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

Synthetic data

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

European Data Protection Supervisor (EDPS)

Public release of example data.

Publish data alongside journal articles to enable reproducibility.

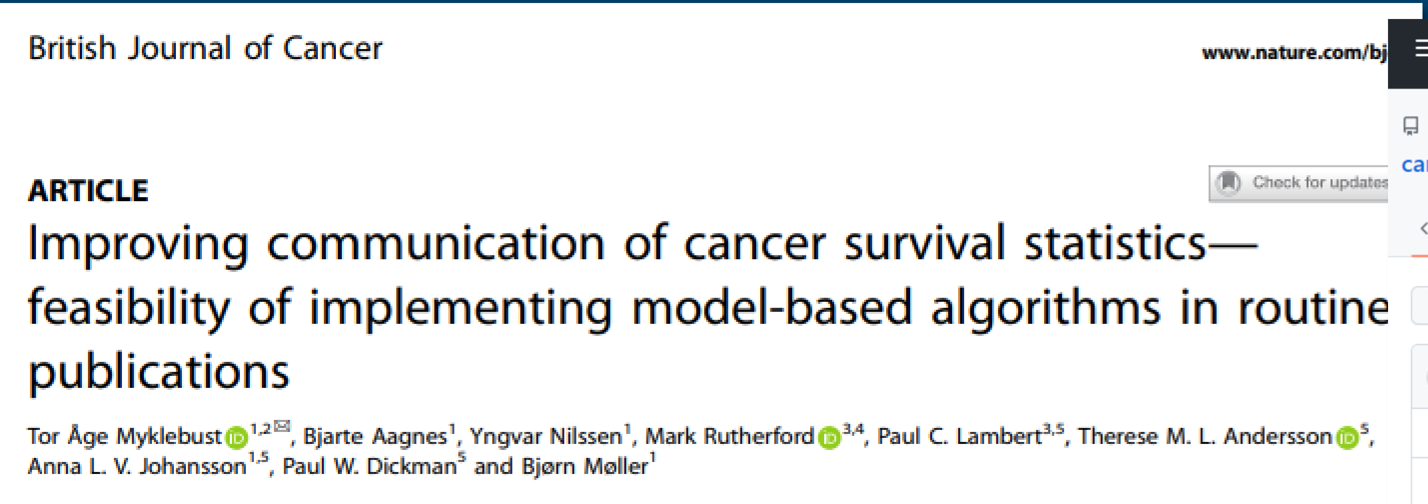
IT development and testing without exposing sensitive information.

Education and training.

Methods and algorithm development.

Cancer Registry of Norway

Motivating dataset



British Journal of Cancer www.nature.com/bj

Check for updates

ARTICLE

Improving communication of cancer survival statistics—feasibility of implementing model-based algorithms in routine publications

Tor Åge Myklebust[1,2✉], Bjarte Aagnes[1], Yngvar Nilssen[1], Mark Rutherford[3,4], Paul C. Lambert[3,5], Therese M. L. Andersson[5], Anna L. V. Johansson[1,5], Paul W. Dickman[5] and Bjørn Møller[1]

© The Author(s), under exclu

BACKGROUND: Routine … e and to inform about prognosis … purposes and targeting different a… tes on a wider range of survival …
METHODS: We used dat… d way of estimating flexible paran… ss in life expectancy across many…
RESULTS: For 21 of 23 … eliable estimates of all desired …
DISCUSSION: It may be … cation of modeling techniques. W… eliable estimates across a range…

British Journal of Cancer

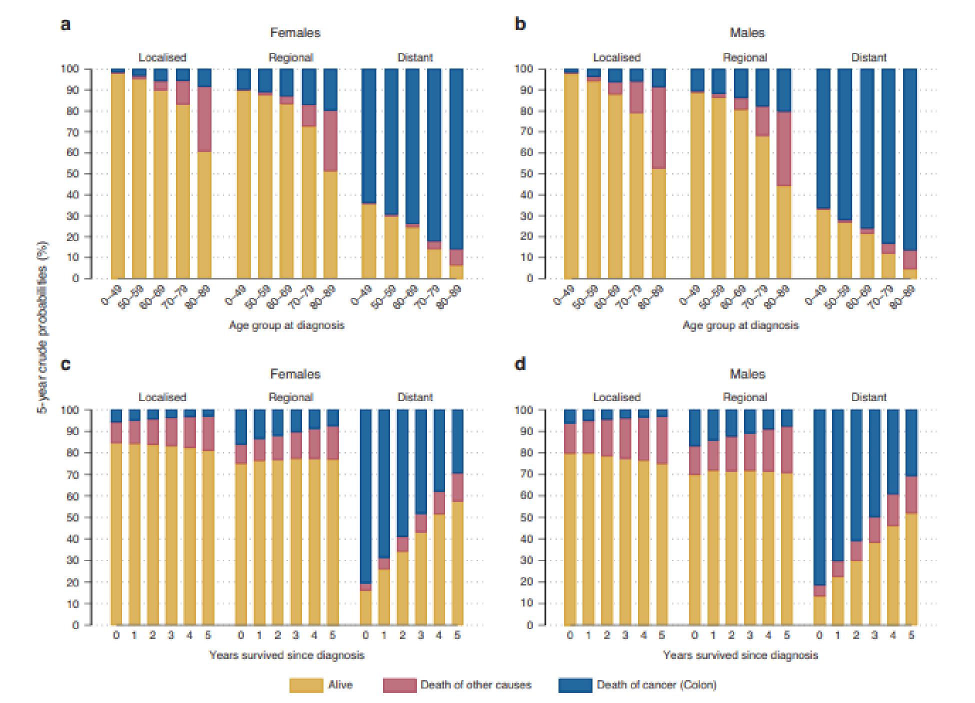
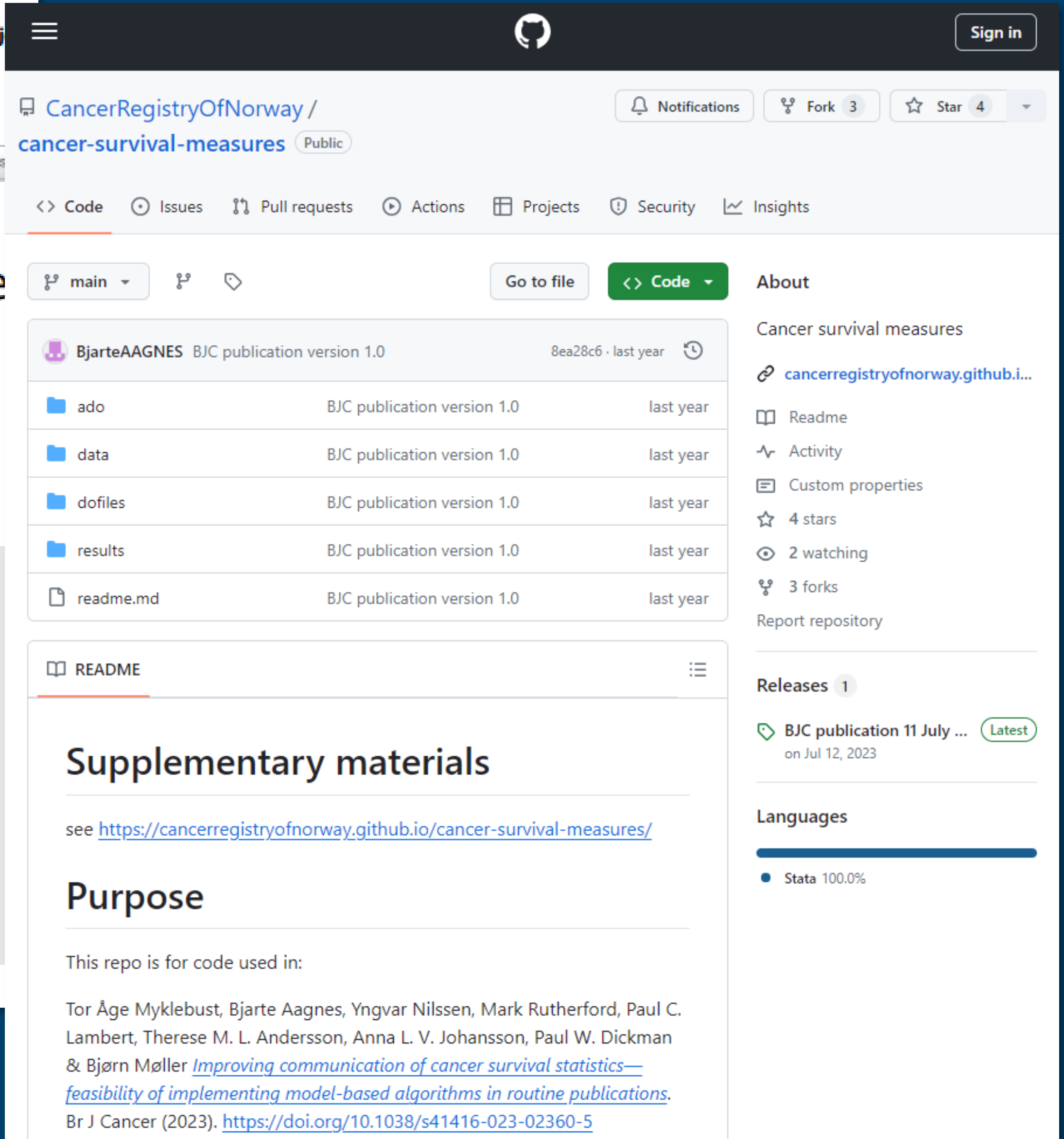
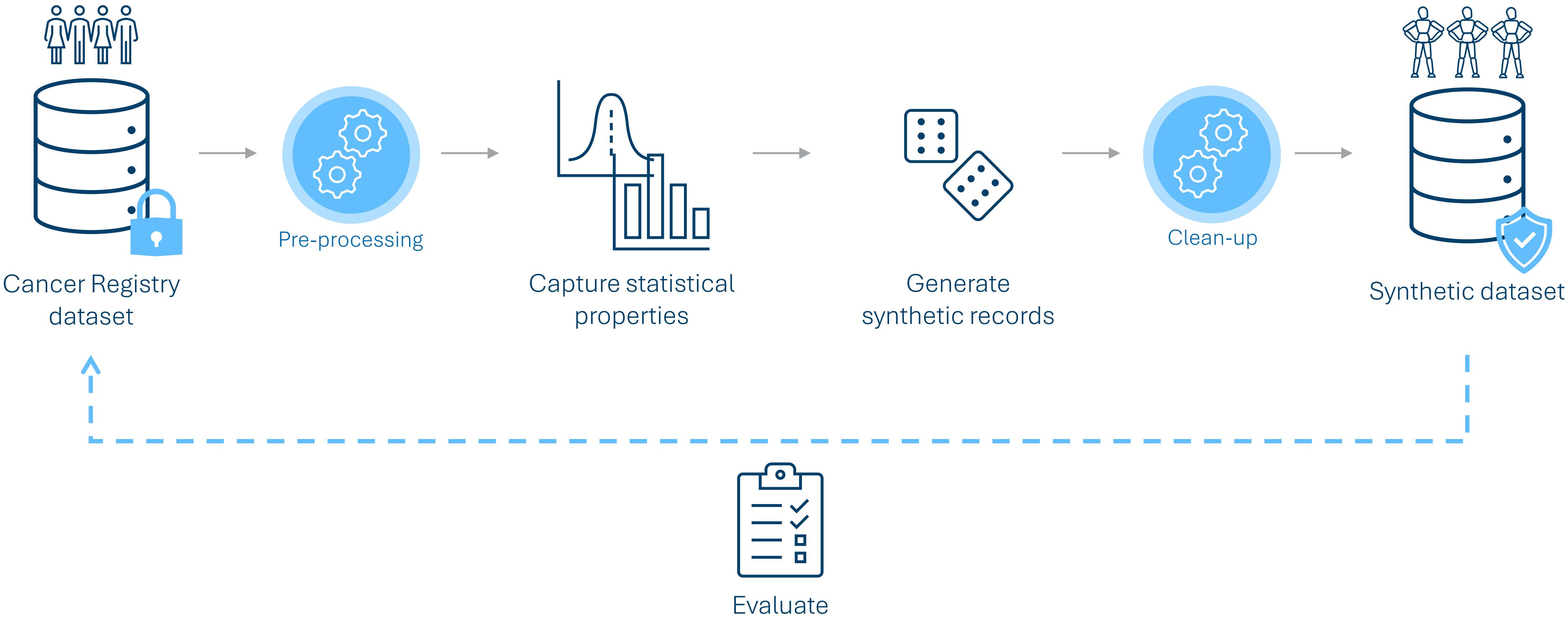


