Scan this QR code to connect with me on LinkedIn

The Utility Costs of Anonymization

Lisa Pilgram, MD

Postdoctoral Fellow at the Electronic Health Information Laboratory

CHEO RESEARCH INSTITUTE INSTITUT DE RECHERCHE

Agenda

Objective: Reducing Uncertainty in the Anonymization of Health Care Data

1. Problem Definition
2. Anonymization
    1. Crucial Considerations
    2. Configuration of a Clinical Case Study
3. Utility Evaluation
    1. Broad Utility and Reproducibility Metrics
    2. Results of the Clinical Case Study
4. Conclusions

CHEO RESEARCH INSTITUTE INSTITUT DE RECHERCHE

Problem Definition

Privacy Concerns are Major Barriers to Access Health Data

- Privacy is considered the most prominent issue in big data research.
    - A. Ferretti et al. “The Challenges of Big Data for Research Ethics Committees: A Qualitative Swiss Study,” J Empir Res Hum Res Ethics, vol. 17, no. 1–2, pp. 129–143, Feb. 2022, doi: 10.1177/15562646211053538
- Privacy concerns act as a barrier to sharing of health data.
    - K. B. Read et al. “Data-sharing practices in publications funded by the Canadian Institutes of Health Research: a descriptive analysis,” Canadian Medical Association Open Access Journal, vol. 9, no. 4, pp. E980–E987, Oct. 2021, doi: 10.9778/cmajo.20200303
    - R. Trestian et al., “Privacy in a Time of COVID-19: How Concerned Are You?,” IEEE Secur. Privacy, vol. 19, no. 5, pp. 26–35, Sep. 2021, doi: 10.1109/MSEC.2021.3092607
- Privacy concerns act as a barrier to seeking health care.
    - Pool J, Akhlaghpour S, Fatehi F, Gray LC. Data privacy concerns and use of telehealth in the aged care context: An integrative review and research agenda. Int J Med Inform. 2022;160:104707. doi:10.1016/j.ijmedinf.2022.104707

Exploring Privacy Concerns in Theory

1. Linking
   - Voter registration list for Cambridge Massachusetts *$20*
   - Group Insurance Commission (GIC) in Massachusetts *$0*
2. Uniqueness
   - William Weld (former governor of Massachusetts)

87% of Americans are probably unique by the combination of 5-digit zip code, sex and birth date.

L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

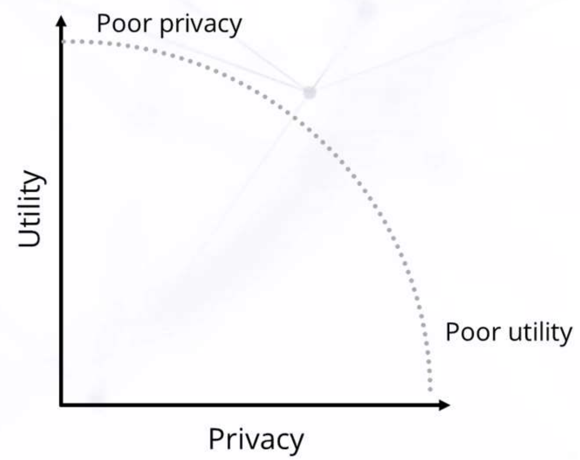
Most re-identification attacks are on improperly anonymized data.

K. El Emam et al. A systematic review of re-identification attacks on health data [published correction appears in PLoS One. 2015;10(4):e0126772]. *PLoS One*. 2011;6(12):e28071. doi:10.1371/journal.pone.0028071

Mitigating Privacy Concerns

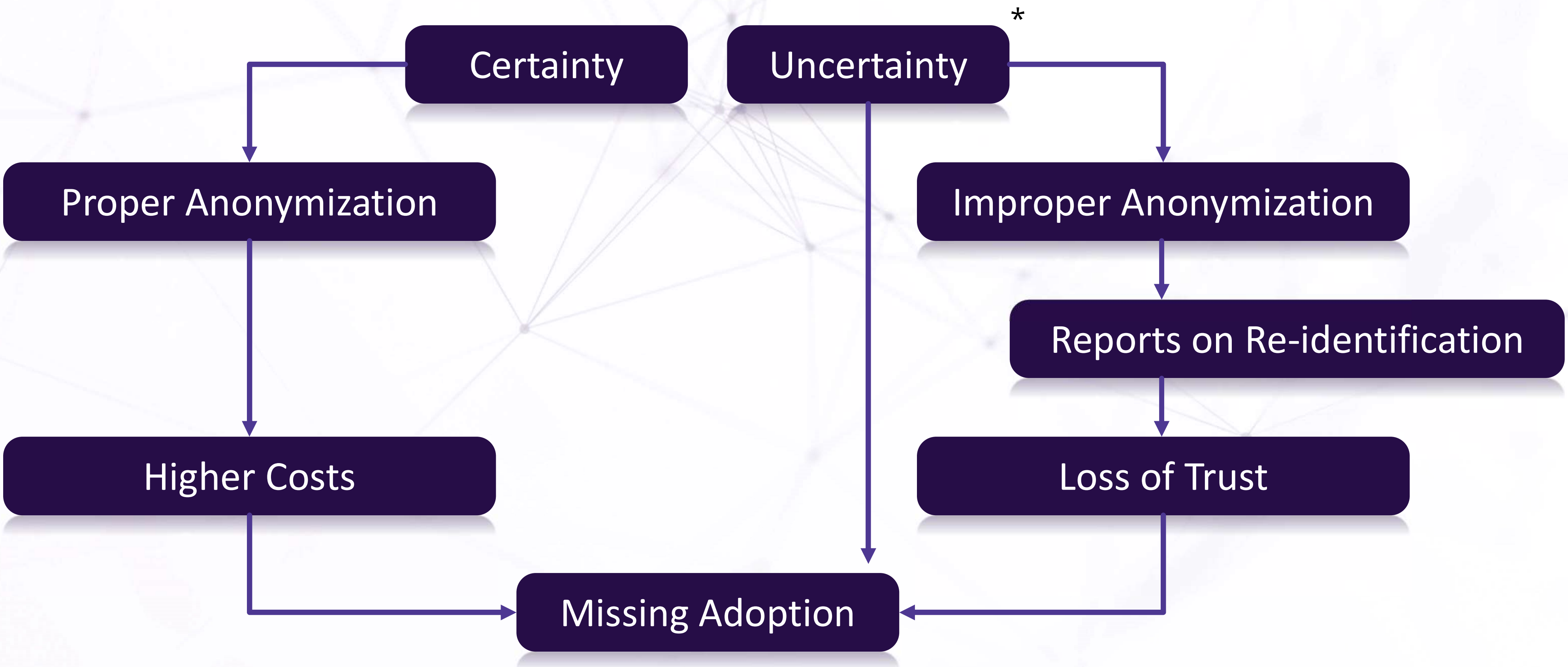
- Controlled (remote/on-site) access
- Remote execution
- Remote queries
- Secure Computation

