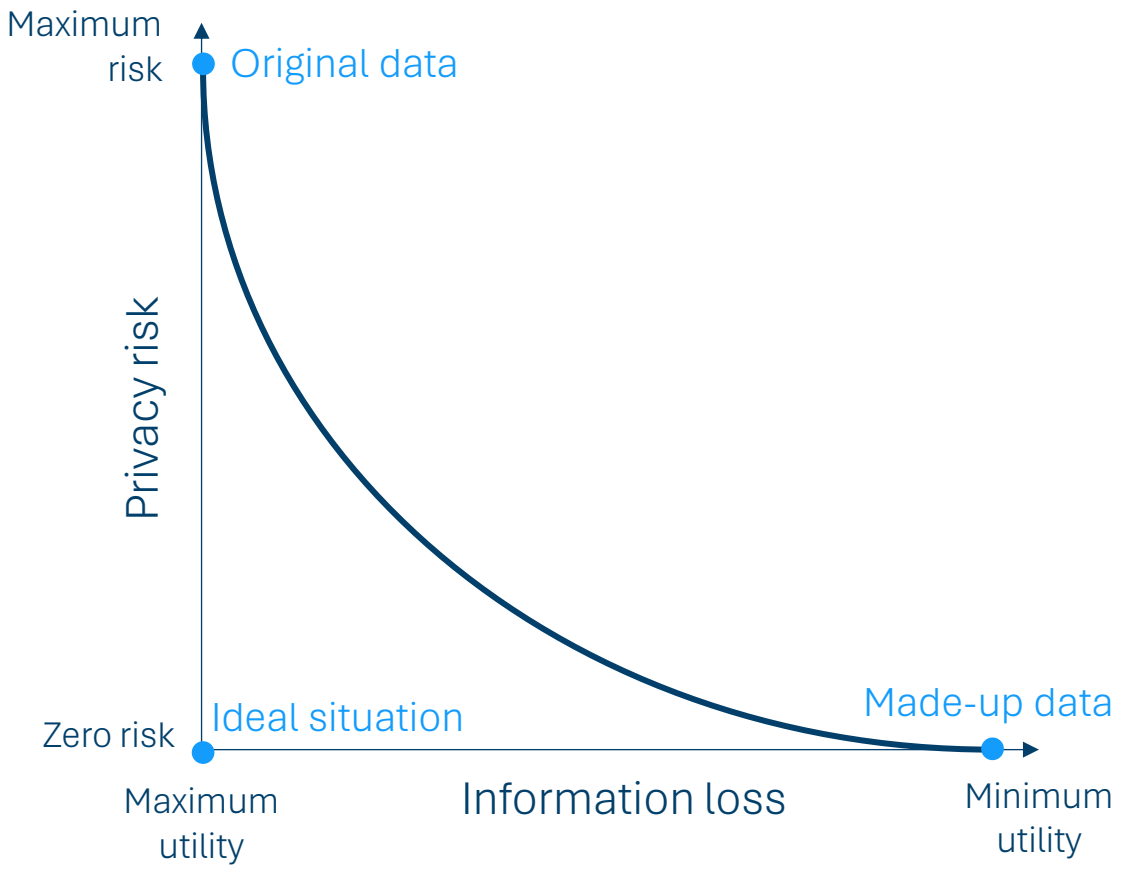
Education and training.

Methods and algorithm development.

Synthetic data generation



Privacy-utility trade-off



Synthetic data generators

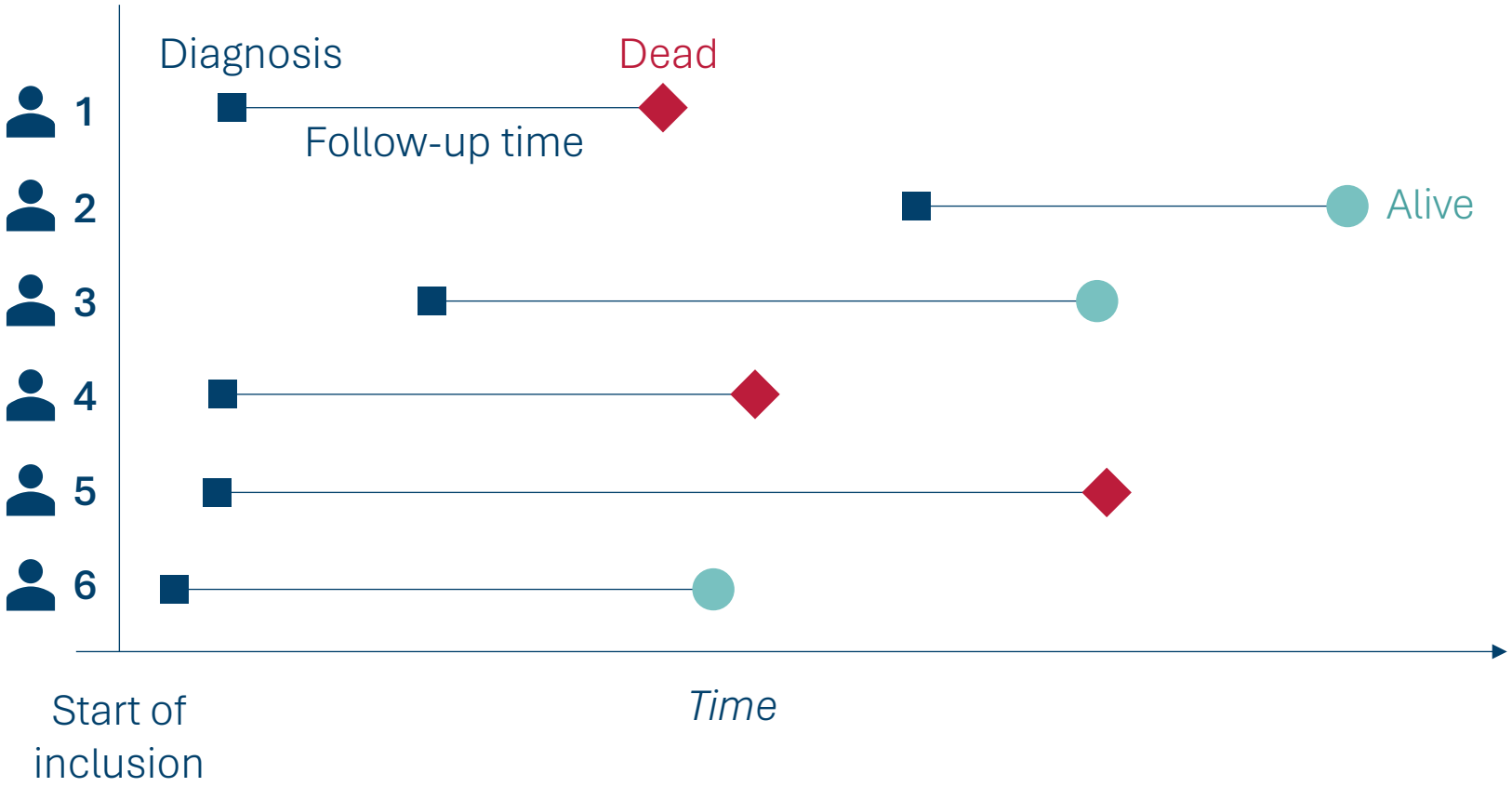
Statistical methods

- Imputation-based methods
- Bayesian networks
- Copula-based methods

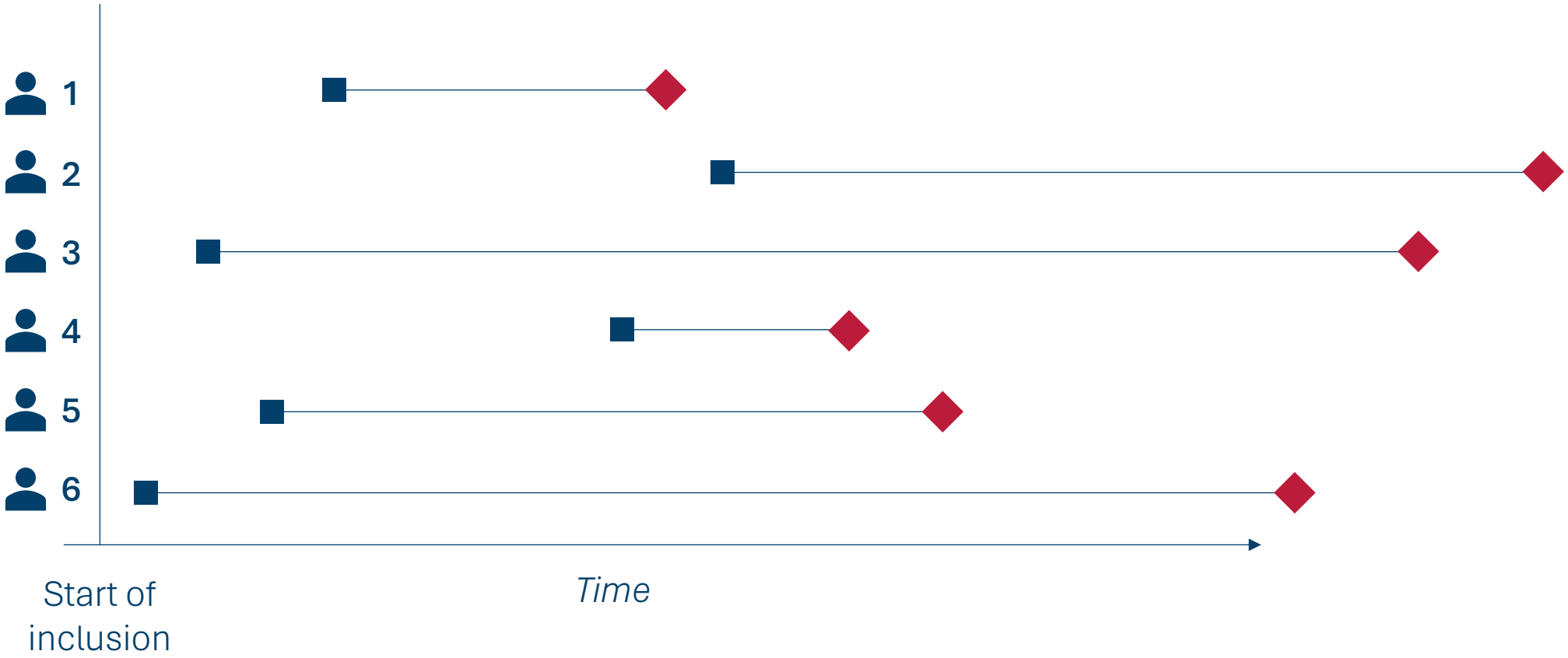
Machine learning methods

- Generative Adversarial Networks (GAN)
- Variational Auto Encoders (VAE)
- Transformer-based models