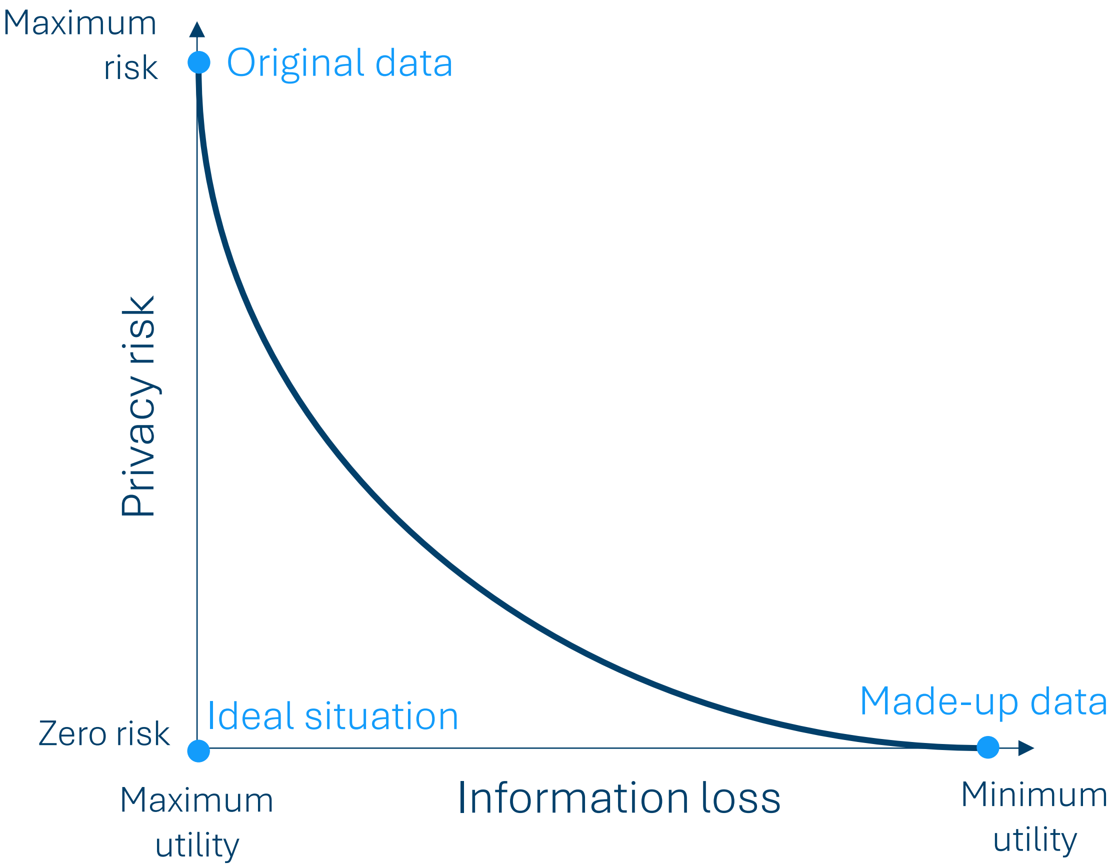
Fig. 1 **The figure shows estimated 5-year crude probabilities of dying from colon cancer, dying from other causes and being alive.** Results are stratified by sex, stage and age group at diagnosis (**a**, **b**) and by sex, stage and number of years survived since diagnosis (**c**, **d**). Crude probabilities are also referred to as cumulative incidence or "real-world probabilities".