- **Anonymization**
- Synthetic Data Generation



Electronic Health Information Laboratory, University of Ottawa and Children's Hospital of Eastern Ontario Research Institute

Missing Adoption of Anonymization



* technical but also regulatory uncertainty

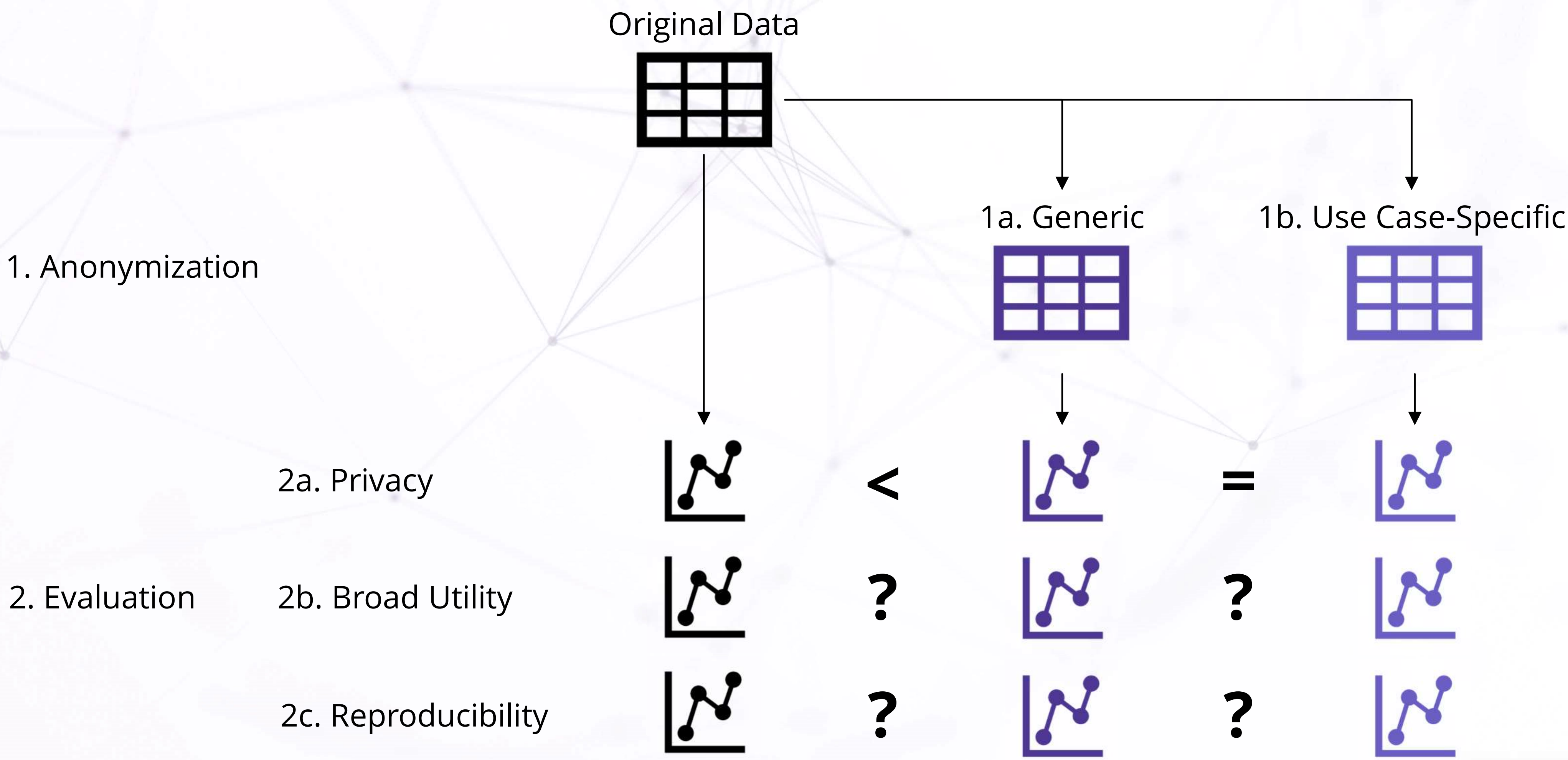
CHEO RESEARCH INSTITUTE INSTITUT DE RECHERCHE

Reducing Uncertainty in the Anonymization of Health Care Data

Research Questions

1. Can we reproduce scientific results in health research with anonymized data?
2. How relevant is use case-specific anonymization for reproducibility?
3. Do broad utility metrics reflect reproducibility?