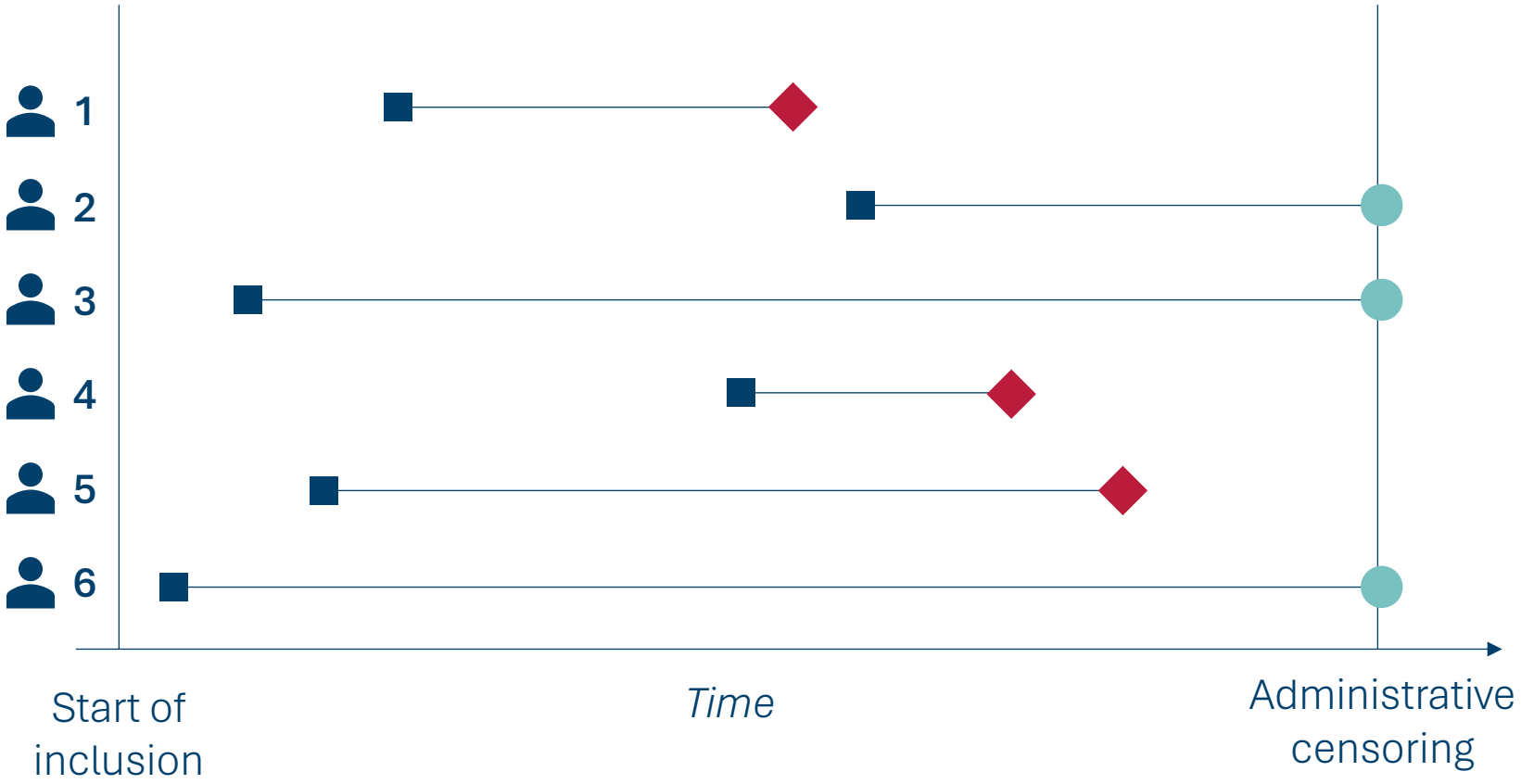
Time-to-event data



Time-to-event data



Time-to-event data



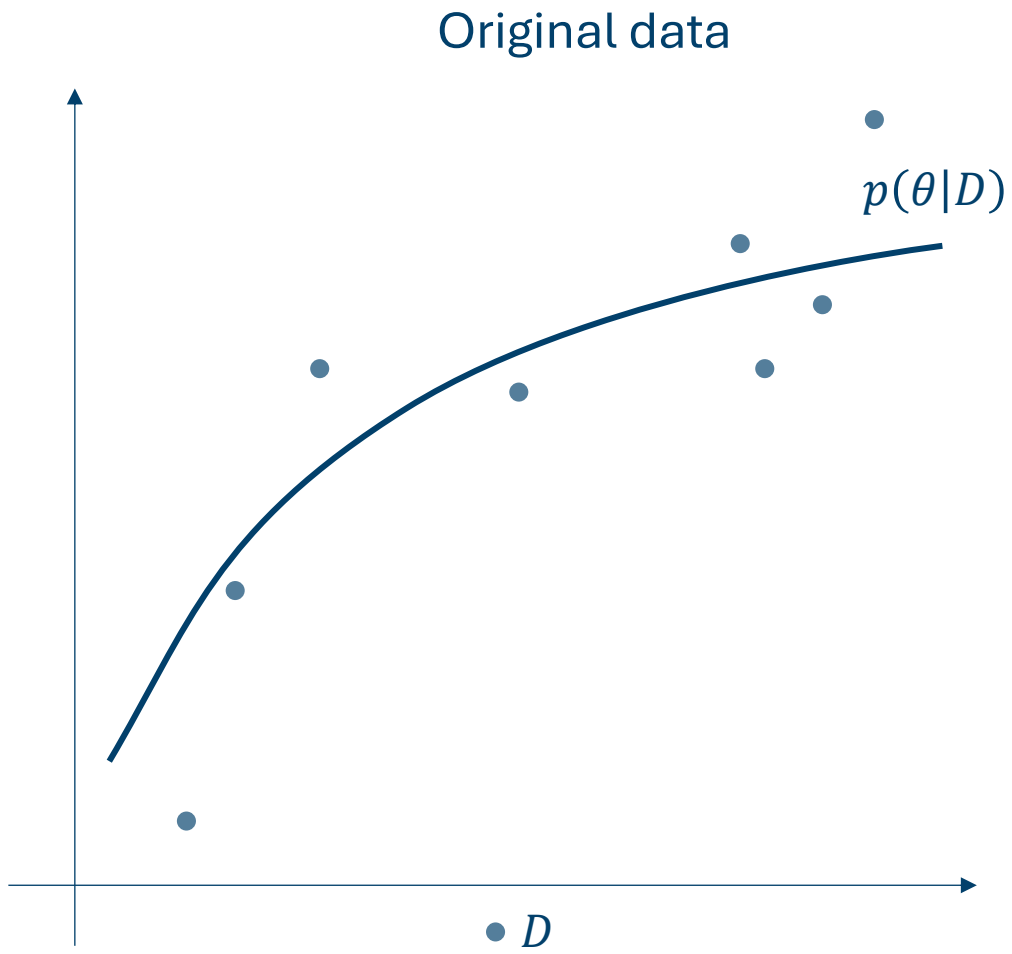
Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Original data: Colon cancer
50 135 patients
Diagnosed 2002-2021
Administrative censoring 31.12.2021

Capture statistical properties

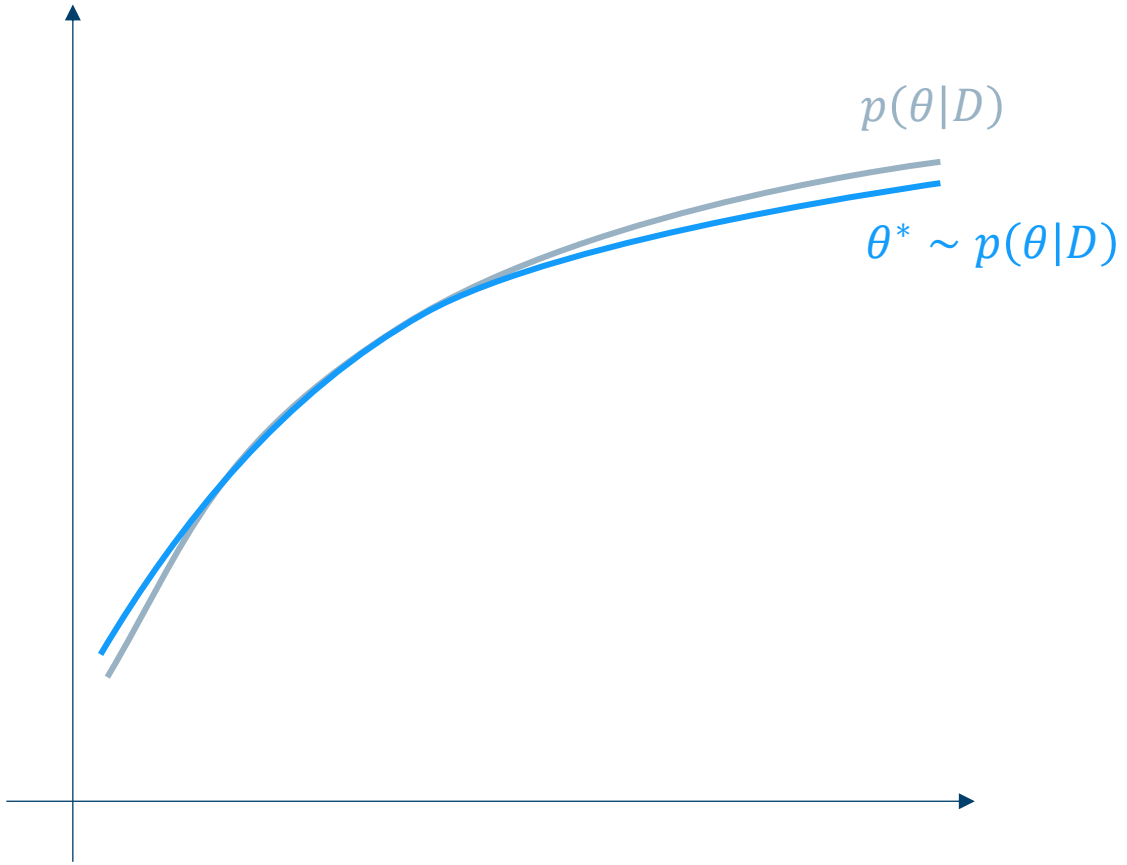
Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead



Capture statistical properties

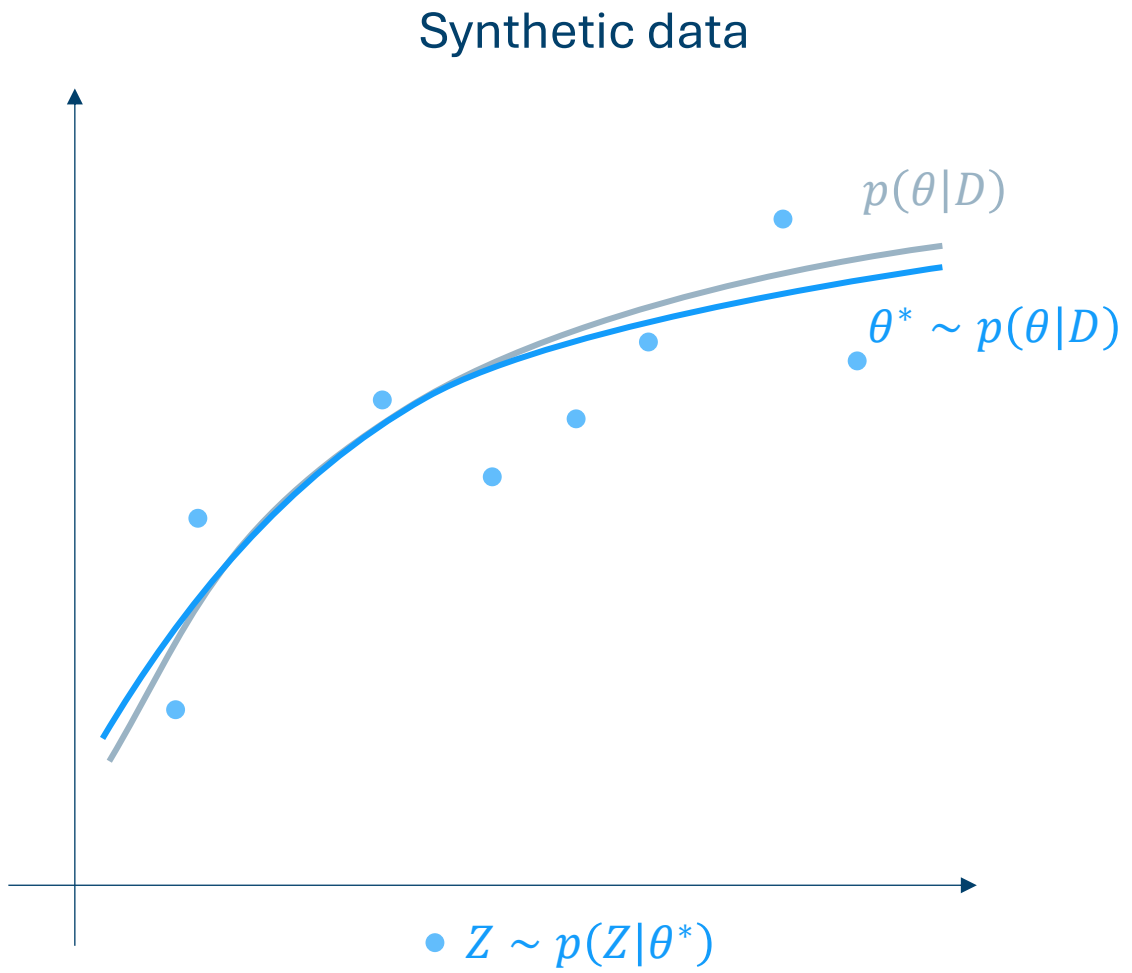
Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Statistical model



Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead



Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

$$\text{Age}_i$$

Synthetic records:

54					

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

Synthetic records:

54					

P(Female | Age = 54) = 0.43
P(Male | Age = 54) = 0.57

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

Synthetic records:

54	Male				

P(Female | Age = 54) = 0.43
P(Male | Age = 54) = 0.57

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

Synthetic records:

54	Male				

P(Local | Age = 54, Male) = 0.45
P(Regional | Age = 54, Male) = 0.25
P(Distant | Age = 54, Male) = 0.19
P(Unknown | Age = 54, Male) = 0.11

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

Synthetic records:

54	Male	Regional			

P(Local | Age = 54, Male) = 0.45
P(Regional | Age = 54, Male) = 0.25
P(Distant | Age = 54, Male) = 0.19
P(Unknown | Age = 54, Male) = 0.11

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

$$\text{Year}_i \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i$$

Synthetic records:

54	Male	Regional			

P(2002 | Age = 54, Male, Regional) = 0.04
P(2003 | Age = 54, Male, Regional) = 0.06
P(2004 | Age = 54, Male, Regional) = 0.05
P(2005 | Age = 54, Male, Regional) = 0.07
…

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

$$\text{Year}_i \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i$$

Synthetic records:

54	Male	Regional	06.03.2005		

P(2002 | Age = 54, Male, Regional) = 0.04
P(2003 | Age = 54, Male, Regional) = 0.06
P(2004 | Age = 54, Male, Regional) = 0.05
P(2005 | Age = 54, Male, Regional) = 0.07
…

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Synthetic records:

54	Male	Regional	06.03.2005		

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

$$\text{Year}_i \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i$$

$$t_i^* \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i, \text{Year}_i$$

$$t_i = \min(t_i^*, C_i)$$

$$\text{Status}_i = \begin{cases} \text{Dead}, & t_i^* \leq C_i \\ \text{Alive}, & t_i^* > C_i \end{cases}$$

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Capture statistical properties

Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
77	Female	Distant	05.03.2002	5.12	Dead
50	Male	Local	21.08.2011	9.51	Alive
63	Female	Unknown	30.12.2018	3.02	Alive
46	Male	Regional	02.07.2005	3.45	Dead
60	Male	Local	09.10.2016	13.83	Dead

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

$$\text{Year}_i \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i$$

$$t_i^* \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i, \text{Year}_i$$

Synthetic records:

54	Male	Regional	06.03.2005	4.31	Dead

$$t_i = \min(t_i^*, C_i)$$

$$\text{Status}_i = \begin{cases} \text{Dead}, & t_i^* \leq C_i \\ \text{Alive}, & t_i^* > C_i \end{cases}$$

Smith, A., Lambert, P. C., & Rutherford, M. J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol*.

Experimental design

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

$$\text{Year}_i \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i$$

$$t_i^* \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i, \text{Year}_i$$

In all models: Main effects of Age, stage, sex and year(2-year periods).

Model #	Interactions	Time varying coefficients (TVCs)	Degrees of freedom: Baseline hazard	Degrees of freedom: Age	Degrees of freedom: TVCs
1	None	None	4	5	-
2	Stage × Age	Age, stage	4	5	2
3	Stage x Age, Stage × Sex, Age × Sex	Age, stage, sex	4	5	3
4	Stage × Age × Sex	Age, stage, sex	4	5	3
5	Stage × Age × Sex	Age, stage, sex	6	8	6

Experimental design

Sequential regressions:

$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

$$\text{Year}_i \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i$$

$$t_i^* \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i, \text{Year}_i$$

Model #	Interactions	Time varying coefficients (TVCs)	Degrees of freedom: Age	Degrees of freedom: Baseline hazard	Degrees of freedom: TVCs
Independent marginals (lower reference model)					
1	None	None	4	5	-
2	Stage × Age	Age, stage	4	5	2
3	Stage × Age, Stage × Sex, Age × Sex	Age, stage, sex	4	5	3
4	Stage × Age × Sex	Age, stage, sex	4	5	3
5	Stage × Age × Sex	Age, stage, sex	6	8	6
Resampling (upper reference model)					