Synthetic data generation



Privacy-utility trade-off



Synthetic data generators

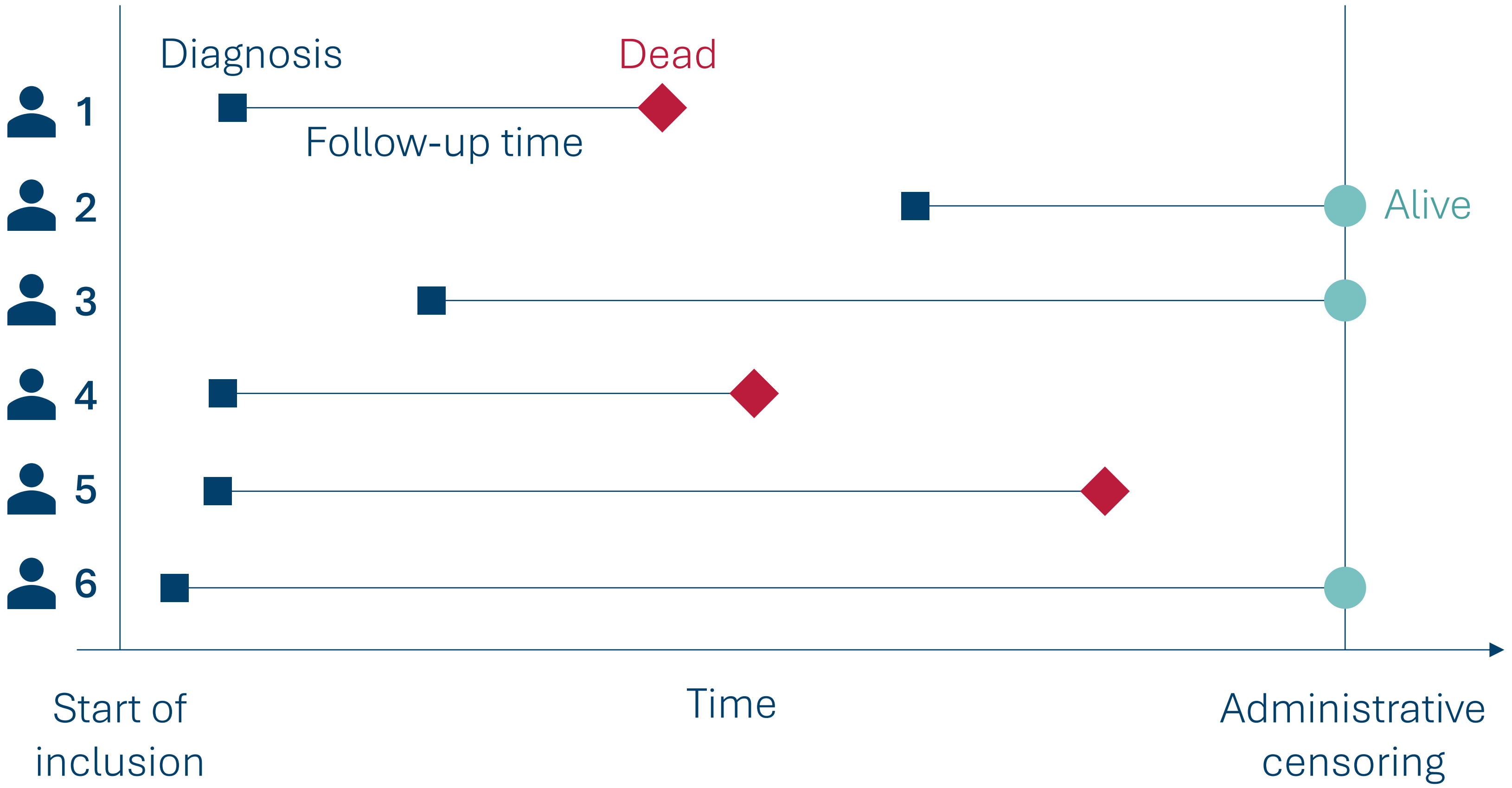
Statistical methods

- 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



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

52 162 patients

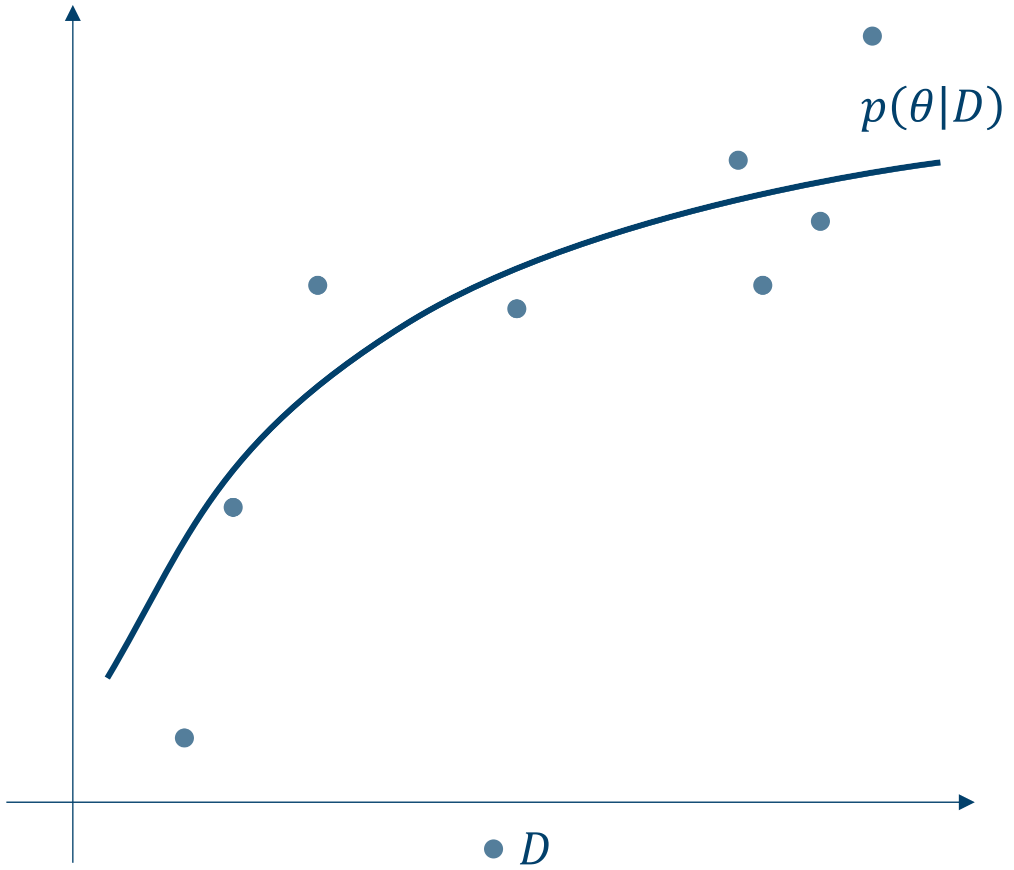
Diagnosed 2002-2021

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

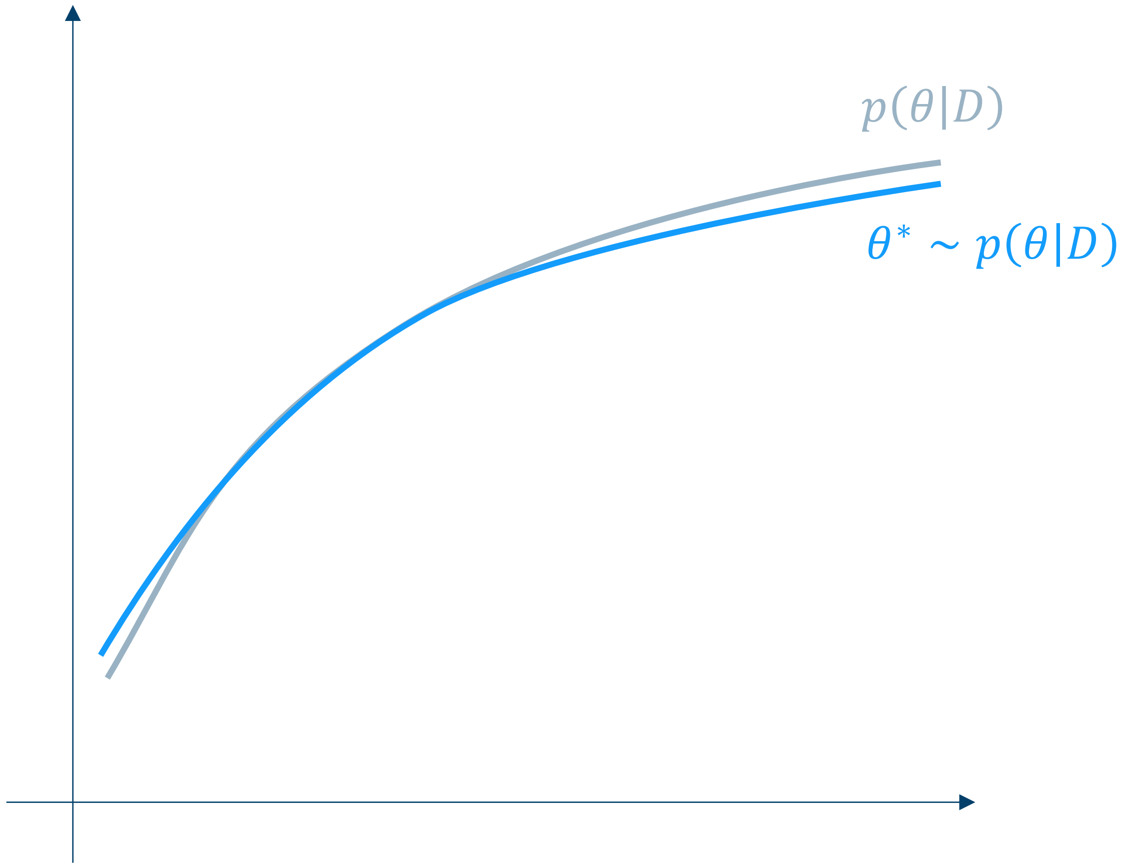
Original data



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

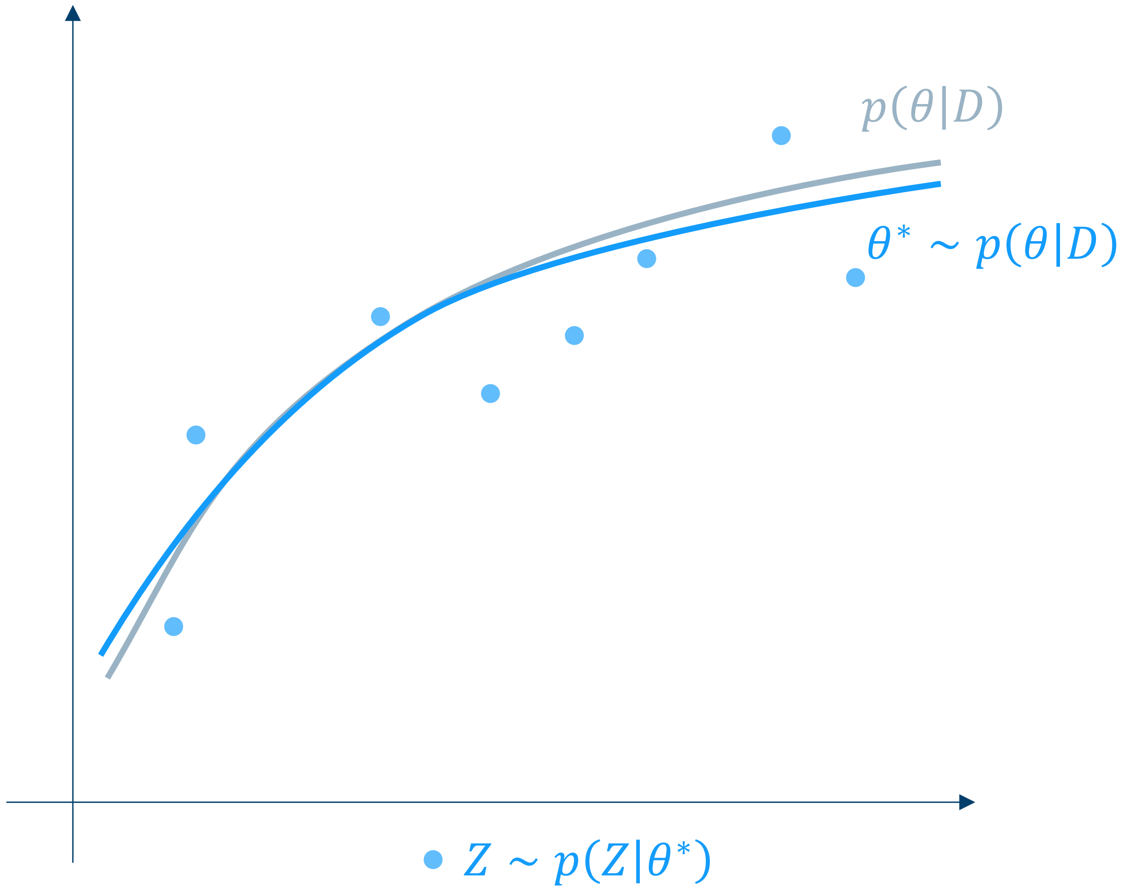
Statistical model



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

Synthetic 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

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:

Age_i

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

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:

$$\mathrm{Age}_i$$

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

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:

$$\mathrm{Age}_i$$

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

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

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:

$$\mathrm{Age}_i$$

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

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

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:

$$\mathrm{Age}_i$$

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

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

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

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:

$$\mathrm{Age}_i$$

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

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

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

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

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		

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

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

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:

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

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

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

Synthetic records:

54	Male	Regional	06.03.2005	4.31	Dead

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:

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(3-year periods).

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

Experimental design

Sequential regressions:

Age_i

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

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

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

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

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

50 datasets from each model

Synthetic data evaluation

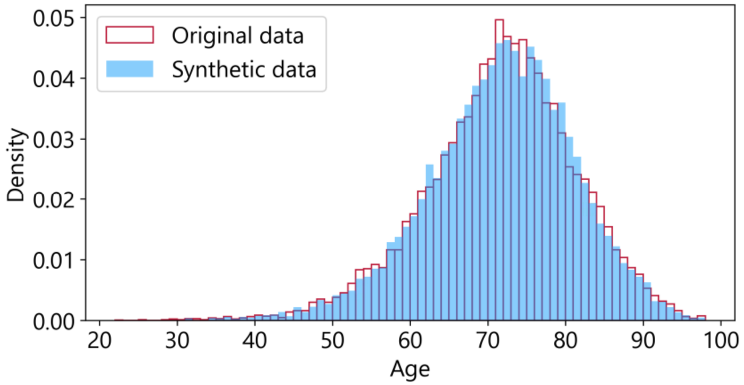
Utility

- Similarity
- Specific utility

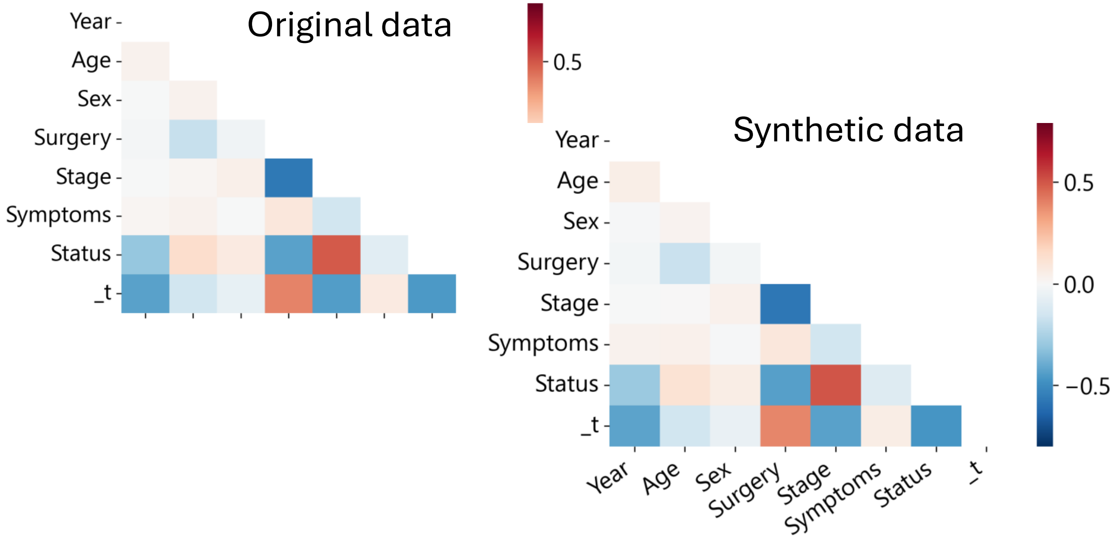
Privacy

Synthetic data utility

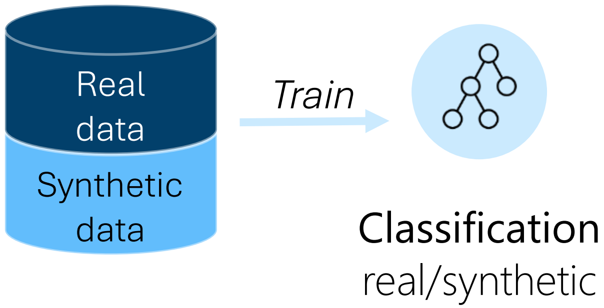
Univariate



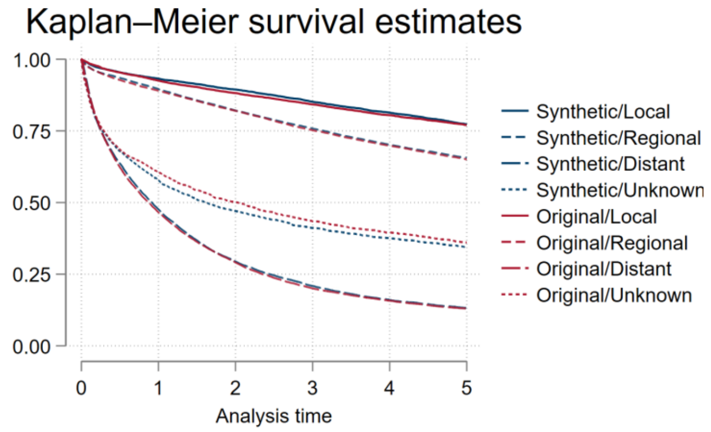
Bivariate



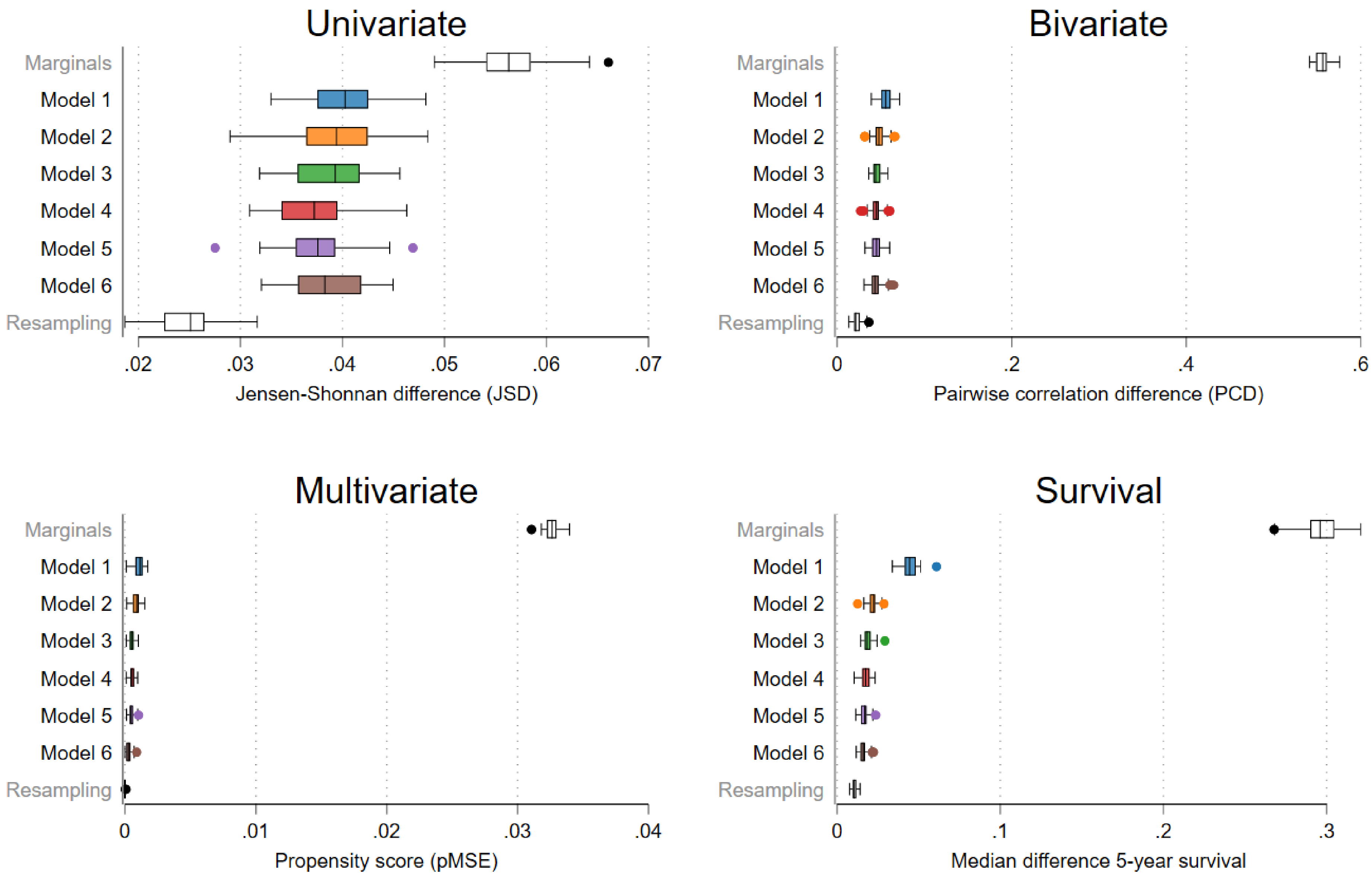
Multivariate



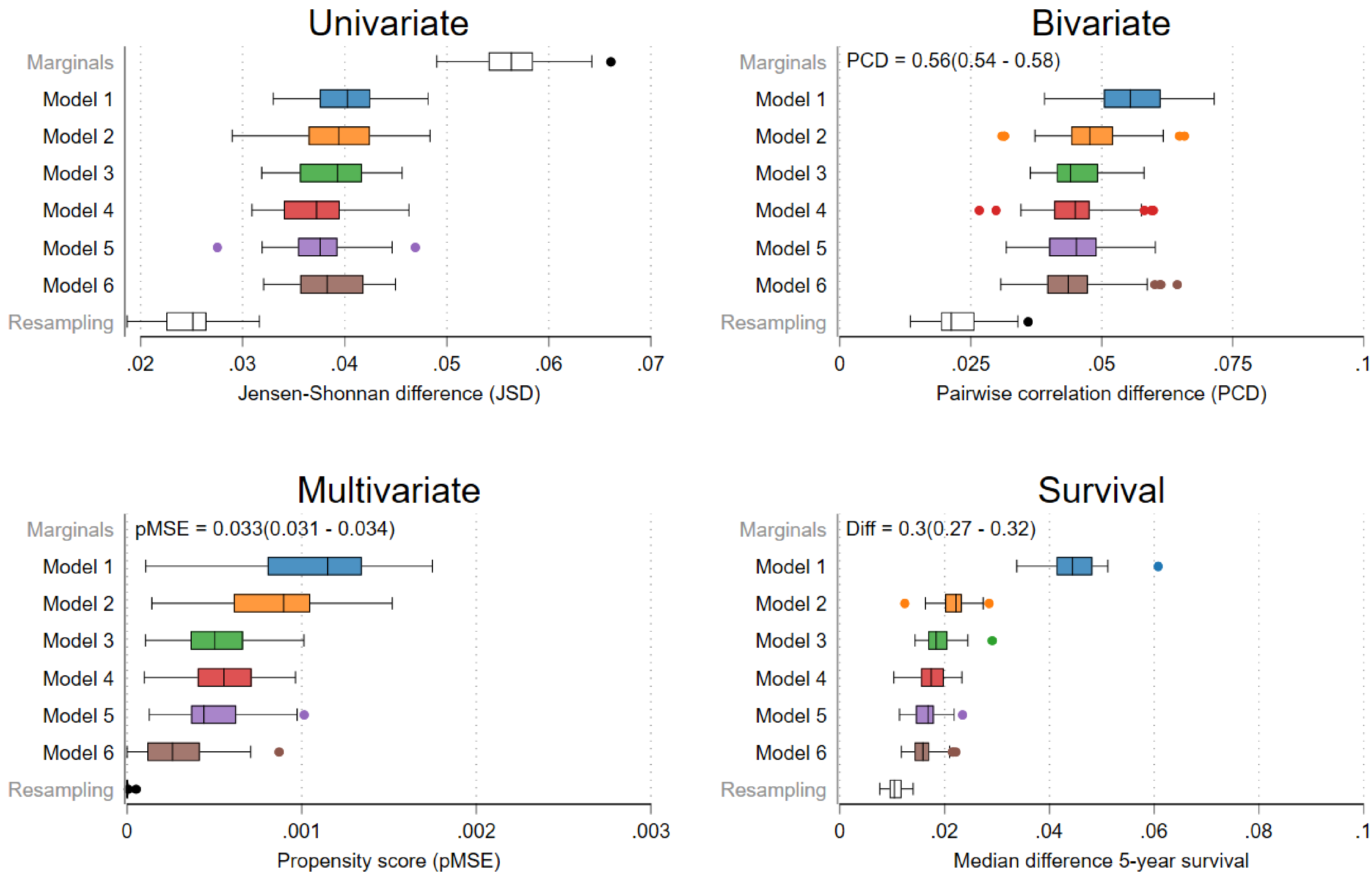
Survival



Synthetic data utility



Synthetic data utility



Synthetic data evaluation

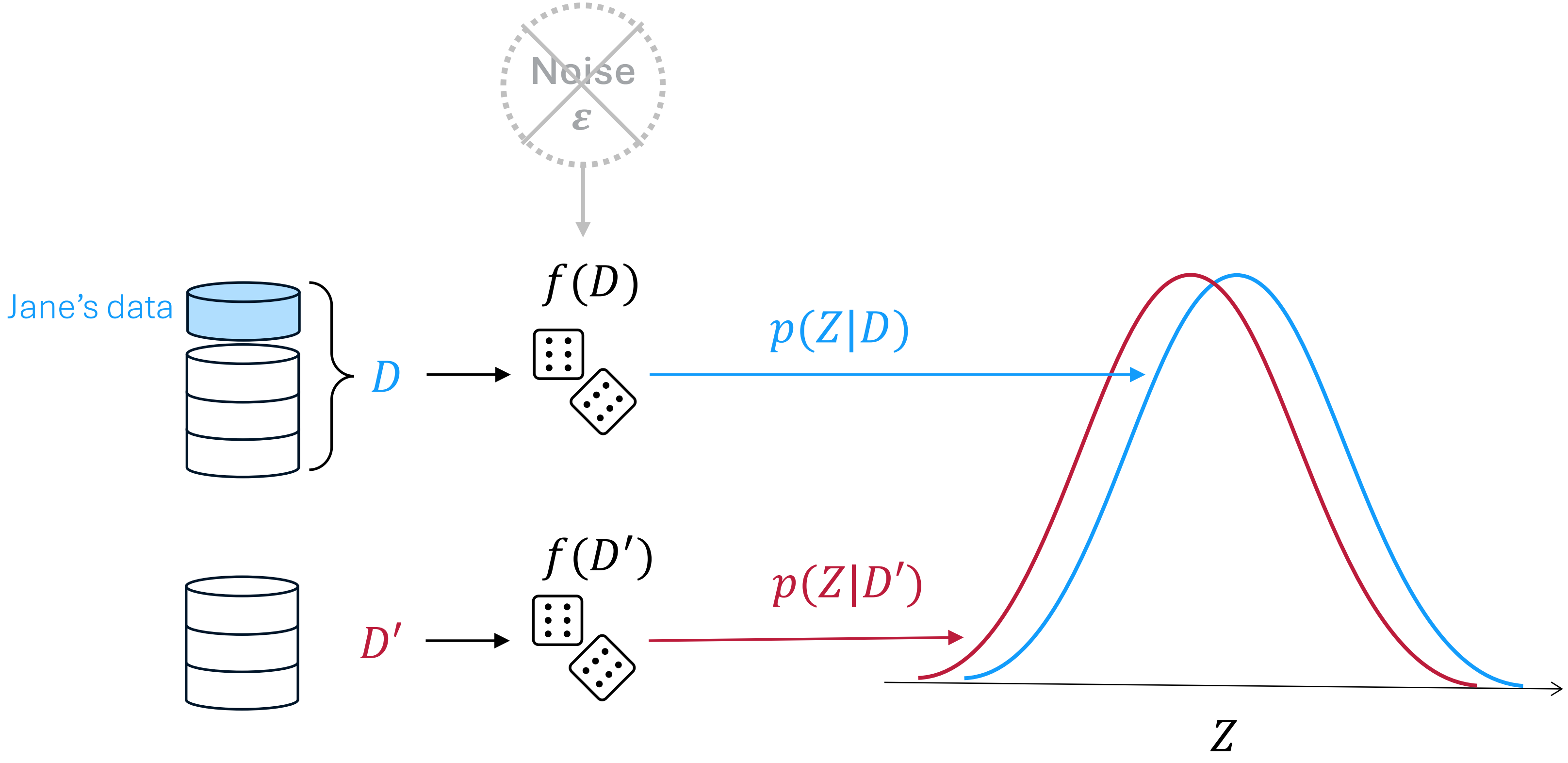
Utility

- Similarity
- Specific utility

Privacy

- Inference attacks
- Mathematical definitions $f(x)$

Differential privacy



For all $Z \in \mathrm{Range}(f)$:

$$\frac{p(Z|D)}{p(Z|D')} \leq e^{\varepsilon}$$

$$\left|\log\left(p(Z|D)\right) - \log\left(p(Z|D')\right)\right| \leq \varepsilon$$

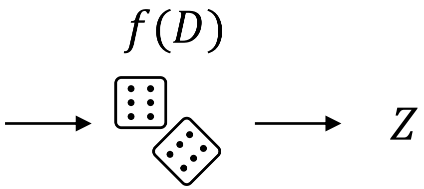
$$\varepsilon_{emp.} = \max\left|\log\left(p(Z|D)\right) - \log\left(p(Z|D')\right)\right|$$

$$\varepsilon_{emp.} \ll \varepsilon$$

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

Differential privacy auditing

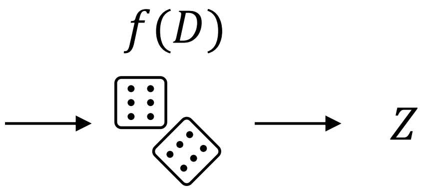
	i	Age	Sex	Stage	Date of diagnosis	Follow-up time	Status
Jane	1	77	Female	Distant	05.03.2002	5.12	Dead
	2	50	Male	Local	21.08.2011	9.51	Alive
	3	63	Female	Unknown	30.12.2018	3.02	Alive
	4	46	Male	Regional	02.07.2005	3.45	Dead
	5	60	Male	Local	09.10.2016	13.83	Dead



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).

Differential privacy auditing

	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

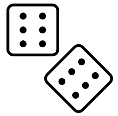


Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Differential privacy auditing

	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

$f(D)$

$\longrightarrow$  $\longrightarrow Z$

Estimate:

$$p(Z|D, \text{stage}_1 = \text{Localised})$$
$$p(Z|D, \text{stage}_1 = \text{Regional})$$
$$p(Z|D) \quad \textit{(stage}_1 = \textit{Distant)}$$
$$p(Z|D, \text{stage}_1 = \text{Unknown})$$

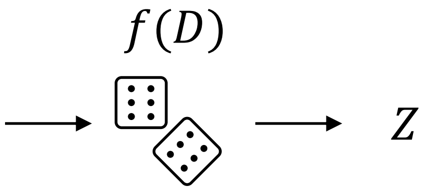
Information leakage:

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

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).

Differential privacy auditing

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:

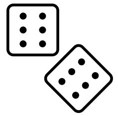