Case Study Using Clinical Data



Anonymization

Equivalence Classes are Defined by Quasi-Identifiers

Age (years)	Gender	BMI	Pulse (bpm)	Obstructive nephropathy
63	Female	23.5	87	Yes
67	Female	30.0	65	Yes
55	Male	35.5	100	Yes
72	Female	27.8	96	No

Quasi-Identifiers (QI)

Re-identification Probability is Based on Equivalence Classes

Birth year	Gender	
1950-1960	Female	Risk: 1/1
1960-1970	Male	Risk: 1/1
1960-1970	Female	Risk: 1/1
1950-1960	Male	Risk: 1/1

Maximum Risk: 1/1 = 1.00

Average Risk: 1/4 * (1/1 + 1/1 + 1/1 + 1/1) = 1.00

Re-identification Probability is Based on Equivalence Classes

Birth year	Gender	
1950-1960	Female	Risk: 1/1
1960-1970	Female	Risk: 1/2
1960-1970	Female	Risk: 1/2
1950-1960	Male	Risk: 1/1

→

Maximum Risk: 1/1 = 1.00

Average Risk: 1/4 * (1/1 + 1/2 + 1/2 + 1/1) = 0.75

Re-identification Probability is Based on Equivalence Classes

Birth year	Gender	
1950-1960	Female	Risk: 1/2
1960-1970	Female	Risk: 1/2
1960-1970	Female	Risk: 1/2
1950-1960	Female	Risk: 1/2