50 datasets from each model

Synthetic data evaluation

Utility

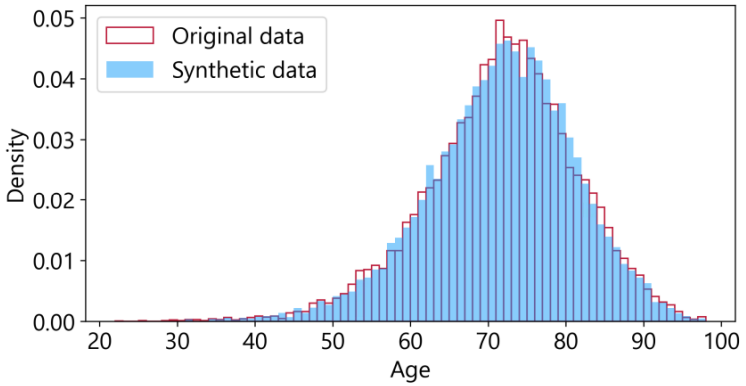
Privacy

Similarity

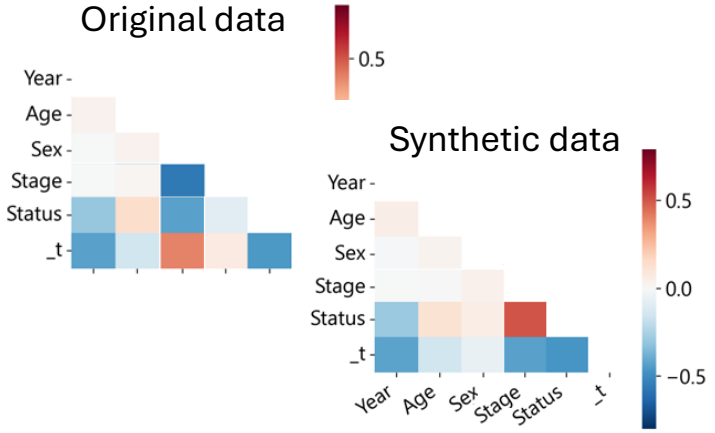
Specific utility

Synthetic data utility

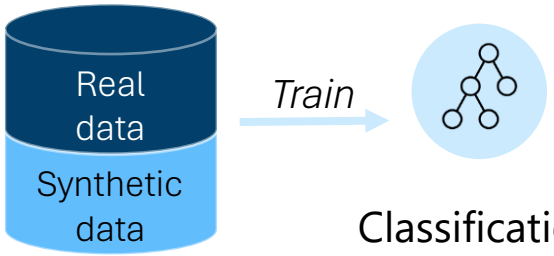
Univariate



Bivariate

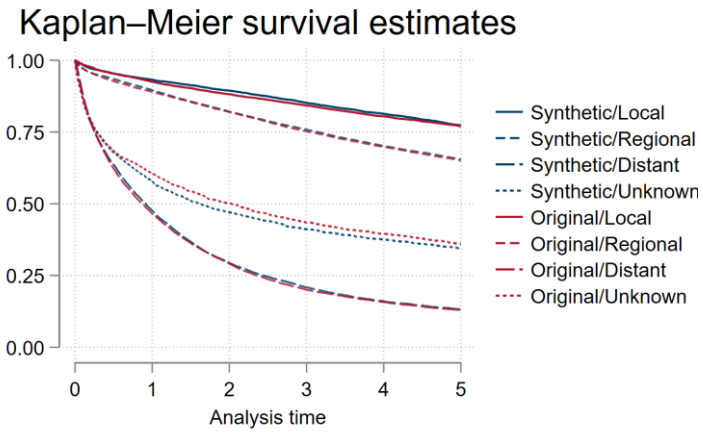


Multivariate

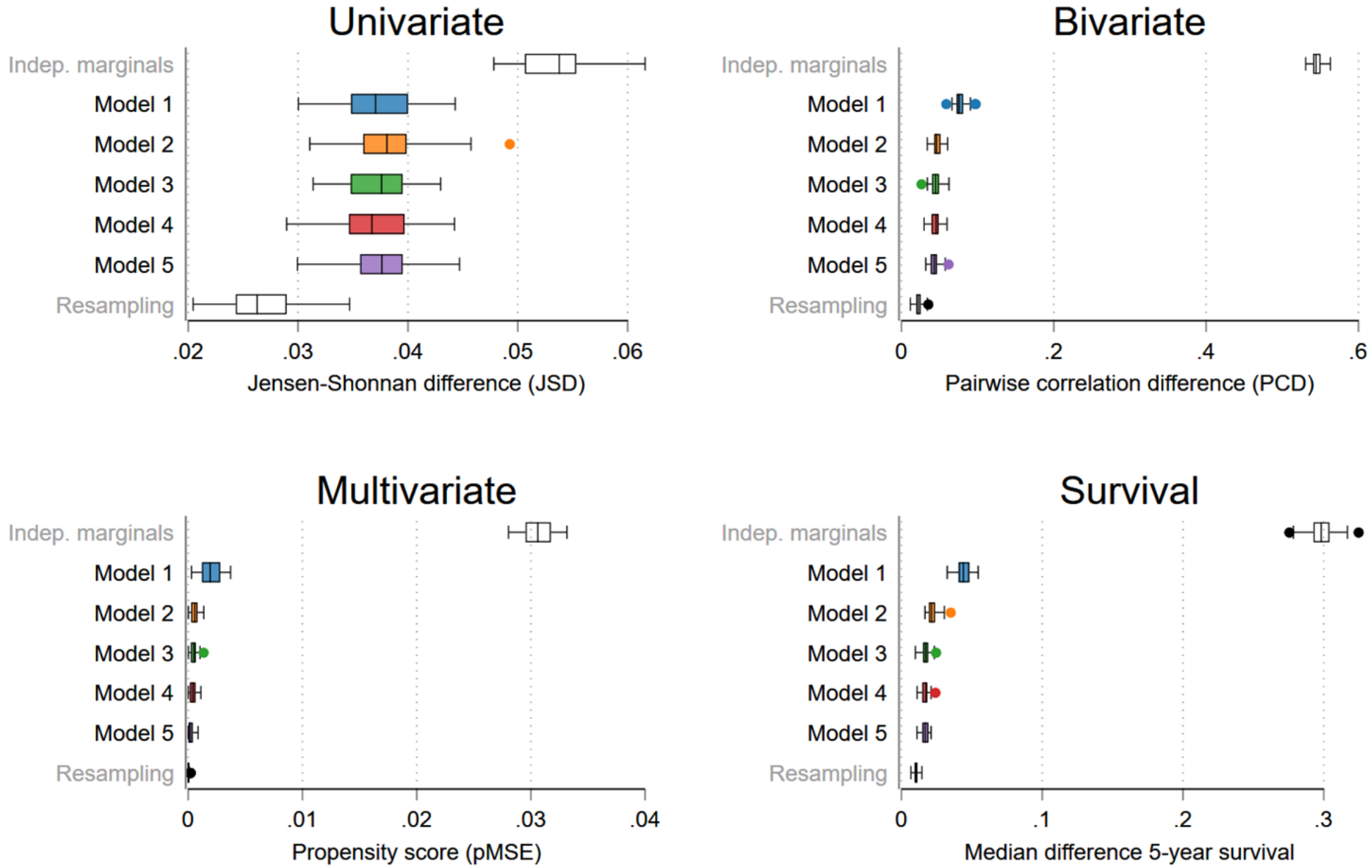


Classification
real/synthetic

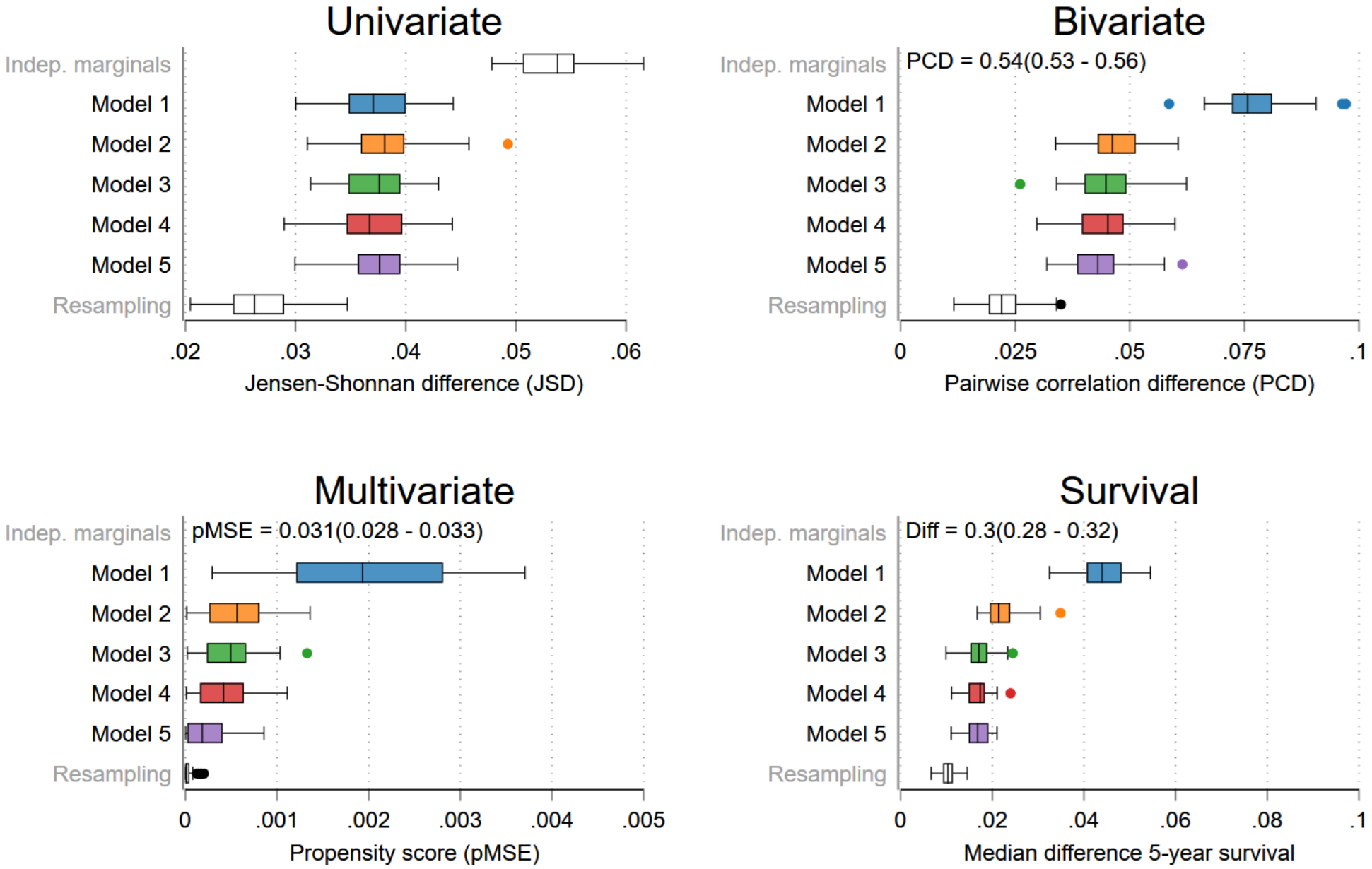
Survival



Synthetic data utility



Synthetic data utility



Synthetic data evaluation

Utility

Similarity

Specific utility

Privacy

Inference attacks

$p(x)$

Probabilistic definitions

Probabilistic privacy evaluation

Toy example

Original data

D_1

ID	v_1	v_2
1	1	1
2	1	0
3	0	0
4	0	1
5	1	0

$p(v_1 = 1) = 60\%$

$p(v_2 = 1|v_1 = 0) = 50\%$

$p(v_2 = 1|v_1 = 1) = 33.3\%$

Possible alternatives

D_2

ID	v_1	v_2
1	1	1
2	1	0
3	0	0
4	0	1
5	1	1

$p(v_1 = 1) = 60\%$

$p(v_2 = 1|v_1 = 0) = 50\%$

$p(v_2 = 1|v_1 = 1) = 66.7\%$

⋮

D_N

Synthetic

Z

ID	v_1	v_2
1	1	0
2	1	0
3	1	1
4	0	1
5	1	0

Probabilistic privacy evaluation

Toy example

Original data

D_1

ID	v_1	v_2
1	1	1
2	1	0
3	0	0
4	0	1
5	1	0

$p(Z|D_1) = 0.26\ \%$

Possible alternatives

D_2

ID	v_1	v_2
1	1	1
2	1	0
3	0	0
4	0	1
5	1	1

$p(Z|D_2) = 0.06\ \%$

$\vdots$

D_N

Z

Synthetic

ID	v_1	v_2
1	1	0
2	1	0
3	1	1
4	0	1
5	1	0

Probabilistic privacy evaluation

Toy example

Original data

D_1

ID	v_1	v_2
1	1	1
2	1	0
3	0	0
4	0	1
5	1	0

$p(Z|D_1) = 0.26\ \%$

Possible alternatives

D_2

ID	v_1	v_2
1	1	1
2	1	0
3	0	0
4	0	1
5	1	1

$p(Z|D_2) = 0.06\ \%$

$\vdots$

D_N

Z

Synthetic

ID	v_1	v_2
1	1	0
2	1	0
3	1	1
4	0	1
5	1	0

$$\frac{p(Z|D_1)}{p(Z|D_2)} = 4$$

$$\frac{p(D_1|Z)}{p(D_2|Z)} = \frac{p(Z|D_1)}{p(Z|D_2)}\frac{p(D_1)}{p(D_2)}$$

Posterior odds = Bayes Factor × Prior odds

Probabilistic privacy evaluation

Toy example

Original data

D_1

ID	v_1	v_2
1	1	1
2	1	0
3	0	0
4	0	1
5	1	0

$p(Z|D_1) = 0.26\ \%$

Possible alternatives

D_3

ID	v_1	v_2
1	1	1
2	1	0
3	0	0
4	0	0
5	1	0

$p(Z|D_3) = 0$

$\vdots$

D_N

Z

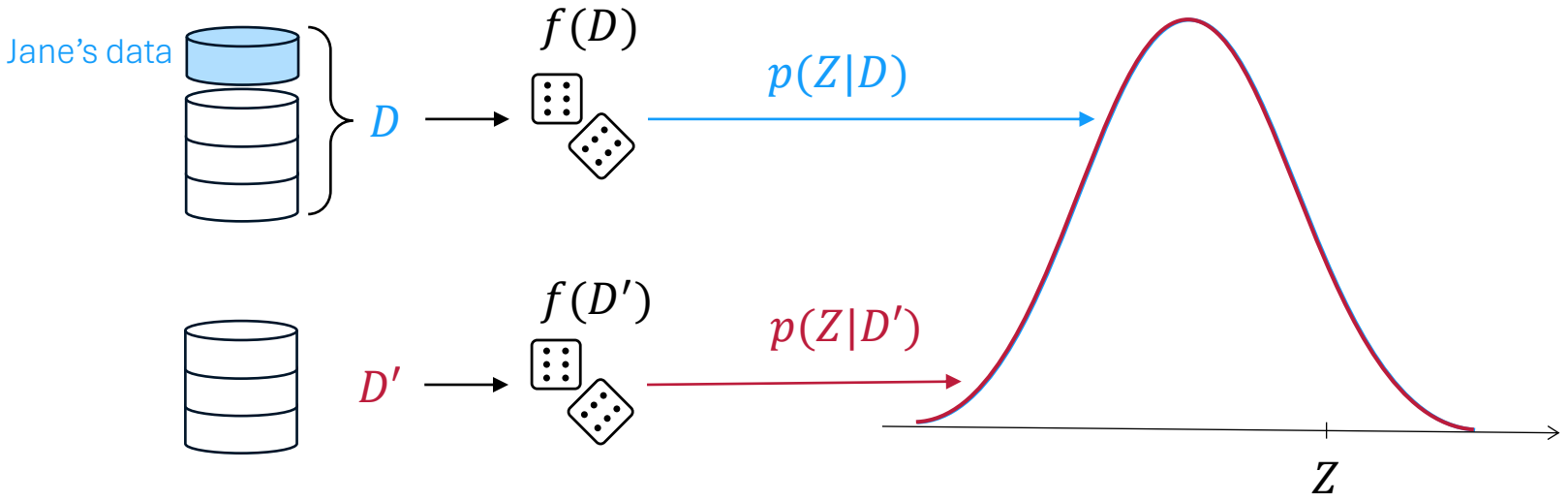
Synthetic

ID	v_1	v_2
1	1	0
2	1	0
3	1	1
4	0	1
5	1	0

$$\frac{p(Z|D_1)}{p(Z|D_3)} = \infty$$

Probabilistic privacy evaluation

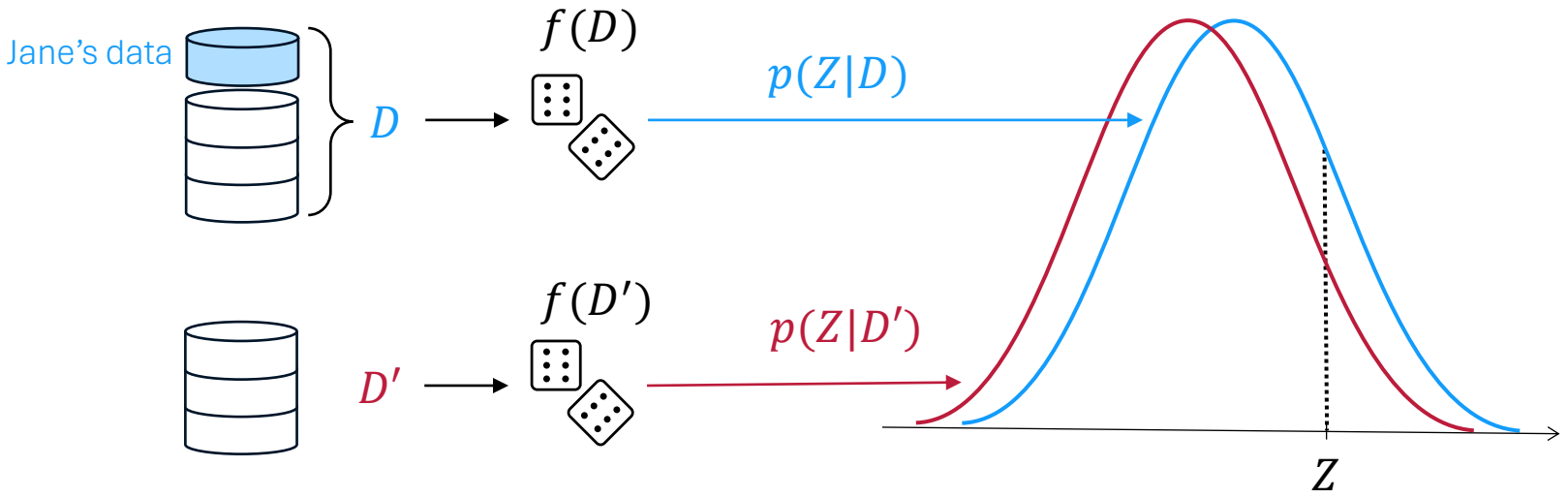
General case



Abowd, J. M., & Vilhuber, L. (2008). How protective are synthetic data? *International Conference on Privacy in Statistical Databases*