$$\max_{y} \left| \log\big(p(Z|D, \text{stage}_2 = y)\big) - \log\big(p(Z|D)\big) \right|$$

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).

Differential privacy auditing

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

$f(D)$

$\longrightarrow$  $\longrightarrow$ Z

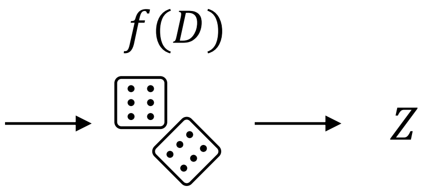
Information leakage:

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

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).

Differential privacy auditing

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



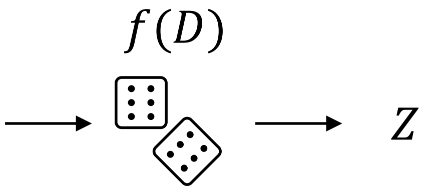
Information leakage:

$$\max_y \left| \log\big(p(Z|D, \text{stage}_4 = y)\big) - \log\big(p(Z|D)\big) \right|$$

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).

Differential privacy auditing

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



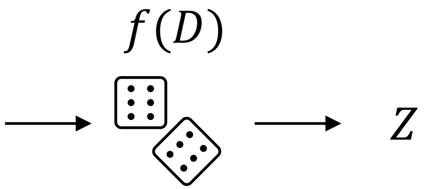
Information leakage:

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

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).

Differential privacy auditing

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



Differential privacy audit:

$$\varepsilon_{emp.} = \max_{i,y} \left| \log\left(p(Z|D, \text{stage}_i = y)\right) - \log\left(p(Z|D)\right) \right|$$

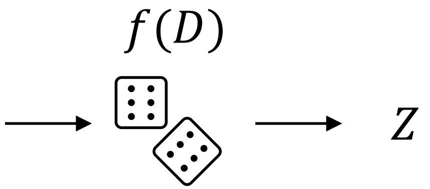
Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Differential privacy auditing

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

⋮

52 162 × 3 = 156 486



Differential privacy audit:

$$\varepsilon_{emp.} = \max_{i,y} \left| \log\big(p(Z|D, \text{stage}_i = y)\big) - \log\big(p(Z|D)\big) \right|$$

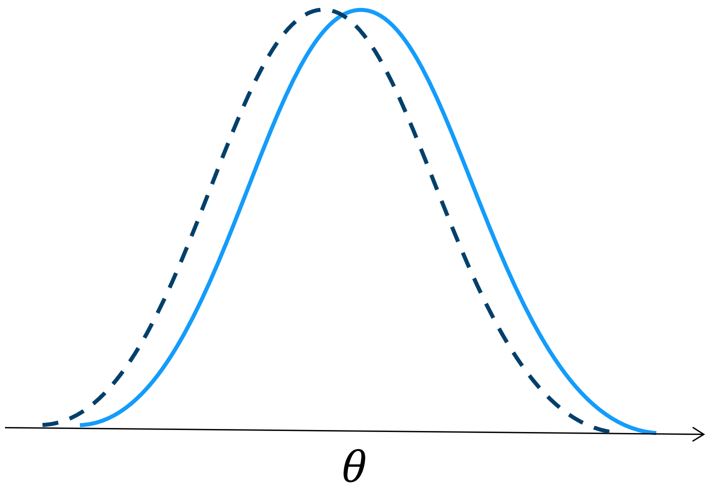
Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1).

Differential privacy auditing

$$\varepsilon_{emp.} = \max_{i,y} \left| \log\left(p(Z|D, \text{stage}_i = y)\right) - \log\left(p(Z|D)\right) \right|$$

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

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



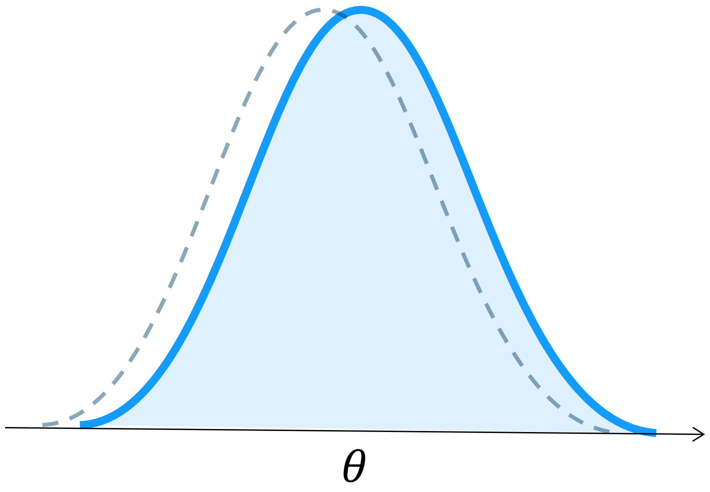
Hu, J., Reiter, J. P., & Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. *International conference on privacy in statistical databases*.

Differential privacy auditing

$$\varepsilon_{emp.} = \max_{i,y} \left| \log\left(p(Z|D, \text{stage}_i = y)\right) - \log\left(p(Z|D)\right) \right|$$

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

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



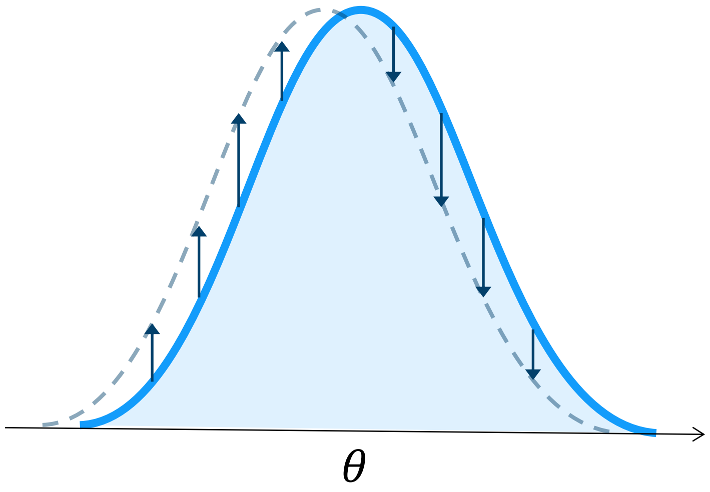
Hu, J., Reiter, J. P., & Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. *International conference on privacy in statistical databases*.