k-anonymity

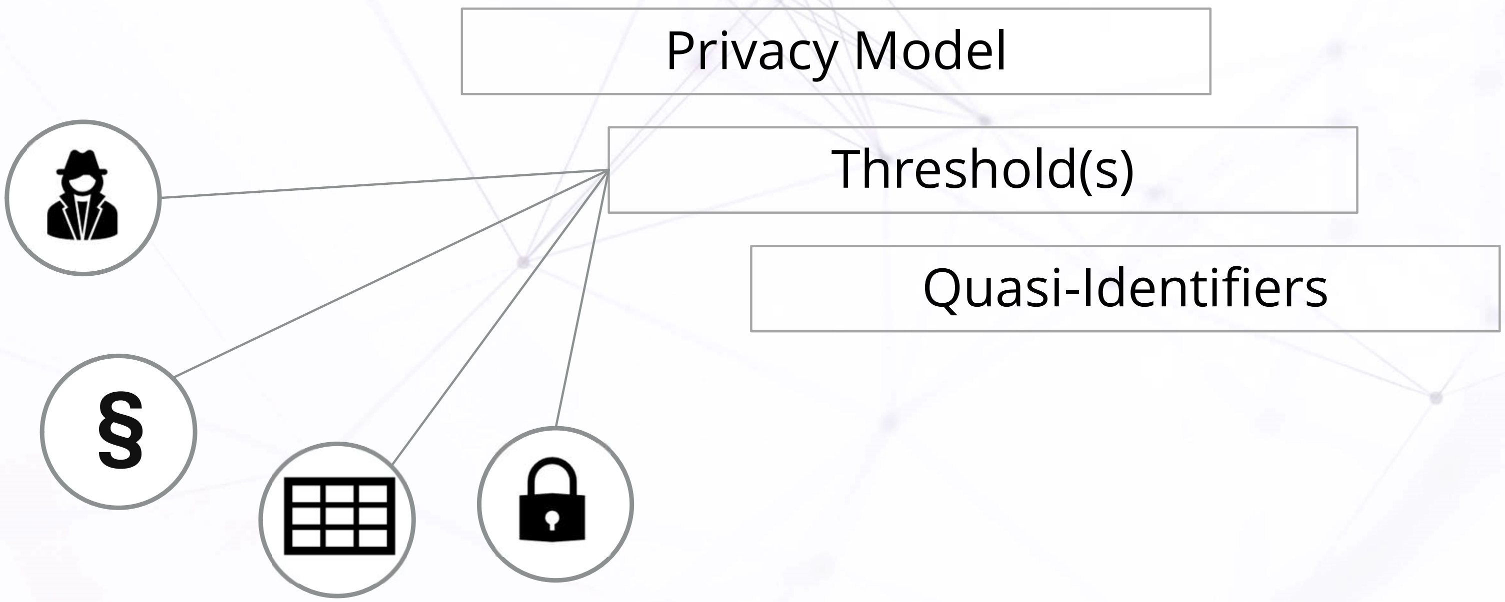
Maximum Risk: 1/2 = 0.5

Average Risk: 1/4 * (1/2 + 1/2 + 1/2 + 1/2) = 0.5

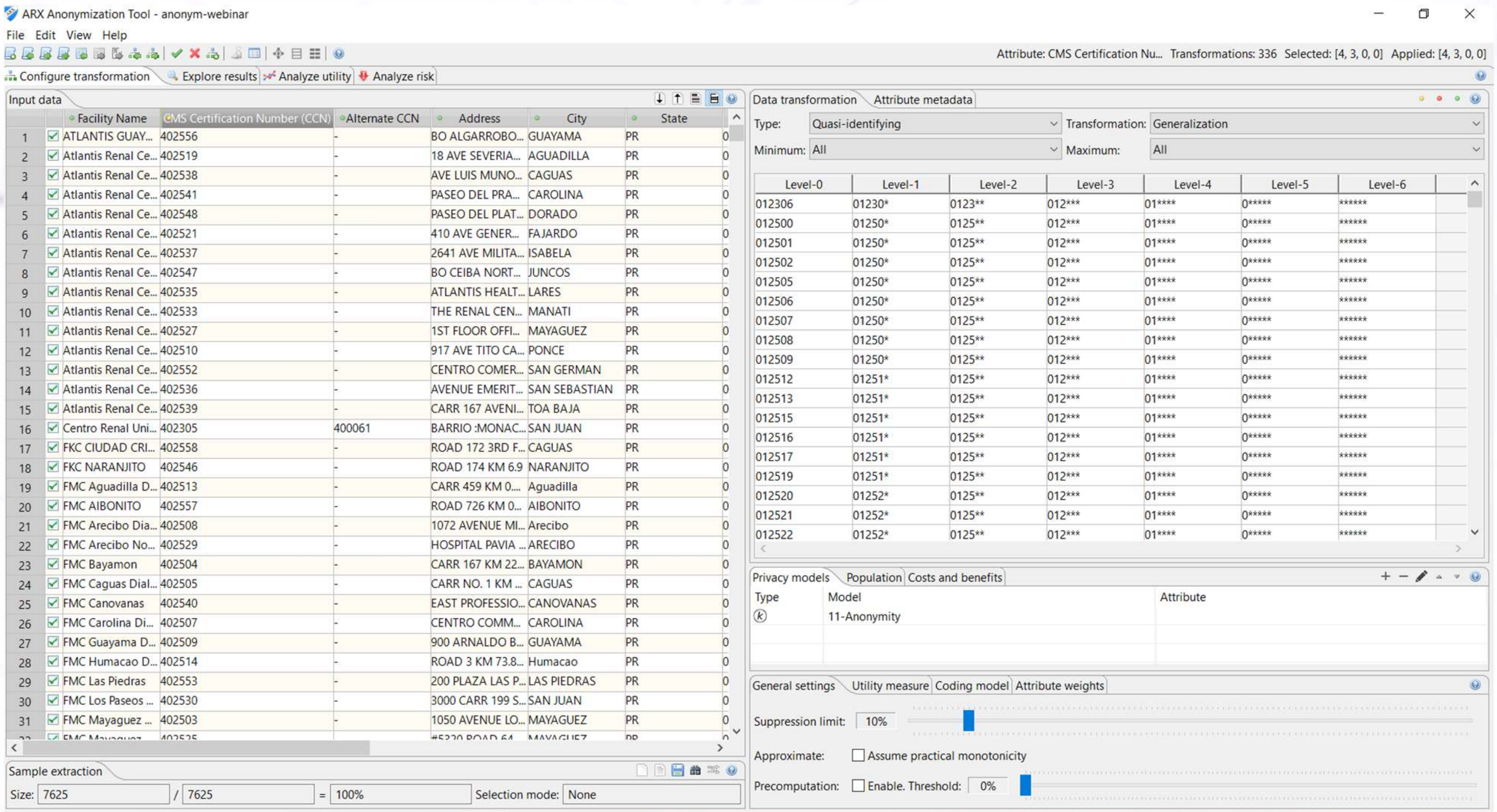
strict-average risk*

* combined with maximum risk (k-anonymity)