Probabilistic privacy evaluation

General case



Information leakage:

$$\frac{p(Z|D)}{p(Z|D')}$$

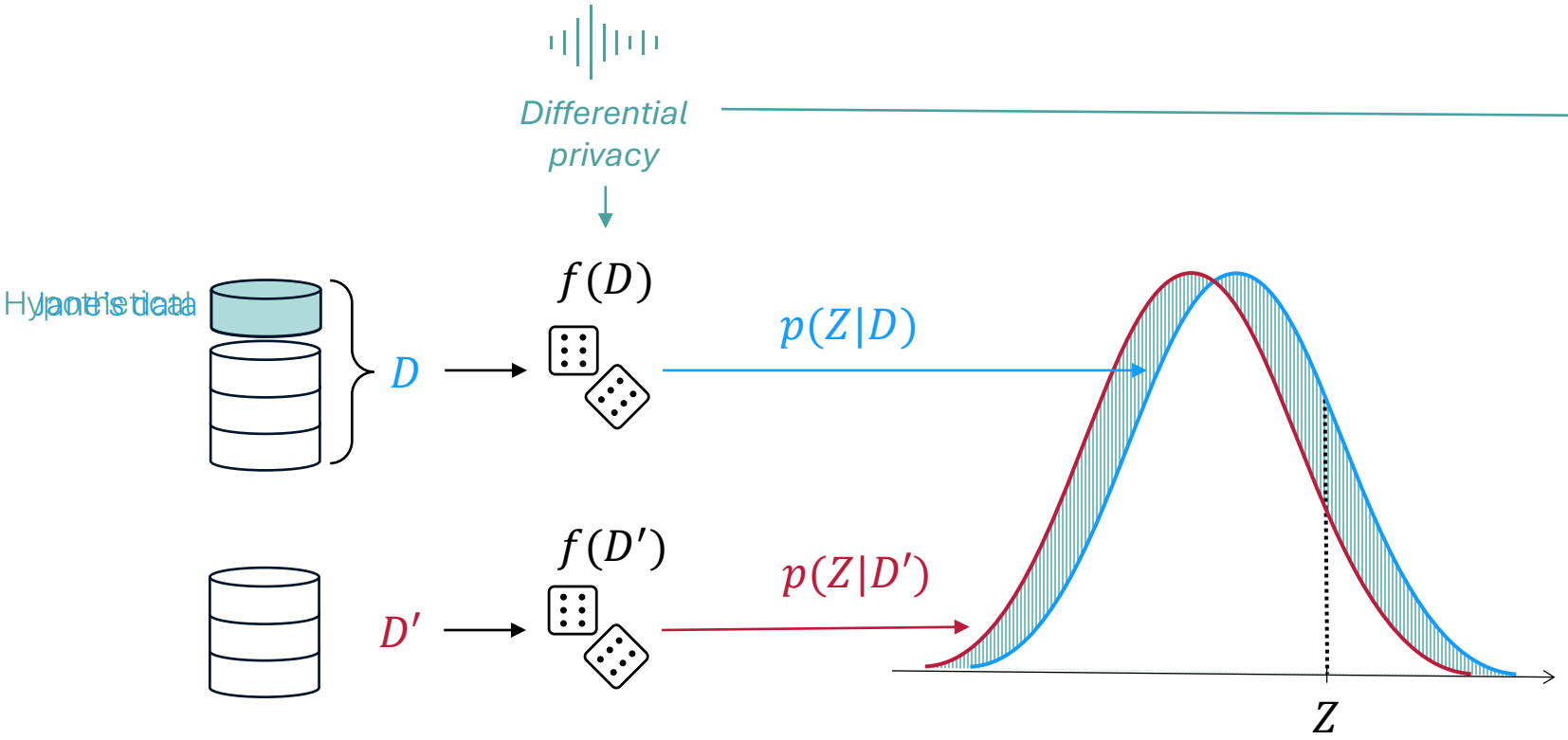
Privacy metric:

$$c(Z|D) = \max_{D'} \left| \log \left(\frac{p(Z|D)}{p(Z|D')} \right) \right|$$

Abowd, J. M., & Vilhuber, L. (2008). How protective are synthetic data? *International Conference on Privacy in Statistical Databases*

Probabilistic privacy evaluation

Relation to differential privacy



$$\forall\, Z \in \text{Range}(f):\ \log\left(\frac{p(Z|D)}{p(Z|D')}\right) \leq \varepsilon$$

Information leakage:

$$\frac{p(Z|D)}{p(Z|D')}$$

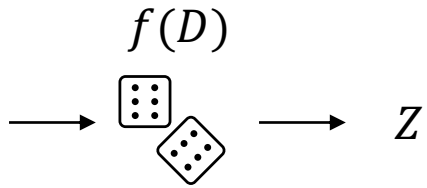
Privacy metric:

$$c(Z|D) = \max_{D'} \left|\log\left(\frac{p(Z|D)}{p(Z|D')}\right)\right|$$

Abowd, J. M., & Vilhuber, L. (2008). How protective are synthetic data? *International Conference on Privacy in Statistical Databases*

Probabilistic privacy evaluation

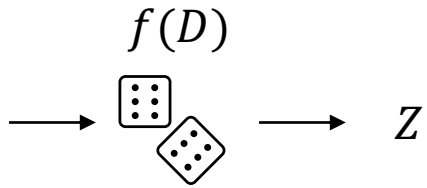
	i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
Jane	1	77	Female	Distant	05.03.2002	5.12	Dead
	2	50	Male	Local	21.08.2011	9.51	Alive
	3	63	Female	Unknown	30.12.2018	3.02	Alive
	4	46	Male	Regional	02.07.2005	3.45	Dead
	5	60	Male	Local	09.10.2016	13.83	Dead



Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

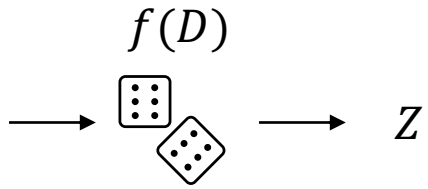
	i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
Jane	1	77	Female	**?**	05.03.2002	5.12	Dead
	2	50	Male	Local	21.08.2011	9.51	Alive
	3	63	Female	Unknown	30.12.2018	3.02	Alive
	4	46	Male	Regional	02.07.2005	3.45	Dead
	5	60	Male	Local	09.10.2016	13.83	Dead



Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

	i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
Jane	1	77	Female	**?**	05.03.2002	5.12	Dead
	2	50	Male	Local	21.08.2011	9.51	Alive
	3	63	Female	Unknown	30.12.2018	3.02	Alive
	4	46	Male	Regional	02.07.2005	3.45	Dead
	5	60	Male	Local	09.10.2016	13.83	Dead



Estimate:

$p(Z|D, \text{stage}_1 = \text{Localised})$

$p(Z|D, \text{stage}_1 = \text{Regional})$

$p(Z|D)$ *(stage$_1$ = Distant)*

$p(Z|D, \text{stage}_1 = \text{Unknown})$

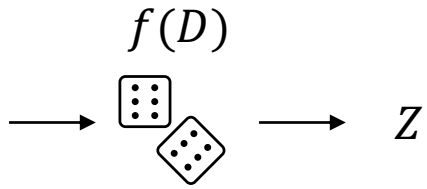
Information leakage:

$$\max_y \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_1 = y)} \right) \right|$$

Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
1	77	Female	Distant	05.03.2002	5.12	Dead
2	50	Male	**?**	21.08.2011	9.51	Alive
3	63	Female	Unknown	30.12.2018	3.02	Alive
4	46	Male	Regional	02.07.2005	3.45	Dead
5	60	Male	Local	09.10.2016	13.83	Dead



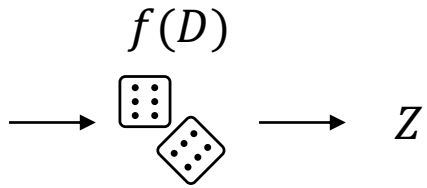
Information leakage:

$$\max_{y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_2 = y)} \right) \right|$$

Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
1	77	Female	Distant	05.03.2002	5.12	Dead
2	50	Male	Local	21.08.2011	9.51	Alive
3	63	Female	**?**	30.12.2018	3.02	Alive
4	46	Male	Regional	02.07.2005	3.45	Dead
5	60	Male	Local	09.10.2016	13.83	Dead



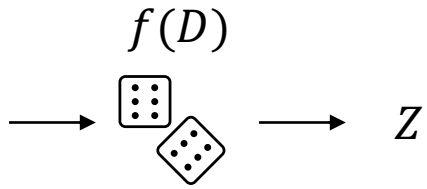
Information leakage:

$$\max_{y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_3 = y)} \right) \right|$$

Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
1	77	Female	Distant	05.03.2002	5.12	Dead
2	50	Male	Local	21.08.2011	9.51	Alive
3	63	Female	Unknown	30.12.2018	3.02	Alive
4	46	Male	**?**	02.07.2005	3.45	Dead
5	60	Male	Local	09.10.2016	13.83	Dead



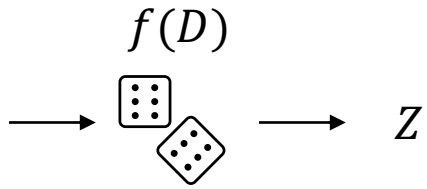
Information leakage:

$$\max_{y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_4 = y)} \right) \right|$$

Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
1	77	Female	Distant	05.03.2002	5.12	Dead
2	50	Male	Local	21.08.2011	9.51	Alive
3	63	Female	Unknown	30.12.2018	3.02	Alive
4	46	Male	Regional	02.07.2005	3.45	Dead
5	60	Male	?	09.10.2016	13.83	Dead



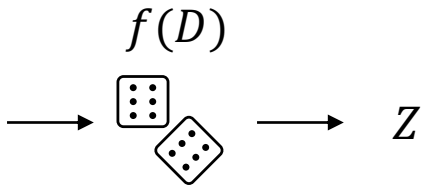
Information leakage:

$$\max_{y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_5 = y)} \right) \right|$$

Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
1	77	Female	Distant	05.03.2002	5.12	Dead
2	50	Male	Local	21.08.2011	9.51	Alive
3	63	Female	Unknown	30.12.2018	3.02	Alive
4	46	Male	Regional	02.07.2005	3.45	Dead
5	60	Male	Local	09.10.2016	13.83	Dead



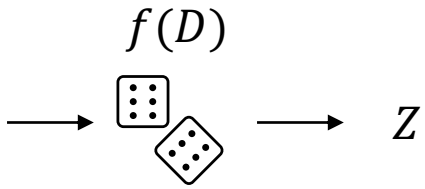
Privacy metric:

$$c(Z|D) = \max_{i,y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_i = y)} \right) \right|$$

Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
1	77	Female	Distant	05.03.2002	5.12	Dead
2	50	Male	Local	21.08.2011	9.51	Alive
3	63	Female	Unknown	30.12.2018	3.02	Alive
4	46	Male	Regional	02.07.2005	3.45	Dead
5	60	Male	Local	09.10.2016	13.83	Dead
⋮						



