Making biomedical research data publicly available while protecting confidentiality

Sabine Hoffmann, Sarah Friedrich, Jan Kapar, Marvin Wright

September 19, 2024

Motivation

- Increasing awareness that data sharing:
  - Improves transparency, credibility and reproducibility
  - Increases reuse potential of scientific studies
  - Makes evidence synthesis more efficient

Motivation

- Increasing awareness that data sharing:
  - Improves transparency, credibility and reproducibility
  - Increases reuse potential of scientific studies
  - Makes evidence synthesis more efficient

$\Rightarrow$ Journals and funders are increasingly incentivizing or even requiring data sharing practices

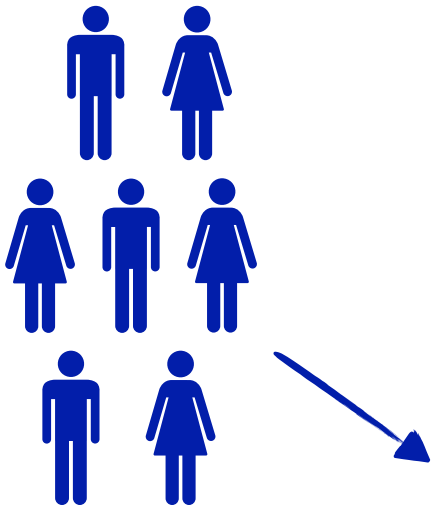
Motivation

- Increasing awareness that data sharing:
  - Improves transparency, credibility and reproducibility
  - Increases reuse potential of scientific studies
  - Makes evidence synthesis more efficient

$\Rightarrow$ Journals and funders are increasingly incentivizing or even requiring data sharing practices

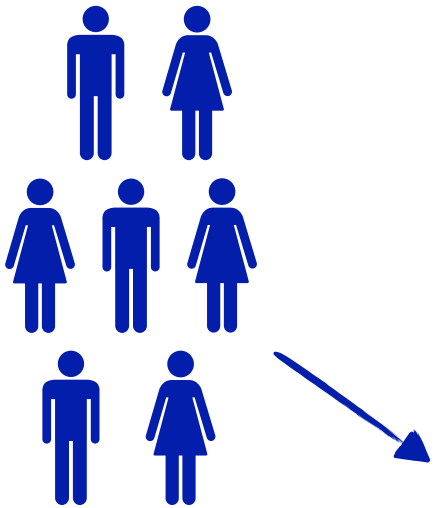
$\Rightarrow$ Many researchers lack skills and knowledge to make their data publicly available while protecting confidentiality

Challenges when making research data publicly available



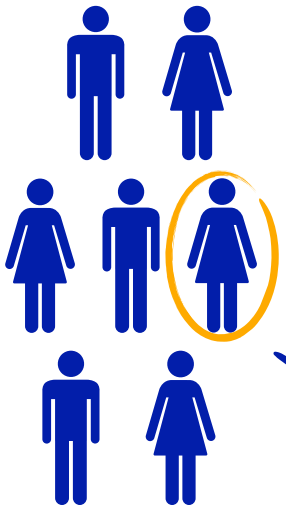
Name	Date of birth	Location	Sensitive X1	Sensitive X2	Heart rate

Challenges when making research data publicly available



ID	Date of birth	Location	Sensitive X1	Sensitive X2	Heart rate

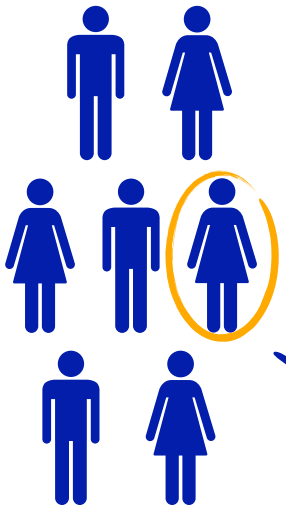
Challenges when making research data publicly available



Identity disclosure

ID	Date of birth	Location	Sensitive X1	Sensitive X2	Heart rate

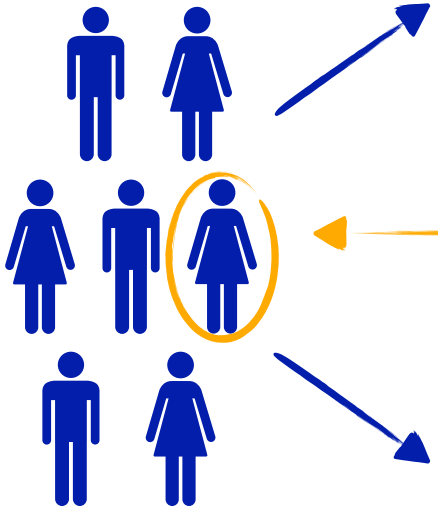
Challenges when making research data publicly available



Attribute disclosure

ID	Date of birth	Location	Sensitive X1	Sensitive X2	Heart rate

Challenges when making research data publicly available



Identity disclosure

Attribute disclosure

ID	Date of birth	Location	Sensitive X1	Sensitive X2	Heart rate

Challenges when making research data publicly available

- How to share biomedical research data?