Threat Modeling

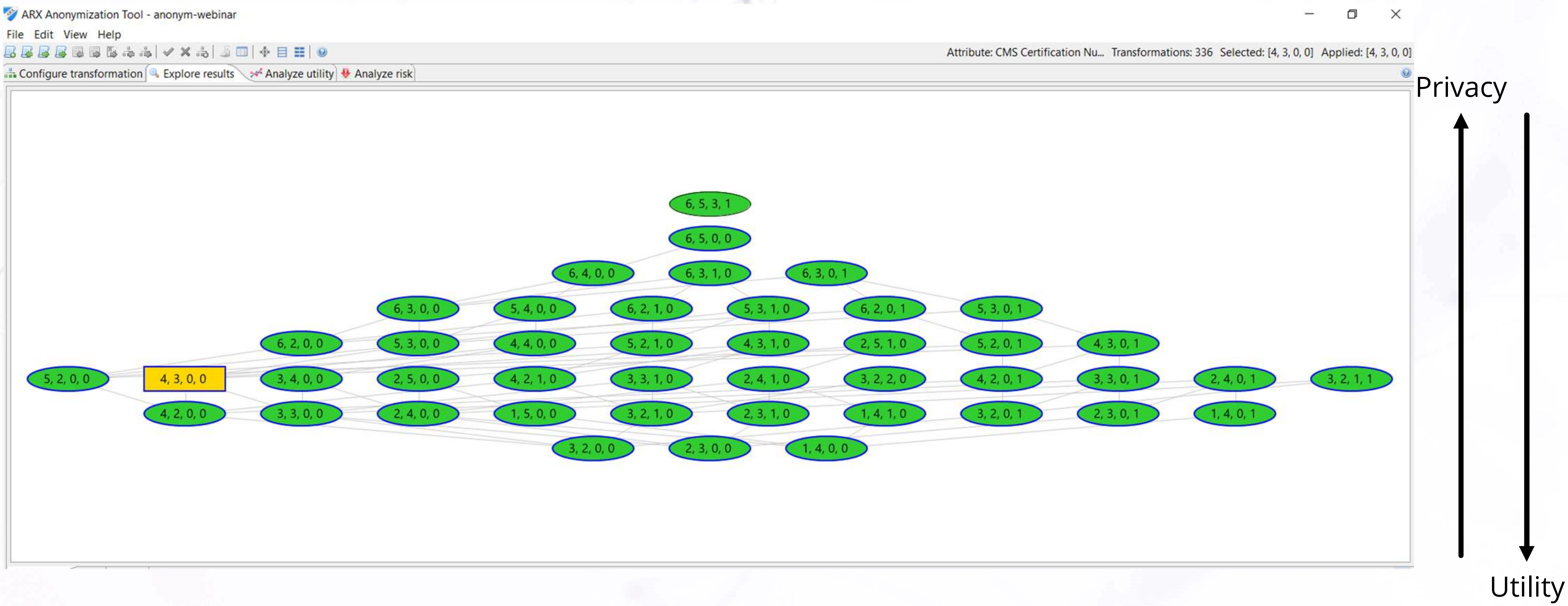


Translating Concepts Into Tools



Tool: Reference: Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX--A Comprehensive Tool for Anonymizing Biomedical Data. AMIA Annu Symp Proc. 2014;2014:984-993. Published 2014 Nov 14. https://arx.deidentifier.org/

Searching for the Optimal Solution

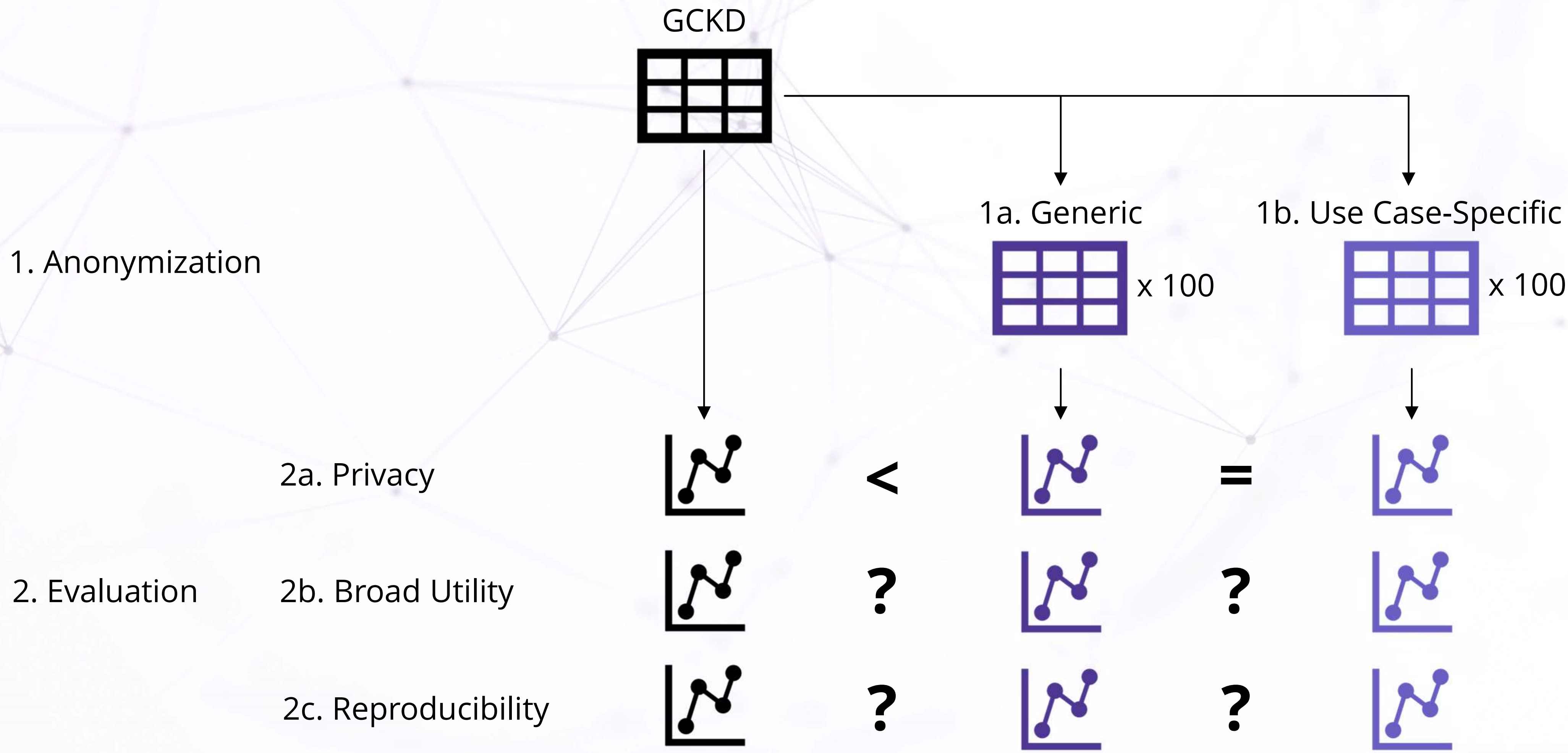


Tool: Reference: Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX--A Comprehensive Tool for Anonymizing Biomedical Data. AMIA Annu Symp Proc. 2014;2014:984-993. Published 2014 Nov 14. https://arx.deidentifier.org/

Configuring the Case Study Using Clinical Data

- Original Data: German Chronic Kidney Disease (GCKD), n = 5,217
- Anonymization: generic scenario, use case-specific scenario
- Privacy models: k-anonymity, strict-average risk
- Thresholds: k between 1 and 50
- Quasi-Identifiers: age, gender, height, weight, BMI, history of renal biopsy
- Transformation models: generalization, suppression (MaxSup: 10%)
- Reproducibility: disease burden and risk profile of patients with CKD

