Utility-Preserving Anonymization in a Real-World Scenario: Evidence from the German Chronic Kidney Disease (GCKD) Study

Lisa PILGRAMa,b,1, Elke SCHÄFFNERc, Kai-Uwe ECKARDTad, Fabian PRASSERb, and GCKD Investigators

aDepartment of Nephrology and Medical Intensive Care, Charité - Universitätsmedizin Berlin, Berlin, Germany
bBerlin Institute of Health at Charité – Universitätsmedizin Berlin, BIH Biomedical Innovation Academy, BIH Charité Junior Digital Clinician Scientist Program, Berlin, Germany
cInstitute of Public Health, Charité – Universitätsmedizin Berlin, Berlin, Germany
dDepartment of Nephrology and Hypertension, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany

ORCiD ID: Lisa Pilgram https://orcid.org/0000-0002-1020-0650
Elke Schäffner https://orcid.org/0000-0002-7925-4577,
Kai-Uwe Eckardt https://orcid.org/0000-0003-3823-0920
Fabian Prasser https://orcid.org/0000-0003-3172-3095

Abstract. Data sharing provides benefits in terms of transparency and innovation. Privacy concerns in this context can be addressed by anonymization techniques. In our study, we evaluated anonymization approaches which transform structured data in a real-world scenario of a chronic kidney disease cohort study and checked for replicability of research results via 95% CI overlap in two differently anonymized datasets with different protection degrees. Calculated 95% CI overlapped in both applied anonymization approaches and visual comparison presented similar results. Thus, in our use case scenario, research results were not relevantly impacted by anonymization, which adds to the growing evidence of utility-preserving anonymization techniques.

Keywords. k-anonymity, privacy preserving technique, health data sharing, anonymization, de-identification

1. Introduction

The secondary use of health data requires responsible handling of personal data as defined by the European General Data Protection Regulation (GDPR). Through anonymization techniques, data can be altered in a way that it cannot be related to a person anymore. This is performed via so-called transformation models e.g. suppression

1 Corresponding Author: Lisa Pilgram, Department of Nephrology and Medical Intensive Care, Charité - Universitätsmedizin Berlin; e-mail: lisa.pilgram@charite.de.
and generalization of values. However, this manipulation can lead to utility loss, which needs to be traded off against the degree of protection achieved. Privacy models such as k-anonymity can exhibit good performance in both, privacy and general-purpose utility measures [1,2]. However, the degree to which general-purpose utility measures correlate with actual utility is less frequently reported.

In our study, we provide a comprehensive evaluation of differently anonymized datasets and evaluate their utility in a concrete application example based on the German Chronic Kidney Disease (GCKD) study.

2. Methods

2.1. Data & anonymization

With more than 10 years of data collection and over 50 peer-reviewed publication GCKD represented a granular and validated dataset with the typical diversity and complexity of a medical dataset. It was composed of 5217 records and 69 variables. We performed qualitative and semi-quantitative risk assessment as proposed by Malin et al. [3] and identified 5 variables that could be used for re-identification, so-called quasi-identifiers (QI), namely age, gender, weight, body-mass index (BMI) and renal biopsy.

Anonymization was realized via full-domain generalization and suppression using ARX with its implemented globally-optimal search algorithm [1]. We defined generalization hierarchies in alignment with the use case. We chose 11-anonymity as conservative and (11,2)-strict-average risk as moderate privacy model [4,5]. Both models put restrictions on the uniqueness of records regarding the QIs, where k-anonymity focuses on the highest uniqueness and strict-average-risk on average uniqueness. Thus, the minimum class size defined by k (11 and 2 respectively) can be translated to the maximum attacker risk (9.09% and 50.0%). In (11,2)-strict-average risk, k’ (11) describes the average group size leading to an average risk of 9.09%.

2.2. Evaluation framework

General-purpose utility was obtained using the granularity and discernibility metrics [1]. We further evaluated use case specific utility by replicating already published research results which identified the risk profile of CKD patients [6]. Exemplary analyses were performed on the original and the two anonymized datasets. We calculated proportional or mean 95% confidence intervals (CI) with the Wilson score interval or t-test respectively and reported the overlap in 95% CI lengths as measure of replicability [7]. Age, weight and BMI experienced scale transformation during the anonymization process and were compared visually.

2.3. Ethical statement

All local ethics committees approved the GCKD study (ethics committee Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, no 3831). It is registered in the national registry of clinical studies (DRKS 00003971).
3. Results

3.1. General-purpose utility

(11,2)-strict-average risk exhibited better general-purpose utility: 84.9% granularity (versus 77.3% in 11-anonymity) and 91.0% discernibility (versus 88.6% in 11-anonymity). The same tendency was observed when investigating granularity on QI-level. In QI with scale transformation during the anonymization process (age, weight, BMI), granularity was affected the most ranging between 76.7% and 85.2% in (11,2)-strict-average risk and between 63.2% and 77.6% in 11-anonymity.