Challenges when making research data publicly available

- How to share biomedical research data?
- How to evaluate the quality of the shared data set in terms of disclosure risk and utility?

Challenges when making research data publicly available

- How to share biomedical research data?
- How to evaluate the quality of the shared data set in terms of disclosure risk and utility?
- What are the possible uses of shared data sets?

Approaches to limit statistical disclosure

Reduce information

- Dropping variables
- Categorizing continuous variables or aggregating categories
- Censoring

Approaches to limit statistical disclosure

Reduce information

- Dropping variables
- Categorizing continuous variables or aggregating categories
- Censoring

Data perturbation

- Data swapping
- Adding noise

Approaches to limit statistical disclosure

Reduce information

- Dropping variables
- Categorizing continuous variables or aggregating categories
- Censoring

Data perturbation

- Data swapping
- Adding noise

Generate synthetic data

Approaches to limit statistical disclosure

Reduce information

- Dropping variables
- Categorizing continuous variables or aggregating categories
- Censoring

Data perturbation

- Data swapping
- Adding noise

Generate synthetic data

- Full or partial synthesis

Approaches to limit statistical disclosure

Reduce information

Data perturbation

Generate synthetic data

- Full or partial synthesis
- Methods:
  - Parametric methods

Approaches to limit statistical disclosure

Reduce information

Data perturbation

Generate synthetic data

- Full or partial synthesis
- Methods:
  - Parametric methods
  - Deep learning:
    - Autoencoders
    - Generative Adversarial Networks

Approaches to limit statistical disclosure

Reduce information

Data perturbation

Generate synthetic data

- Full or partial synthesis
- Methods:
  - Parametric methods
  - Deep learning:
    - Autoencoders
    - Generative Adversarial Networks
  - Tree-based methods
    - Synthpop [Nowok et al., 2016]
    - Adversarial Random Forests [Watson et al., 2023]