Case Study Using Clinical Data: 100 Study Points Per Scenario



Utility Evaluation

Measuring Utility

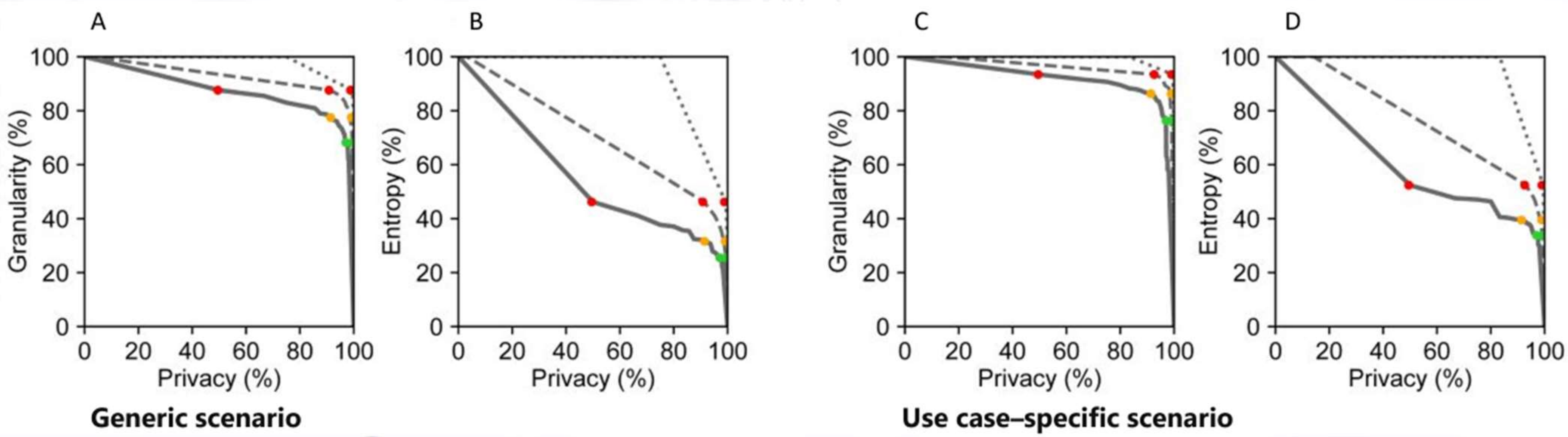
- Broad Utility
  - Granularity: coverage of the original value space
  - Entropy: differences in the distribution
- Reproducibility
  - Estimate agreement
  - 95% CI overlap

$$J_k = \frac{1}{2}\left[\frac{U_{\mathrm{over},k} - L_{\mathrm{over},k}}{U_{\mathrm{orig},k} - L_{\mathrm{orig},k}} + \frac{U_{\mathrm{over},k} - L_{\mathrm{over},k}}{U_{\mathrm{rel},k} - L_{\mathrm{rel},k}}\right]$$

From: Karr AF, et al. A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician 2006, 60:3:224-232. doi: 10.1198/000313006X124640

CHEO RESEARCH INSTITUTE INSTITUT DE RECHERCHE

Utility loss was more pronounced for entropy than for granularity.



Privacy-utility curves based on general-purpose utility metrics.
From: Pilgram et al. The Costs of Anonymization: Case Study Using Clinical Data. J Med Internet Res (forthcoming). doi:10.2196/49445
http://dx.doi.org/10.2196/49445

CHEO RESEARCH INSTITUTE INSTITUT DE RECHERCHE

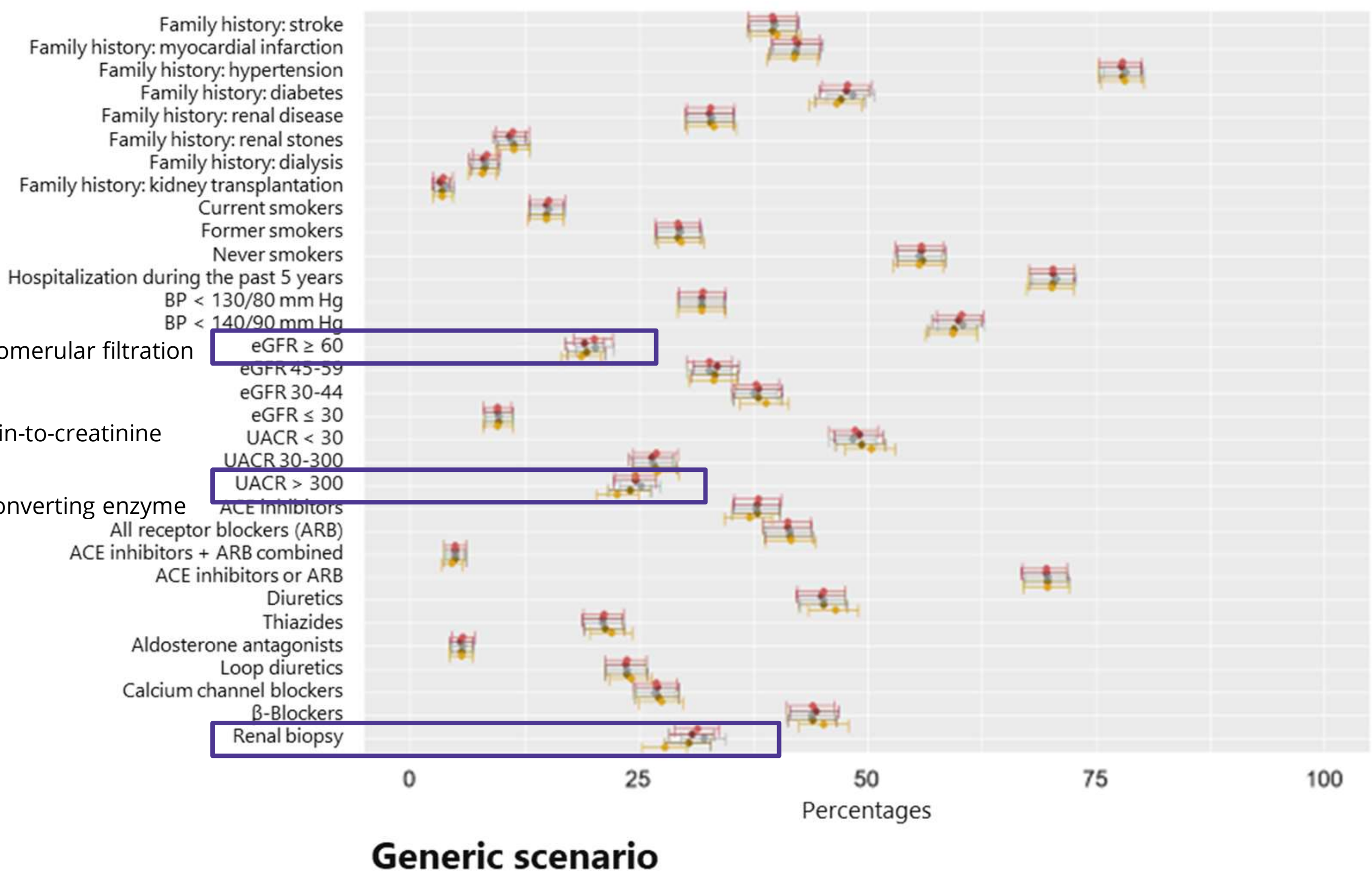
Most estimates in anonymized data had a 95% CI overlap of over 50%.