50 135 × 3 = 150 405

Privacy metric:

$$c(Z|D) = \max_{i,y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_i = y)} \right) \right|$$

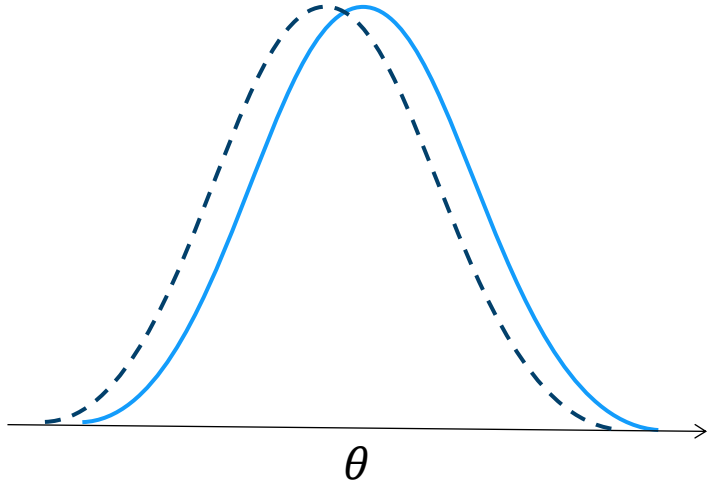
Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Probabilistic privacy evaluation

$$c(Z|D) = \max_{i,y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_i = y)} \right) \right|$$

$$D_{\text{stage}_i = y} \approx D$$

$$p\left(\theta \middle| D_{\text{stage}_i = y}\right) \approx p(\theta|D)$$



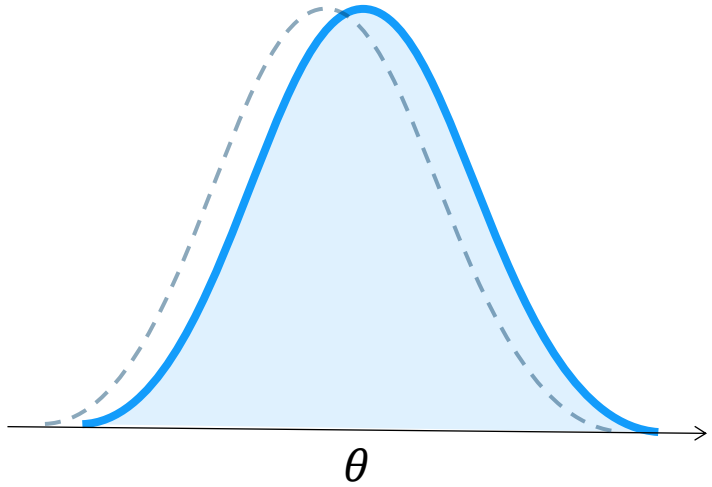
Hu, J., Reiter, J. P., & Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. *International conference on privacy in statistical databases*.

Probabilistic privacy evaluation

$$c(Z|D) = \max_{i,y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_i = y)} \right) \right|$$

$$D_{\text{stage}_i = y} \approx D$$

$$p\left(\theta \middle| D_{\text{stage}_i = y}\right) \approx p(\theta|D)$$



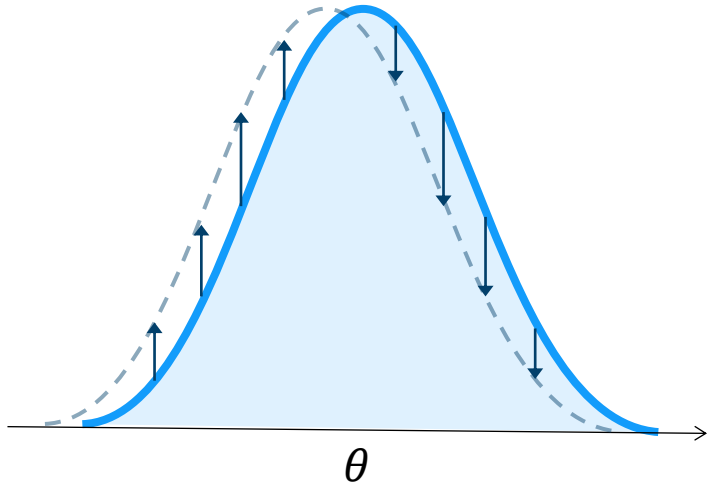
Hu, J., Reiter, J. P., & Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. *International conference on privacy in statistical databases*.

Probabilistic privacy evaluation

$$c(Z|D) = \max_{i,y} \left| \log \left(\frac{p(Z|D)}{p(Z|D, \text{stage}_i = y)} \right) \right|$$

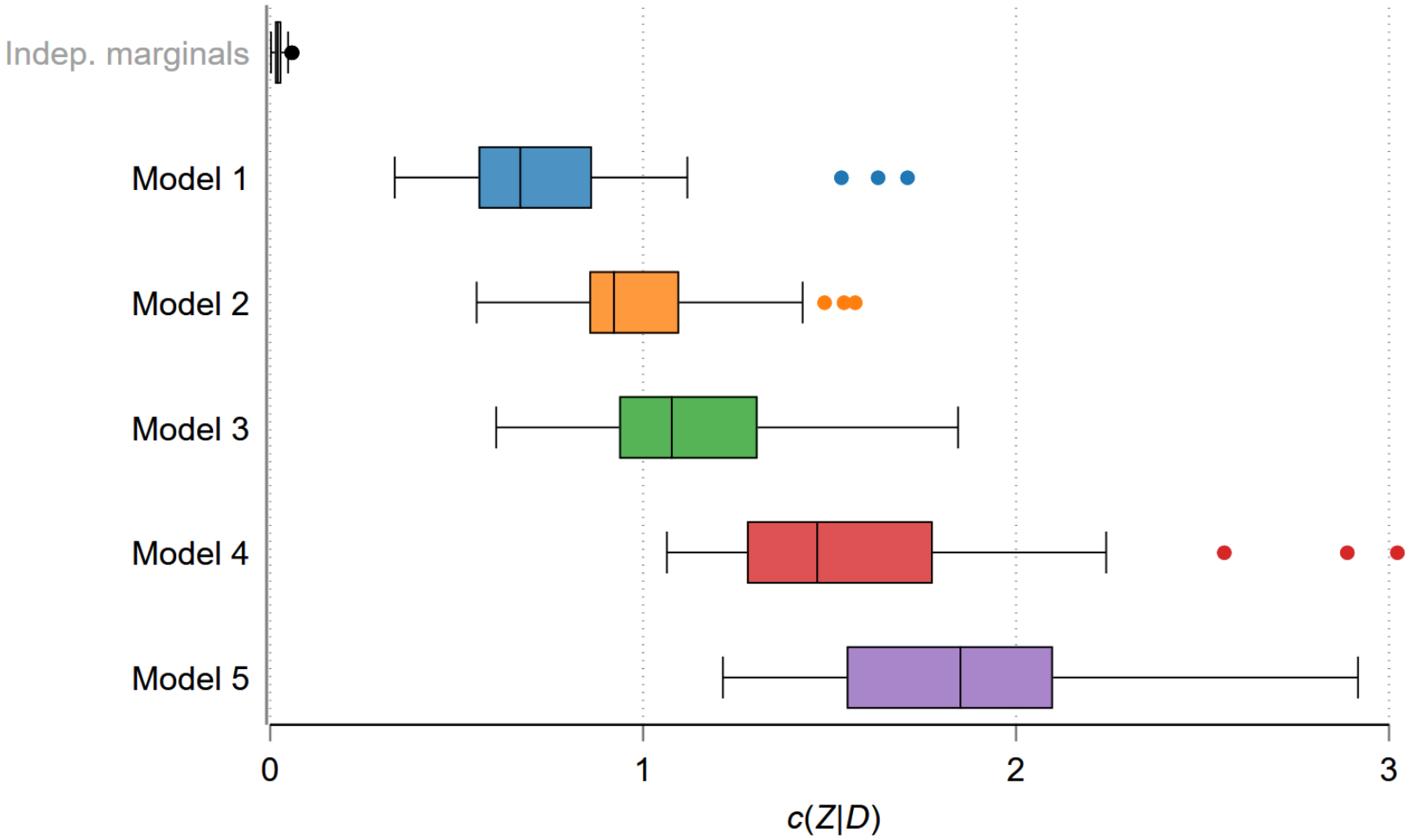
$$D_{\text{stage}_i = y} \approx D$$

$$p\left(\theta \middle| D_{\text{stage}_i = y}\right) \approx p(\theta|D)$$



Hu, J., Reiter, J. P., & Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. *International conference on privacy in statistical databases*.

Probabilistic privacy evaluation



Cancer Registry of Norway

Probabilistic privacy evaluation

$$p(D, \text{stage}_i = y|Z) \propto p(Z|D, \text{stage}_i = y)p(D, \text{stage}_i = y)$$