3.2. Use case specific utility

We analyzed disease burden stratified by gender and the presence of diabetes mellitus. Results consist of 200 proportion and mean estimates from 40 variables (excluding variables with scale transformation) in 5 subsets. Across all 200 estimates, 95% CI lengths of the original and anonymized data overlapped. Table 1 shows the effects on selected variables of the subset of female non-diabetics. The 95% CI length overlap was lowest in renal biopsy, eGFR and systolic blood pressure. It averaged 90.4% in (11,2)-strict-average risk and 90.2% in 11-anonymity in this subset. For the overall analysis including male diabetics, male non-diabetics, female diabetics, female non-diabetics and the total cohort, the average 95% CI length overlap reached 91.1% for (11,2)-strict-average risk and 89.1% for 11-anonymity. 23.5% (47/200) of the reported estimates were not affected by anonymization at all (95% CI length overlaps 100%). These estimates were (apart from creatinine and cystatin) in the subset of the total cohort where without any stratification only renal biopsy as QI was altered during anonymization.

Table 1. Baseline characteristics of the subset of female non-diabetics within the GCKD cohort across the different datasets. Bpm: beats per minute; (S)(D)BP: (systolic)(diastolic) blood pressure; eGFR: estimated glomerular filtration rate; ARB: angiotensin receptor blockers; n: counts; SD: standard deviation; CI: confidence interval.
Serum creatinine (mg/dl)  
1.3 0.4 1.2-1.3 100.0 100.0

eGFR (ml/min)  
49.3 18.4 48.4-50.3 32.5 43.3

eGFR >= 60 ml/min  
292 20.2 18.2-22.3 52.4 55.4

eGFR 45-59 ml/min  
474 32.7 30.3-35.2 79.3 81.3

eGFR 30-44 ml/min  
544 37.5 35.1-40.1 79 81.6

eGFR < 30 ml/min  
139 9.6 8.2-11.2 93.6 96.9

Medication: ACE-inhibitors  
553 37.8 35.3-40.4 95.3 98.1

Medication ARB  
606 41.5 38.9-44.0 96.4 97.2

Medication: diuretics  
659 45.1 42.5-47.7 76.7 89.8

Medication: beta blockers  
640 43.8 41.2-46.4 78.6 78.6

Renal biopsy  
468 32.0 29.7-34.5 24.5 40.8

In a second analysis, we replicated the published age distribution of patients stratified by gender to include an example of scale transformation. Age was given in 5- or 10-years intervals in our anonymized datasets which met the originally reported distribution in 10-years intervals. As demonstrated in Figure 1, results were similar in both anonymized datasets.

![Figure 1](image.png)  
Figure 1. Age distribution of patients enrolled into the GCKD study stratified by gender. Results of the original dataset are illustrated in dark gray followed by the ones of anonymized data using 11-anonymity (middle gray) and (11,2)-strict-average risk (light gray).

4. Discussion

We detected similar tendencies among general-purpose and use case specific utility measures. In both, the moderately strong anonymization approach performed better. Proportion of renal biopsy was impacted most by anonymization in terms of lowest 95% CI length overlaps which might be due to its low case numbers in the individual strata and the relatively high application of suppression to comply to the privacy models.
However, the performance differences between the anonymization approaches seem not to relevantly affect real-world analyses in our scenario. In the exemplary analyses, 95% CI length overlapped, which we see as a confirmation of utility-preserving anonymization achieved by both approaches. The actual choice of the privacy model should consider the exact data sharing scenario (e.g. controlled vs. open access) and further safeguards implemented.

Our results are in line with other real-world applications. In the Lean European Open Survey on SARS-CoV-2 infected patients, for example, anonymization only led to deviations of not more than 0.11% in reported frequencies [8]. Song et al. could replicate prediction accuracy of a machine learning model for early acute kidney injury risk prediction in anonymized data [9]. Such more complex correlations need to be addressed more often in further research. Even though descriptive statistics as we did in our study represent an almost mandatory part of any study, anonymized data must also be able to reproduce clinical implications drawn from more sophisticated analyses.

5. Conclusion

The presented evaluation contributes to the growing evidence of successful utility-preserving anonymization techniques. This evidence is needed to address ambiguities about the trade-off between privacy and utility, and to foster adoption of anonymization into standard data management procedures.

References