BP: blood pressure

eGFR: estimated glomerular filtration rate

UACR: urine albumin-to-creatinine ratio

ACE: angiotensin-converting enzyme

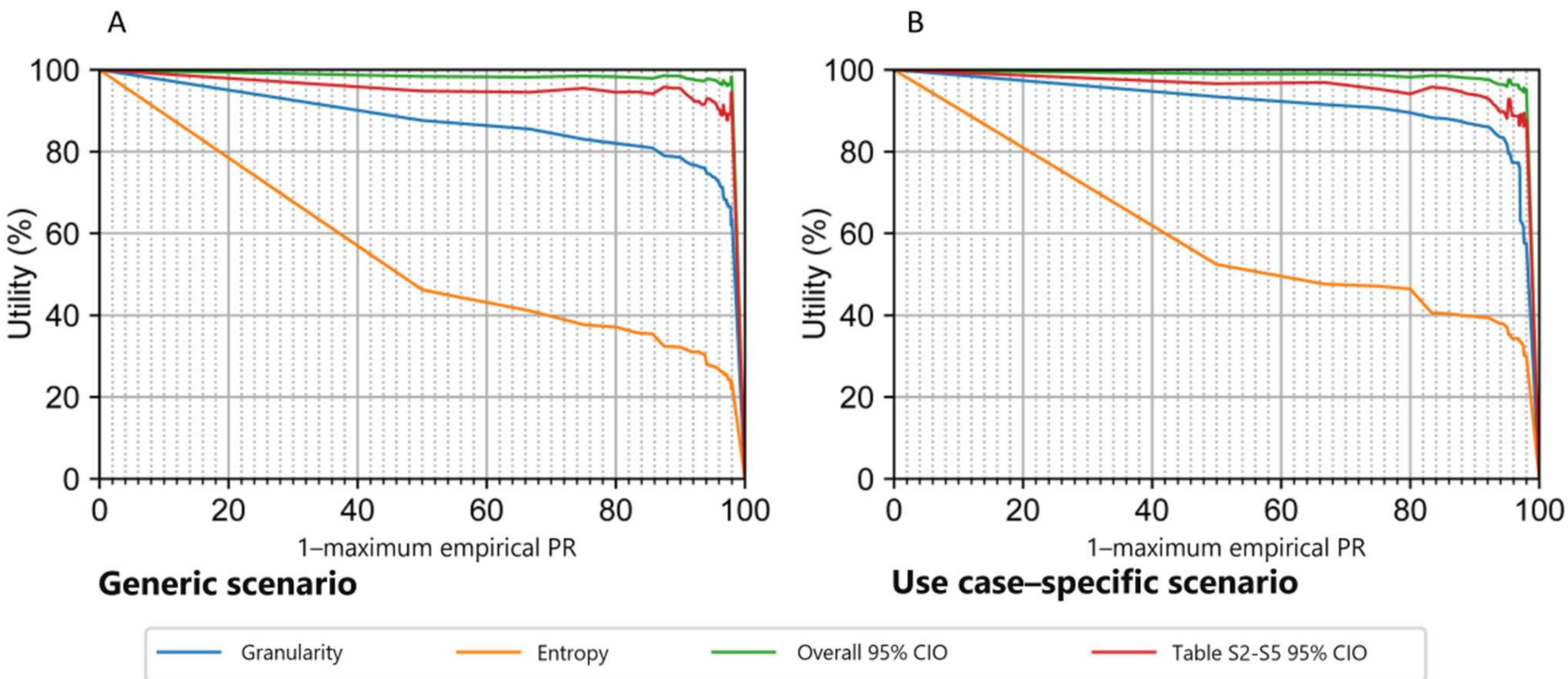


Proportion, CIs, and overlap in the interval lengths for descriptive analyses.