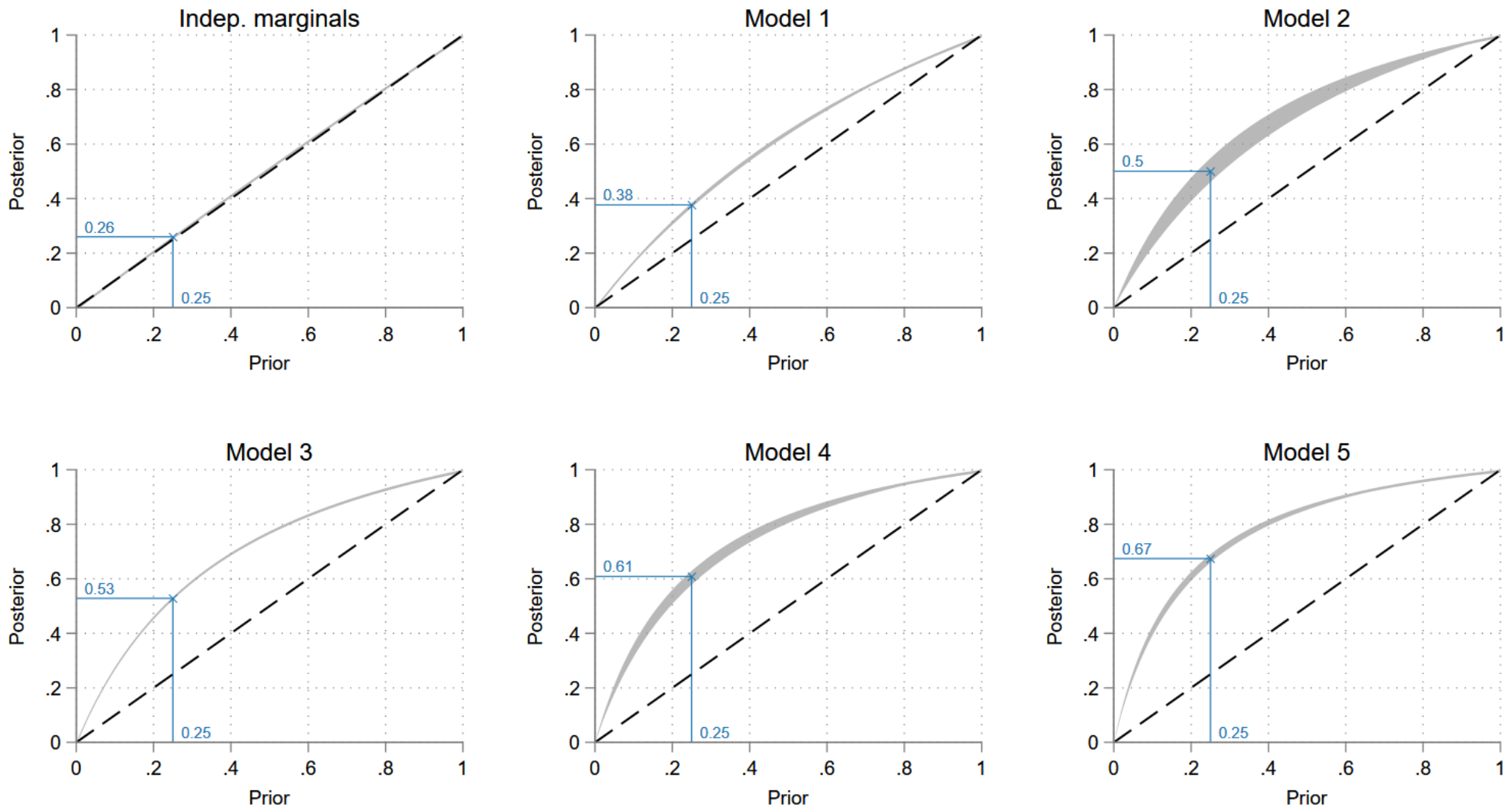
Posterior

Likelihood

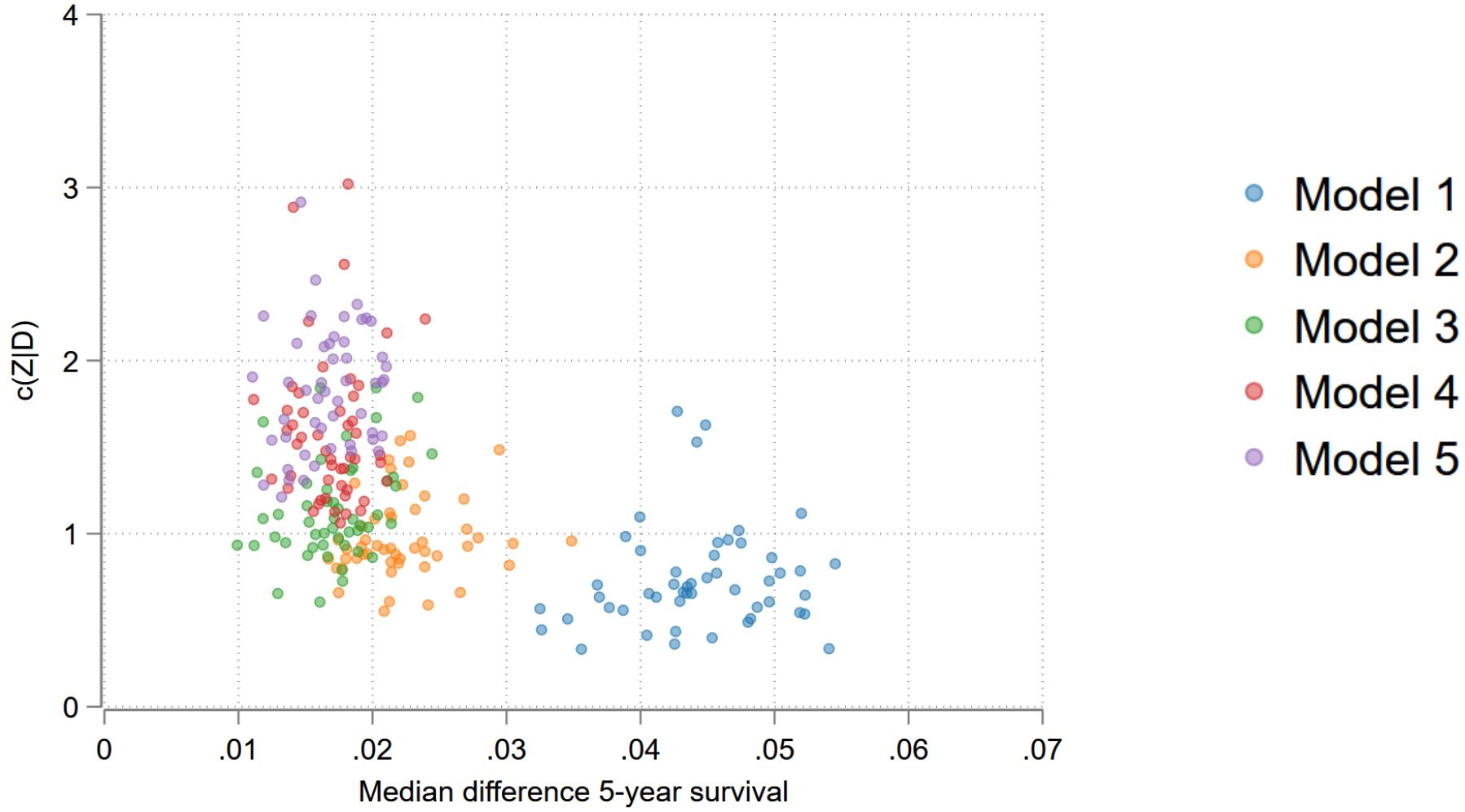
Prior

Cancer Registry of Norway

Probability assigned to correct outcome

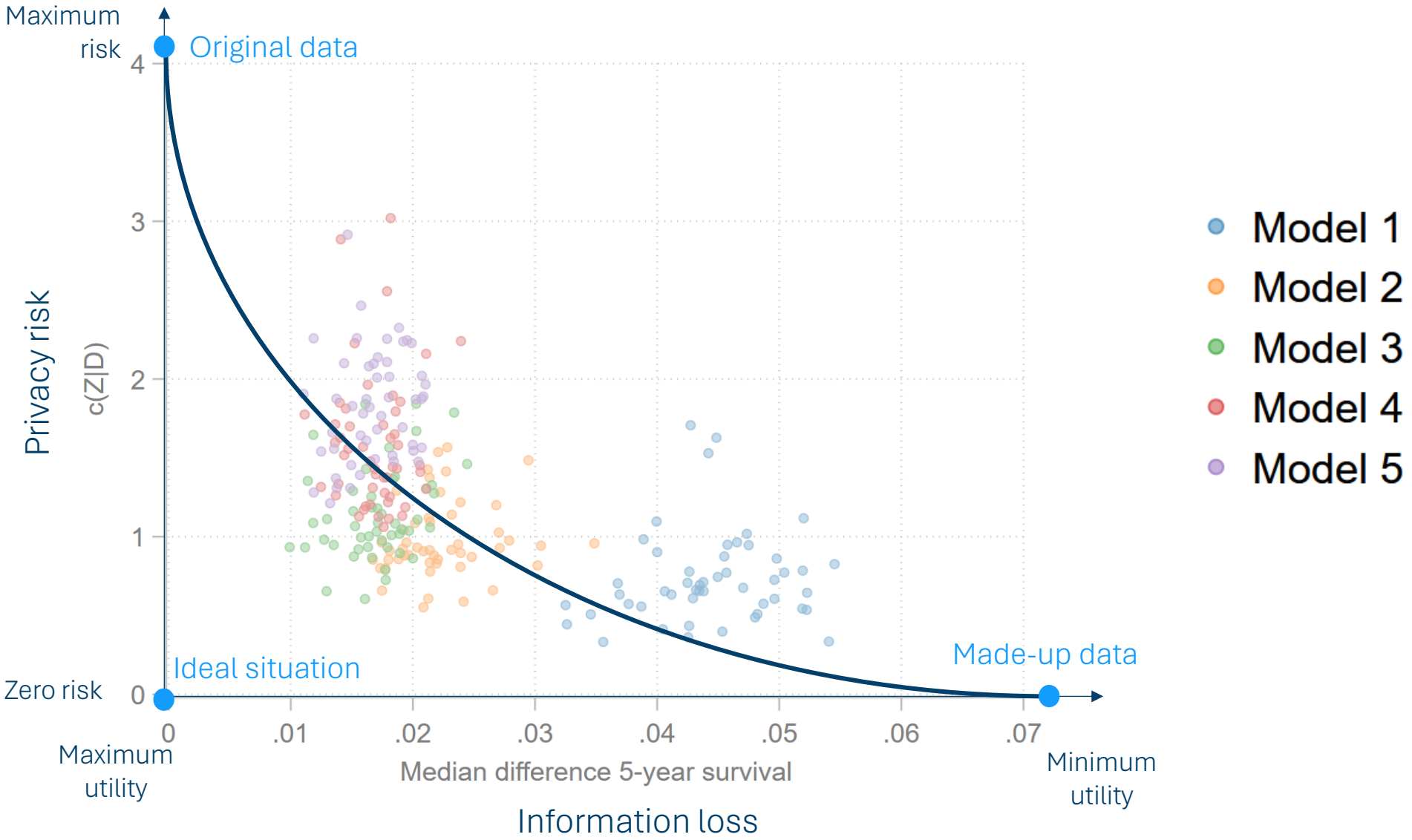


Privacy-utility trade-off



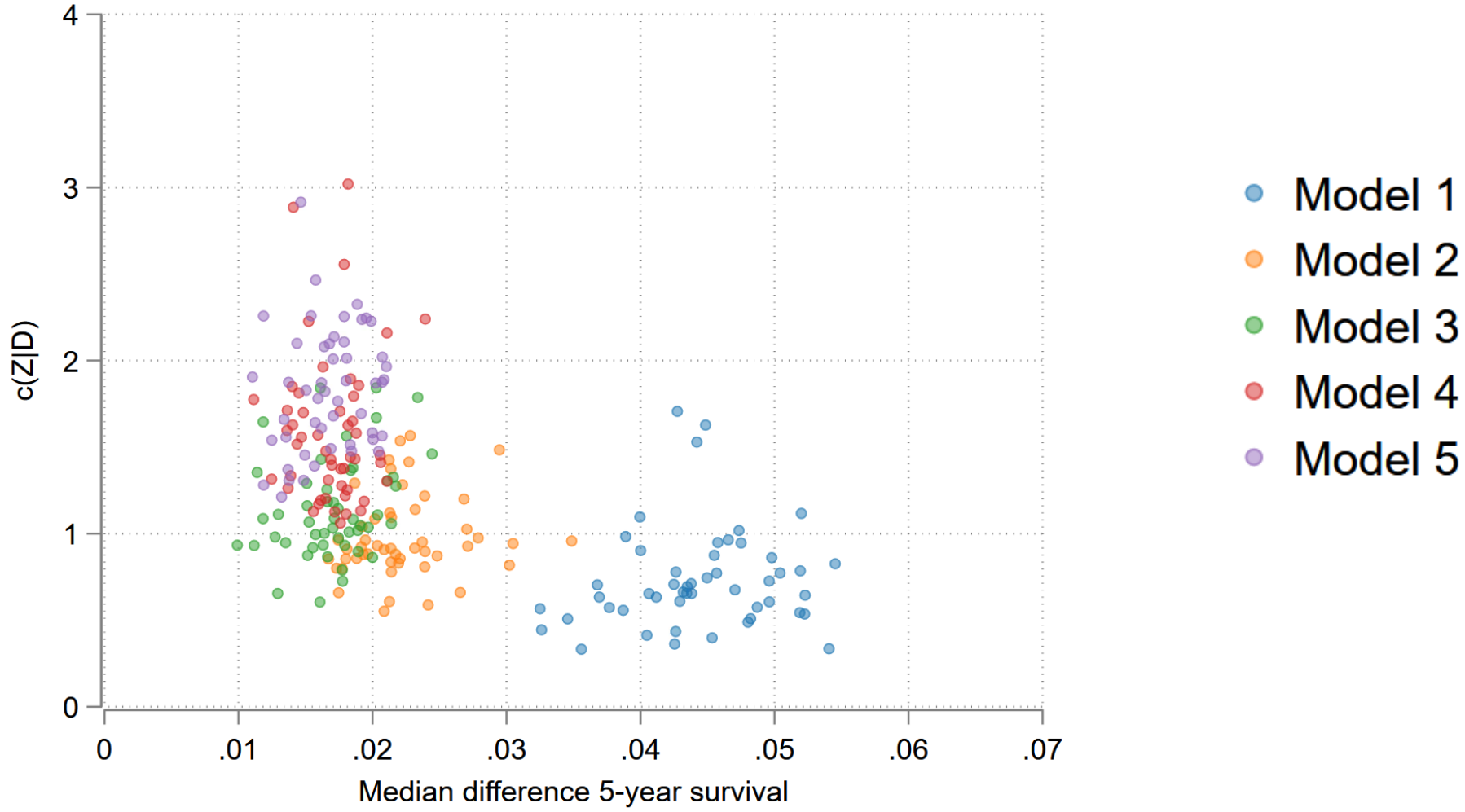
Cancer Registry of Norway

Privacy-utility trade-off



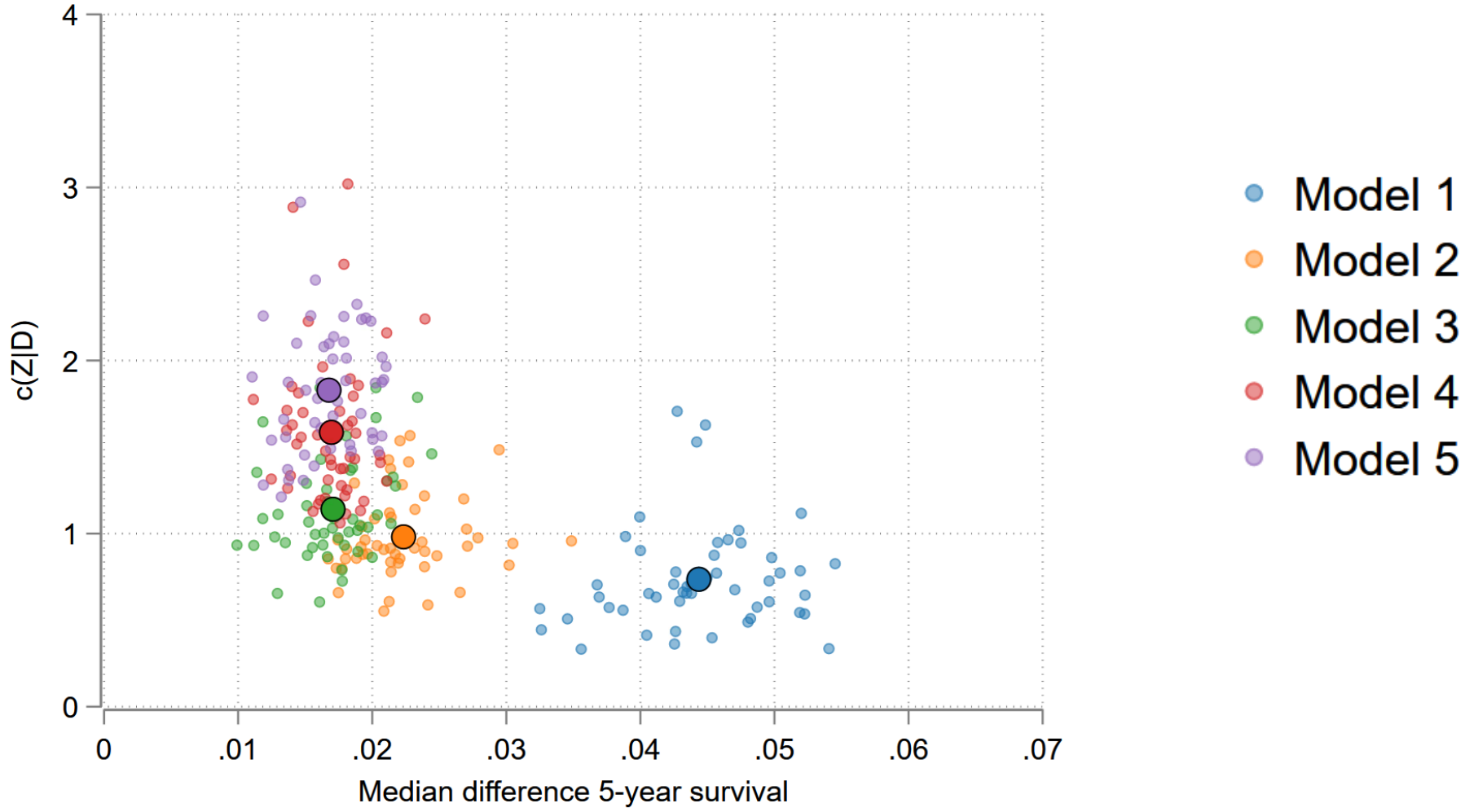
Cancer Registry of Norway

Privacy-utility trade-off



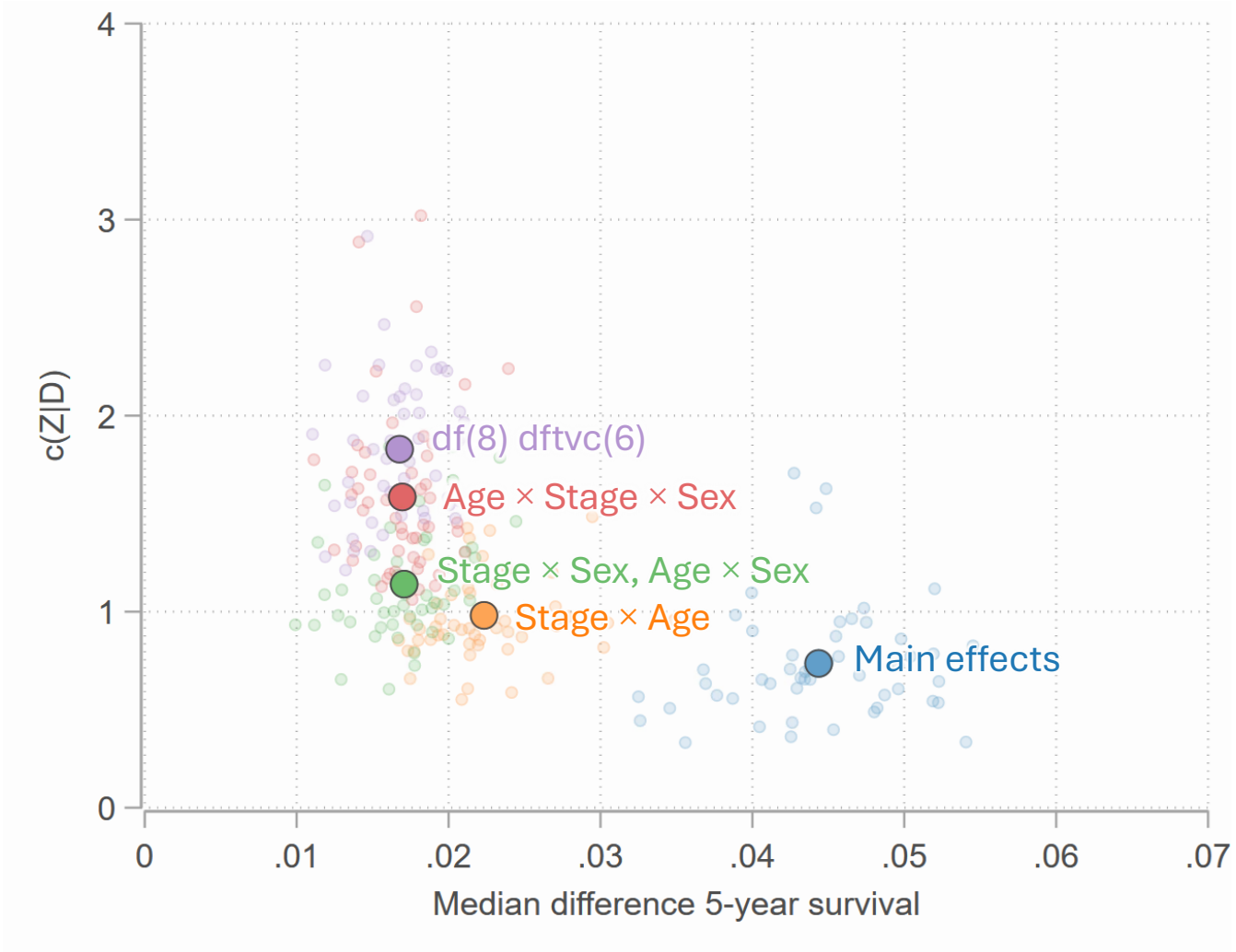
Cancer Registry of Norway

Privacy-utility trade-off



Cancer Registry of Norway

Privacy-utility trade-off