Approaches to limit statistical disclosure

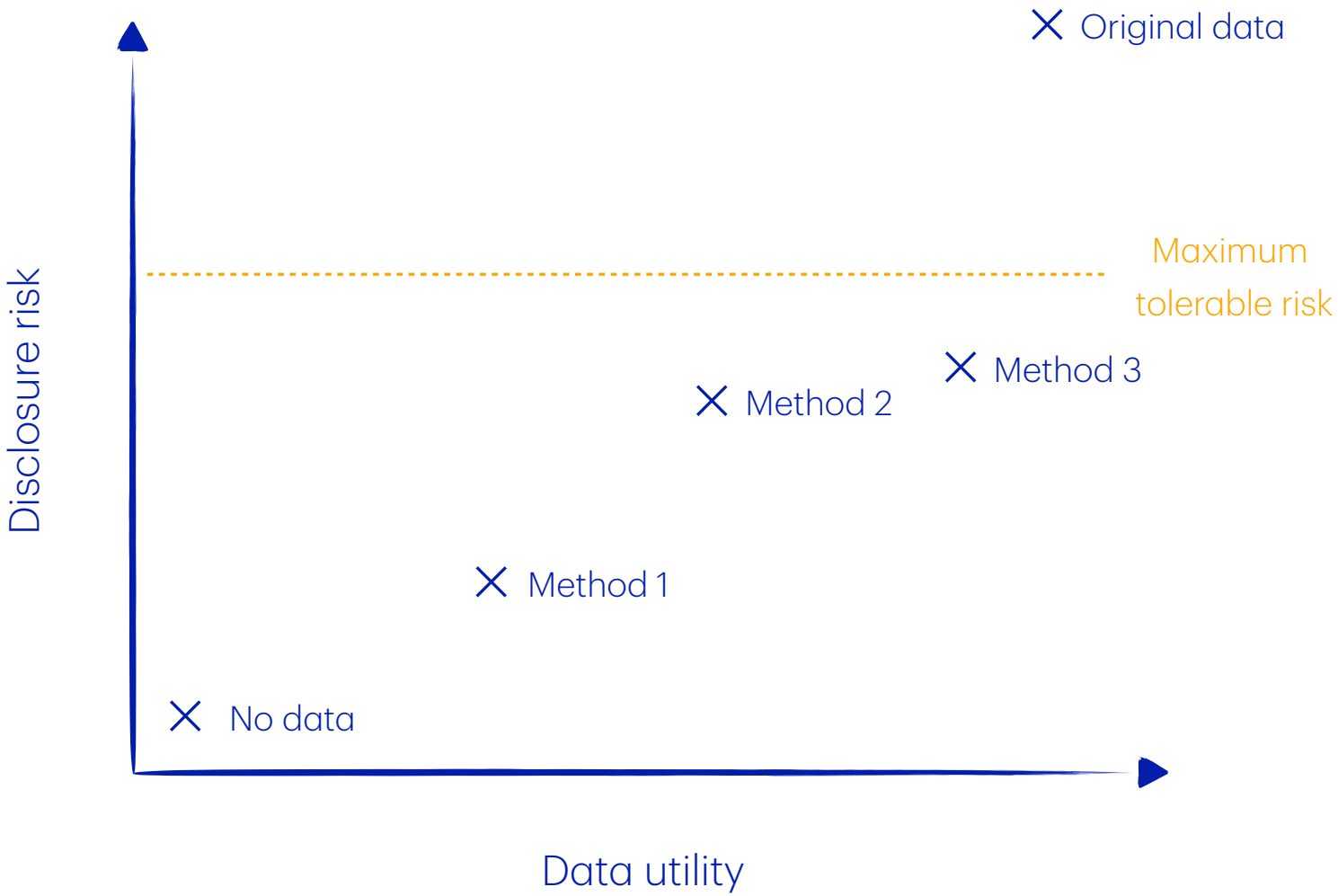
Reduce information

Data perturbation

Generate synthetic data

- Full or partial synthesis
- Methods:
  - Parametric methods
  - Deep learning:
    - Autoencoders
    - Generative Adversarial Networks
  - Tree-based methods
    - Synthpop [Nowok et al., 2016]
    - Adversarial Random Forests [Watson et al., 2023]
  - Bayesian networks