Differential privacy auditing

$$\varepsilon_{emp.} = \max_{i,y} \left| \log\left(p(Z|D, \text{stage}_i = y)\right) - \log\left(p(Z|D)\right) \right|$$

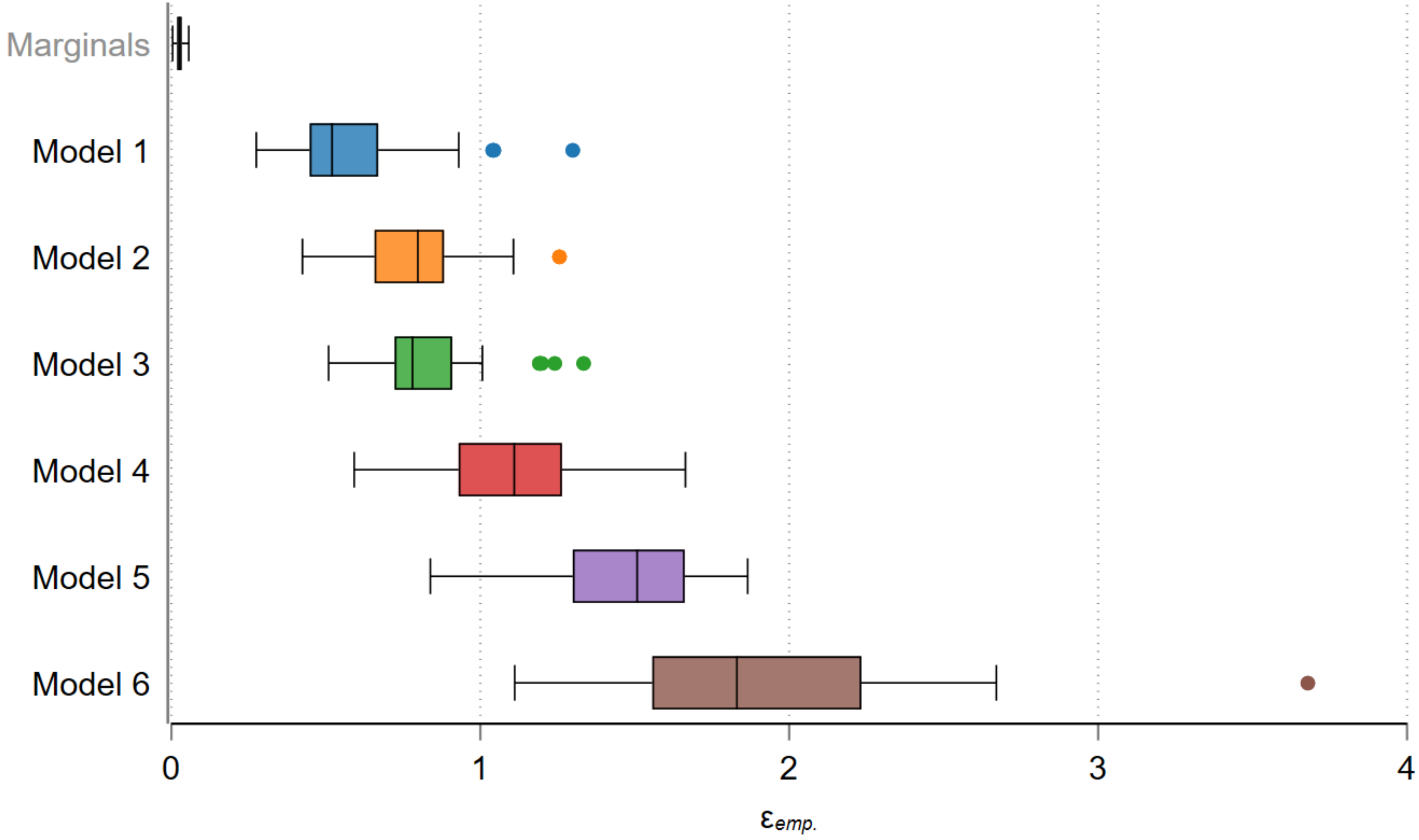
$$D_{\text{stage}_i = y} \approx D$$

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



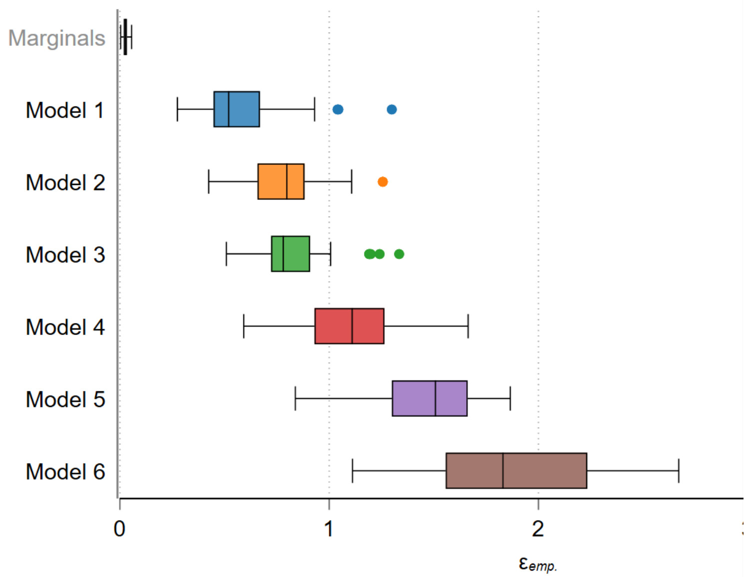
Hu, J., Reiter, J. P., & Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. *International conference on privacy in statistical databases*.