Cancer Registry of Norway

Summary

Sequential regressions:

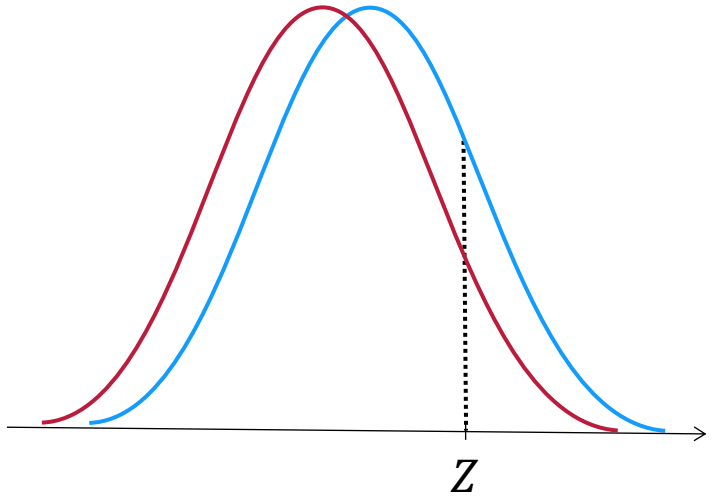
$$\text{Age}_i$$

$$\text{Sex}_i \sim \text{Age}_i$$

$$\text{Stage}_i \sim \text{Age}_i, \text{Sex}_i$$

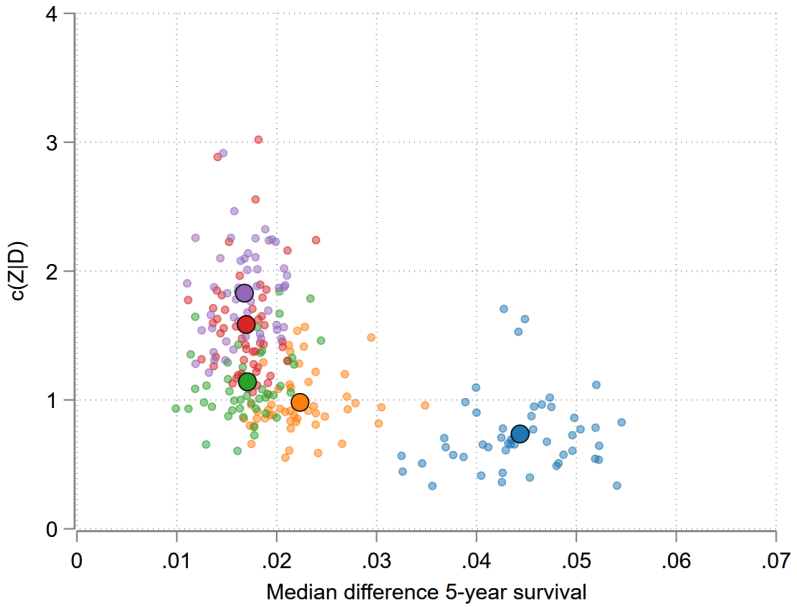
$$\text{Year}_i \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i$$

$$t_i^* \sim \text{Age}_i, \text{Sex}_i, \text{Stage}_i, \text{Year}_i$$



Privacy metric:

$$c(Z|D) = \max_{D'} \left| \log \left(\frac{p(Z|D)}{p(Z|D')} \right) \right|$$



Thank you!

Questions?

GitHub

Cancer Registry of Norway

NIPH

Norwegian Institute of Public Health