Evaluating the quality of the shared data set



Evaluating data utility

- Precision and recall based metrics

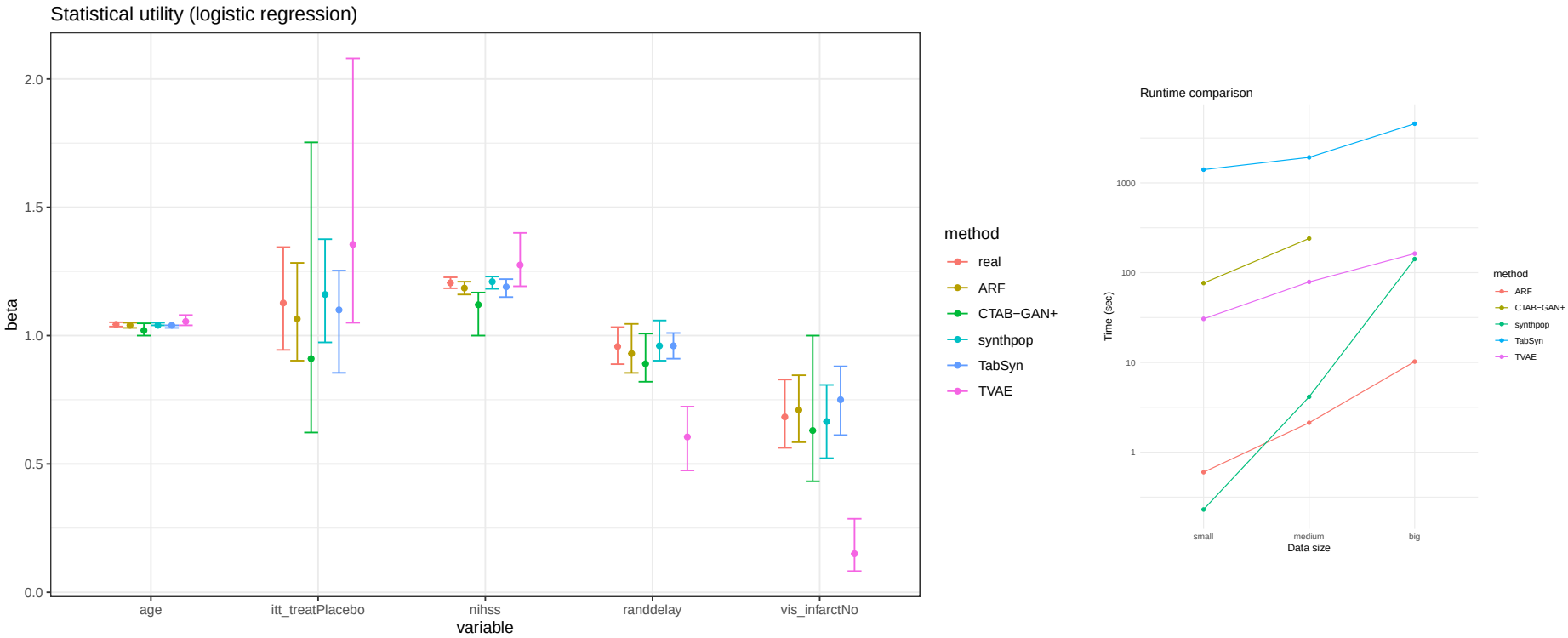
Evaluating data utility

- Precision and recall based metrics
- Classifier two sample test

Evaluating data utility

- Precision and recall based metrics
- Classifier two sample test
- Machine learning efficacy and statistical utility

Evaluating data utility



Evaluating disclosure risk

- Privacy attacks

Evaluating disclosure risk

- Privacy attacks
- Share of identical records

Evaluating disclosure risk

- Privacy attacks
- Share of identical records
- Distance to closest record

Evaluating disclosure risk

- Privacy attacks
- Share of identical records
- Distance to closest record
- Nearest neighbour distance ratio

What are the possible uses of shared data sets?

- **Synthpop:** "The original aim of producing synthetic data has been to provide publicly available datasets that can be used for inference in place of the actual data. (...) Our aim in writing the synthpop package for R is a more modest one of providing test data for users of confidential datasets. (...) These test datasets should resemble the actual data as closely as possible, but would never be used in any final analyses. [Nowok et al., 2016]

What are the possible uses of shared data sets?

- **Synthpop:** "The original aim of producing synthetic data has been to provide publicly available datasets that can be used for inference in place of the actual data. (...) Our aim in writing the synthpop package for R is a more modest one of providing test data for users of confidential datasets. (...) These test datasets should resemble the actual data as closely as possible, but would never be used in any final analyses. [Nowok et al., 2016]
- **Simulacrum:** "The Simulacrum has a similar data structure to the real data in the CAS and maintains many of the statistical properties of the original data with a high degree of accuracy. (...) However, it does have limitations: more complex statistical properties are not so well captured. (...) Therefore, it is important that Simulacrum alone is not used to make epidemiological inferences or clinical decisions."

Open questions

- What are the intended uses of synthetic data sets?

Open questions

- What are the intended uses of synthetic data sets?
- Do we need partial or full synthesis?

Open questions

- What are the intended uses of synthetic data sets?
- Do we need partial or full synthesis?
- Does sampling already provide some confidentiality?

Open questions

- What are the intended uses of synthetic data sets?
- Do we need partial or full synthesis?
- Does sampling already provide some confidentiality?
- Challenges:
  - High-dimensional data
  - Longitudinal data
  - Logical constraints between variables

Open questions

- What are the intended uses of synthetic data sets?
- Do we need partial or full synthesis?
- Does sampling already provide some confidentiality?
- Challenges:
  - High-dimensional data
  - Longitudinal data
  - Logical constraints between variables
- Synthetic data set generation procedure enabling various add-ons:
  - Missing data
  - Flexible functional forms
  - Measurement error
  - Counterfactual outcomes

Open questions

- What are the intended uses of synthetic data sets?
- Do we need partial or full synthesis?
- Does sampling already provide some confidentiality?
- Challenges:
  - High-dimensional data
  - Longitudinal data
  - Logical constraints between variables
- Synthetic data set generation procedure enabling various add-ons:
  - Missing data
  - Flexible functional forms
  - Measurement error
  - Counterfactual outcomes
- Lack of neutral comparison studies

References

B. Nowok, G. M. Raab, and C. Dibben. synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11), 2016. ISSN 1548-7660. doi: 10.18637/jss.v074.i11.

D. S. Watson, K. Blesch, J. Kapar, and M. N. Wright. Adversarial random forests for density estimation and generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 5357–5375. PMLR, 2023.