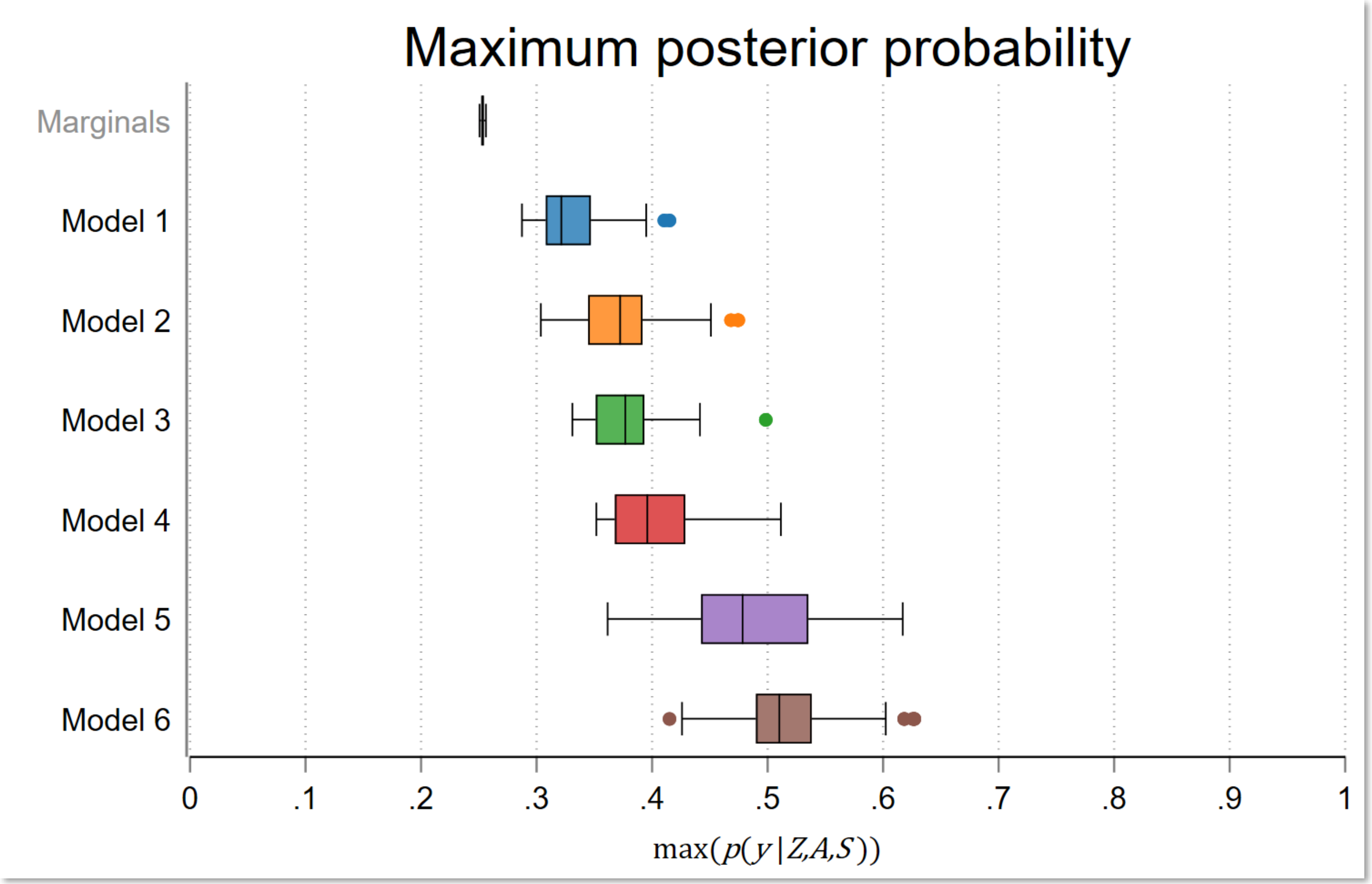
Differential privacy auditing



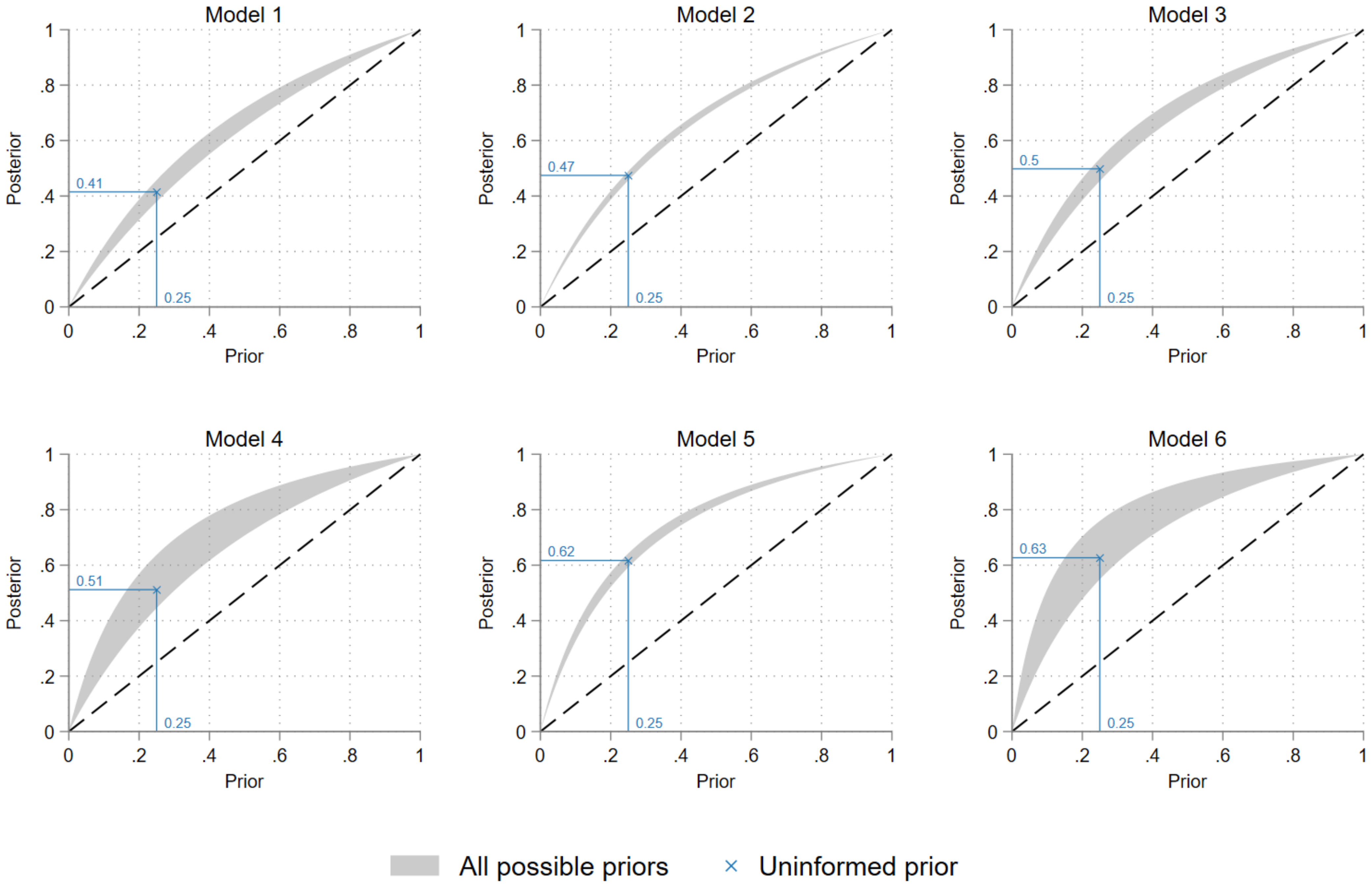
Cancer
Registry of Norway

Differential privacy auditing

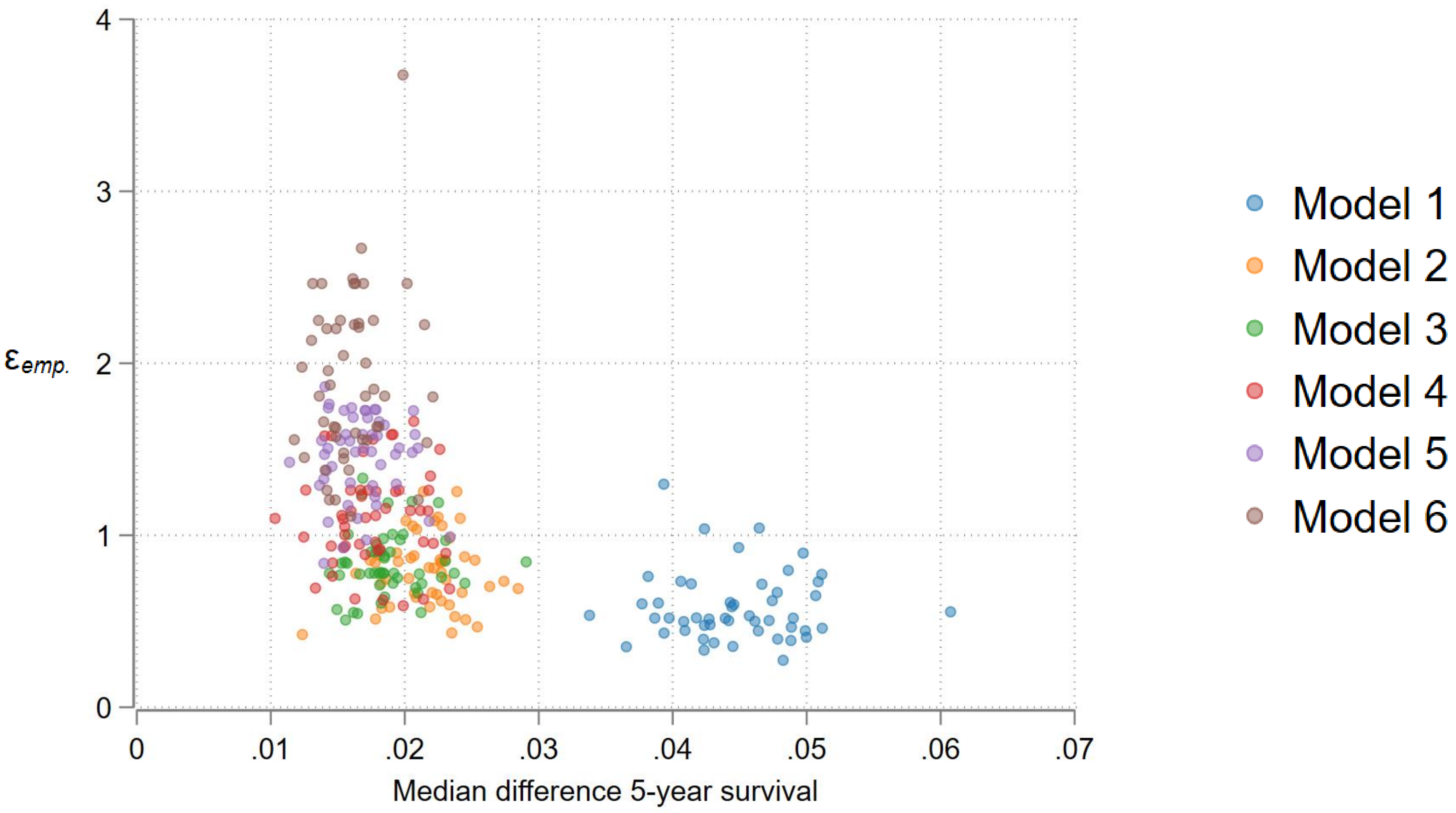




Cancer Registry of Norway

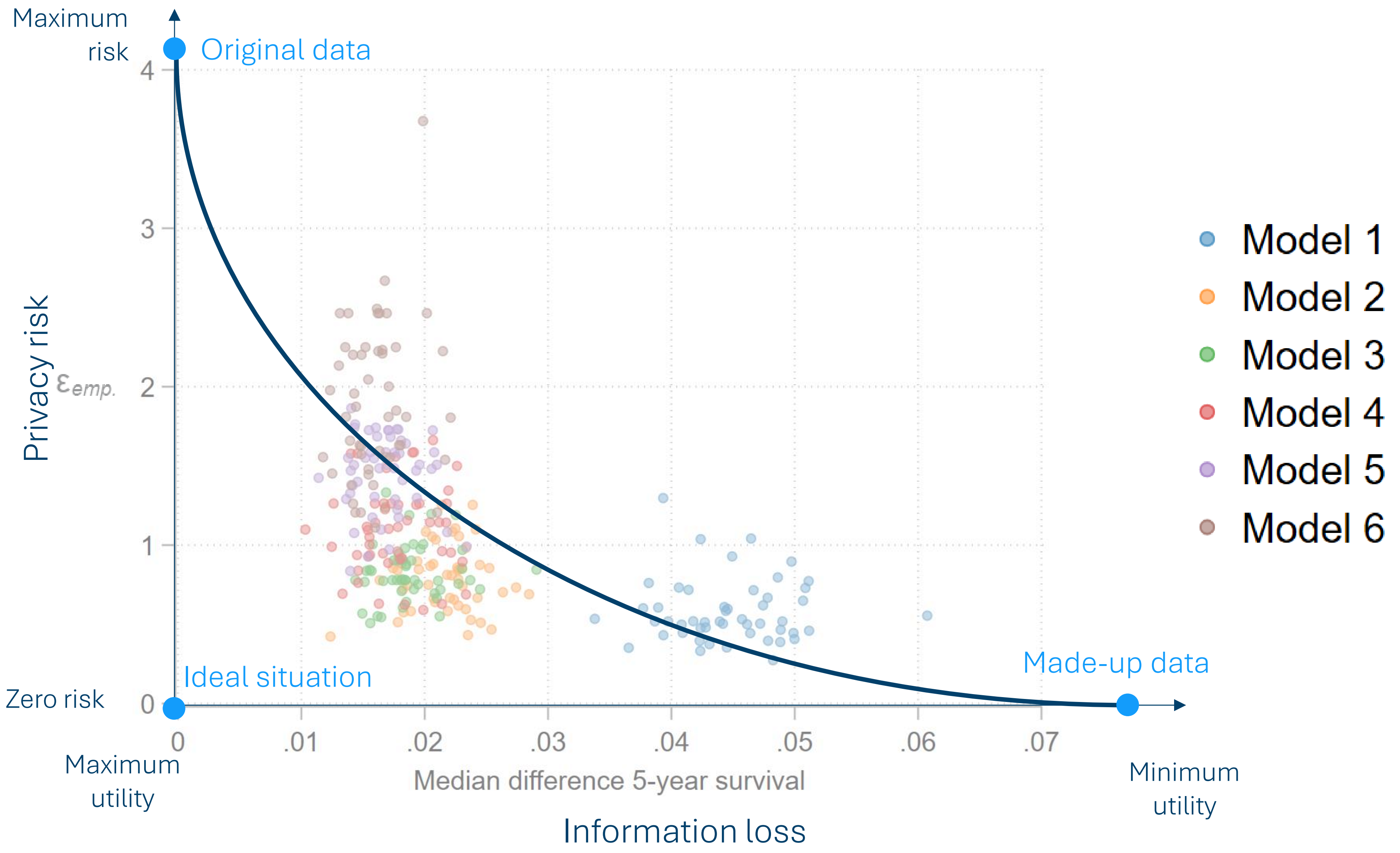

Model 1
Model 2
Model 3
Model 4
Model 5
Model 6
Posterior
Prior
0.41
0.47
0.5
0.51
0.62
0.63
0.25
All possible priors
Uninformed prior