From: Pilgram et al. The Costs of Anonymization: Case Study Using Clinical Data. J Med Internet Res (forthcoming). doi:10.2196/49445

http://dx.doi.org/10.2196/49445

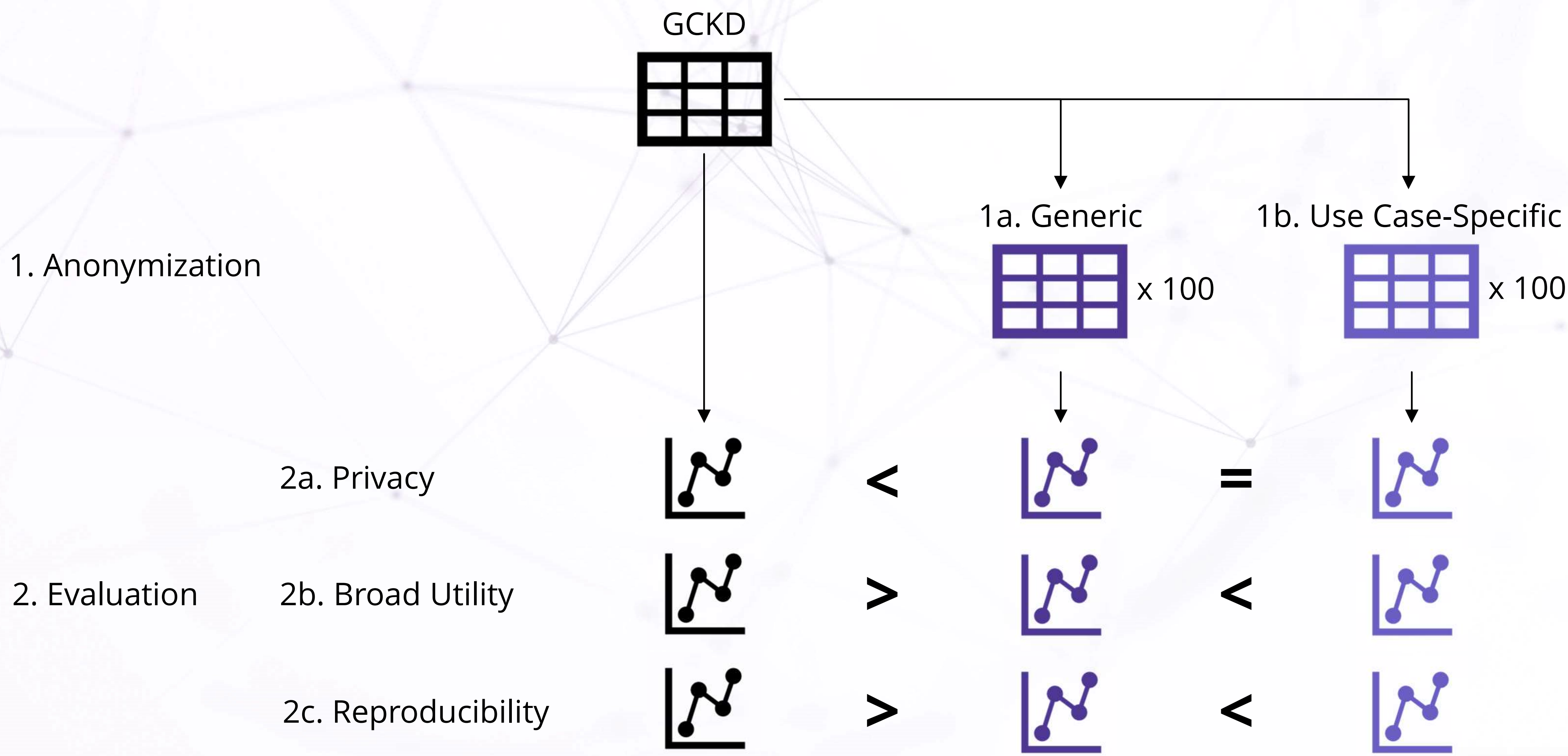
There are differences between the applied utility metrics and scenarios.



Generic and use case–specific utility metrics.
From: Pilgram et al. The Costs of Anonymization: Case Study Using Clinical Data. J Med Internet Res (forthcoming). doi:10.2196/49445
http://dx.doi.org/10.2196/49445

Conclusions

Case Study Using Clinical Data: Summary



CHEO RESEARCH INSTITUTE INSTITUT DE RECHERCHE

Research Questions: Key Findings

1. Can we reproduce scientific results in health research with anonymized data?

Yes. Anonymization of data does not necessarily impair utility for downstream analyses.

2. How relevant is use case-specific anonymization for reproducibility?

Use case-specific anonymization results in better utility for downstream analyses than generic one.

3. Do broad utility metrics reflect reproducibility?

Not necessarily. Broad utility metrics treat all variables equally. Reproducibility might be worse or better than anticipated.

Conclusions

→ Specification of utility requirements should be an integral part of the anonymization process.

→ Anonymized data for multiple likely uses should indicate limitations when implications are drawn from their analyses.

Read more in

L. Pilgram, T. Meurers, B. Malin, GCKD Investigators, E. Schaeffner, K.-U. Eckardt, F. Prasser. The Costs of Anonymization: Case Study Using Clinical Data. J Med Internet Res (forthcoming). doi:10.2196/49445
http://dx.doi.org/10.2196/49445

Special thanks to …

… the coauthors

… the GCKD investigators and study participants

… the BIH Charité Junior Digital Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin, and the Berlin Institute of Health at Charité (BIH)

Questions?