Privacy-utility trade-off



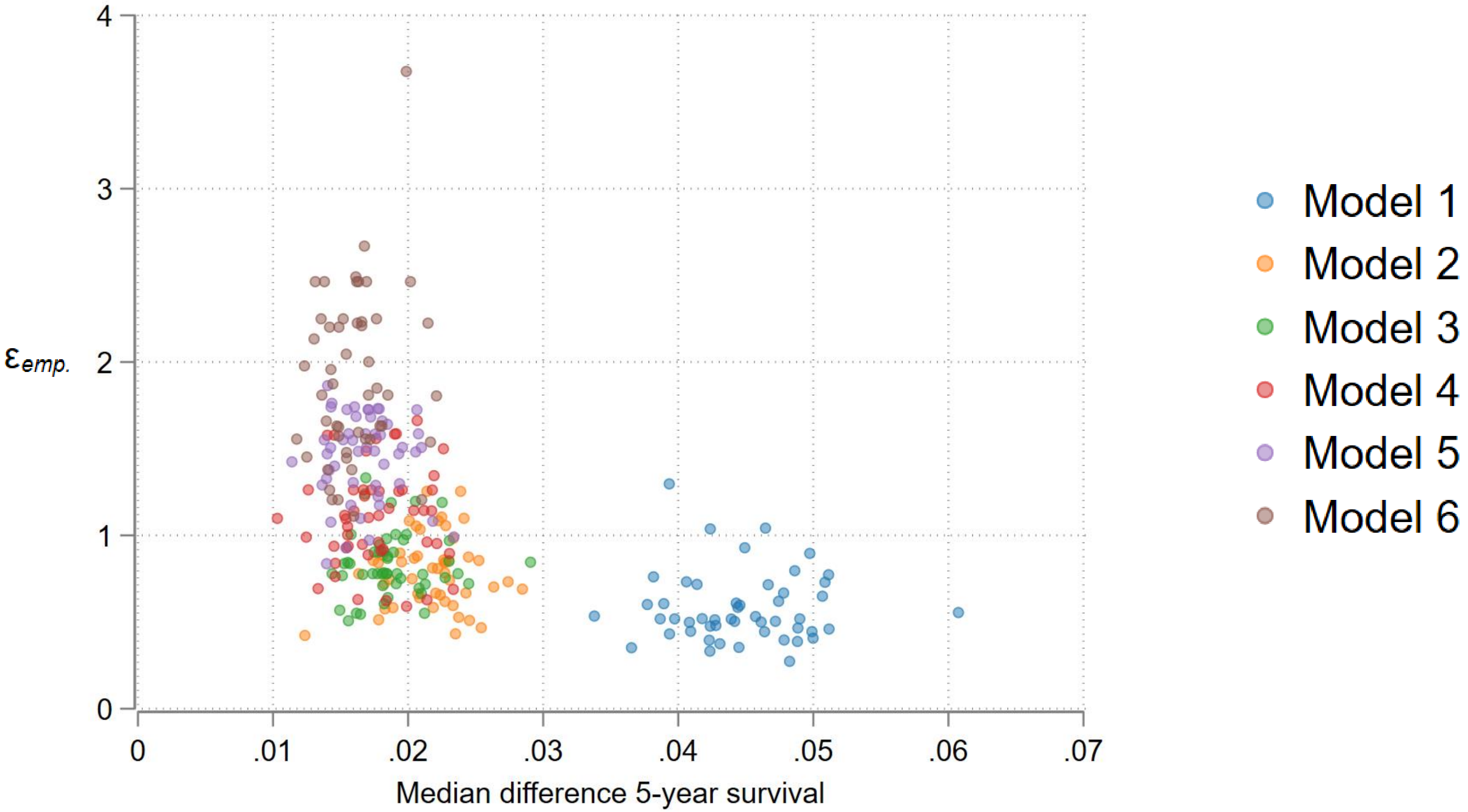
Cancer Registry of Norway

Privacy-utility trade-off



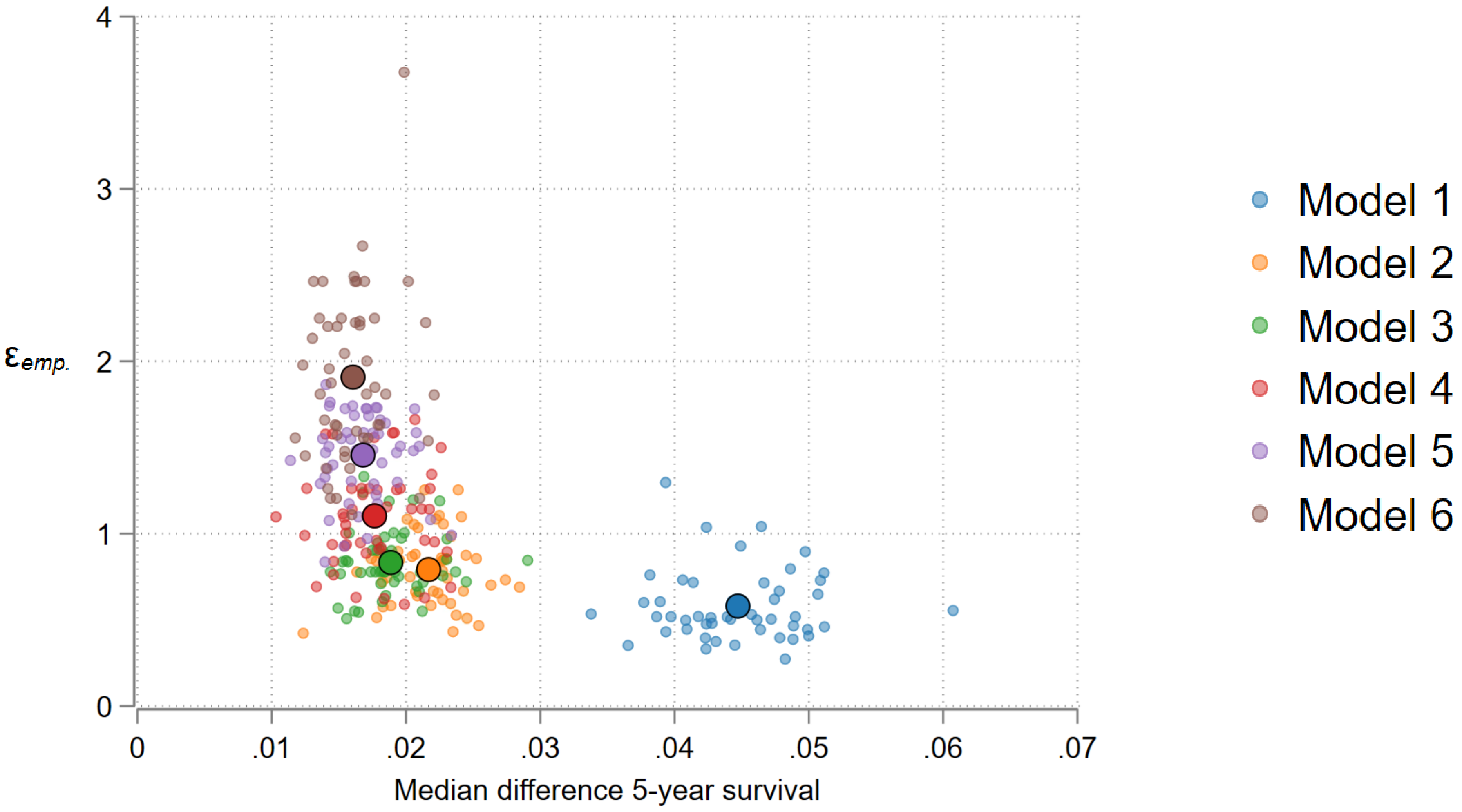
Cancer Registry of Norway

Privacy-utility trade-off



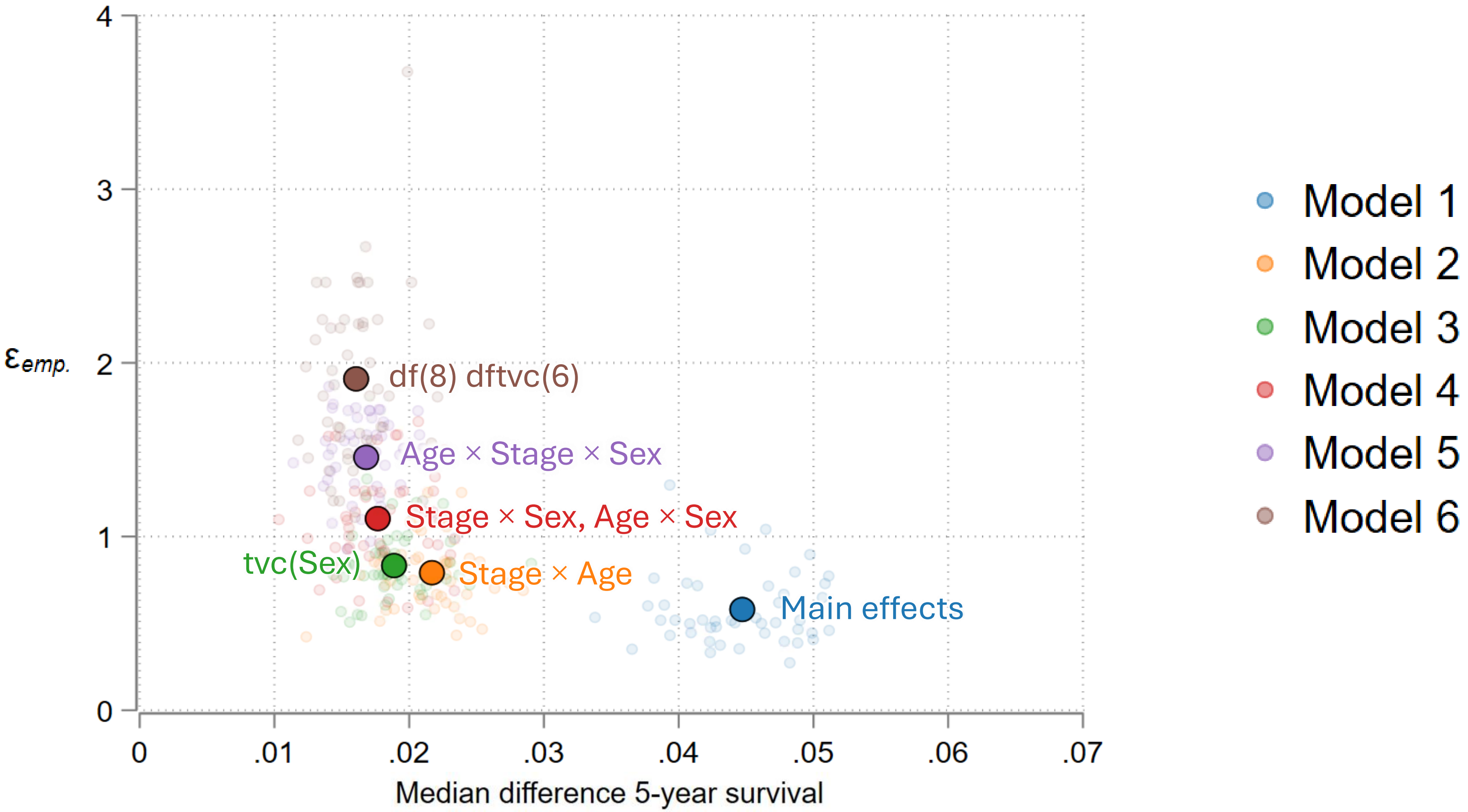
Cancer Registry of Norway

Privacy-utility trade-off



Cancer Registry of Norway

Privacy-utility trade-off



Cancer Registry of Norway

Thank you!

Questions?

GitHub

Cancer Registry of Norway

A part of the Norwegian Institute of Public Health