*ICT for Health Science Research*
*A. Shabo (Shvo) et al. (Eds.)*

*doi:10.3233/978-1-61499-959-1-70*

# An Analysis of Erlangen University Hospital's Billing Data on Utility-Based De-Identification

Marvin O. KAMPF[a,1], Detlef KRASKA[a] and Hans-Ulrich PROKOSCH[a,b]
[a] *Center for Medical Information & Comm., Erlangen University Hospital, Germany*
[b] *Department of Medical Informatics, University of Erlangen, Germany*

**Abstract.** *Background*: To make patient care data more accessible for research, German university hospitals join forces in the course of the Medical Informatics Initiative. In a first step, the administrative data of university hospitals is made available for federated utilization. Project-specific de-identification of this data is necessary to satisfy privacy laws. *Objective*: We want to make a statement about the population uniqueness of the data. By generalizing the data, we try to reduce uniqueness and improve *k*-anonymity. *Methods*: We analyze quasi-identifying attributes of the Erlangen University Hospital's billing data regarding population uniqueness and re-identification risk. We count individuals per equality class (*k*) to measure uniqueness. *Results*: Because of the diagnoses and procedures being particularly unique in combination with sex and age of the patients, the data set is not anonymized in matters of *k*-anonymity with $k > 1$. We are able to reduce population uniqueness with generalization and suppression of unique domains. *Conclusion*: To create *k*-anonymity with $k > 1$ while still maintaining a particular utility of the data, we need to apply further established strategies of de-identification.

**Keywords.** Data privacy, *k*-anonymity, risk, secondary use, population uniqueness

## 1. Introduction

The secondary use of hospital care data for research purposes and quality assurance opens up extensive possibilities for researchers, e.g. towards feasibility studies and patient recruitment. By definition, patient consent for the secondary use of care data for research purposes is rarely available. Hence, this data may only be used anonymized, without reference to the patient's identity [1].

In the course of the Medical Informatics Initiative Germany (MI-I) a minimum set of data elements is defined [2]. This MI-I core data set is modularized and meant to be expanded over time [3,4]. Amongst others, the modules for demographics, associated diagnoses and procedures are strived to made available for federated secondary use [5], taking privacy and applicable law into account. Therefore, factual anonymization needs to be applied on the data set, while substantive utility and quality of the data needs to be maintained.

[1] Corresponding Author: Marvin Kampf, MIK, Universitätsklinikum Erlangen, Krankenhausstraße 12, 91054 Erlangen, Germany; E-Mail: marvin.kampf@uk-erlangen.de

In this paper we describe the analysis of data from the first basic modules of the MI-I core data set, containing standardized billing data from the primary patient care system of the Erlangen University Hospital as a participating site of the MIRACUM consortium [6]. This data is not only used for quality assurance and accounting but also for research purposes. To make this data accessible for explorative research requests it is pseudonymized and uploaded to the clinical data warehouse and specific research data repositories like e.g. i2b2 (Informatics for Integrating Biology & the Bedside) [7] or OHDSI (Observational Health Data Sciences and Informatics) [8] at the Erlangen data integration center.

## 2. Objectives

We examine named data set concerning basic data quality metrics and risk-based anonymizability. We use the risk model of *k*-anonymity [9,10] to make a statement about uniqueness of data records for selected combinations of quasi-identifying attributes, which is an indicator for vulnerability. We further try to improve *k*-anonymity by stepwise generalize and suppress attribute values [11] to reduce re-identification risks through Prosecutor, Journalist or Marketer attacks [12,13].

## 3. Methods

The Erlangen i2b2 instance for MIRACUM contains pseudonymized billing data in the period from 2004 to 2017. We examined the i2b2 data base design and utilized the table of patient demographics and the table of observations such as procedures and diagnoses to obtain each patient's age, sex, ICD-10-encoded diagnoses, diagnosis types and OPS-encoded procedures. We counted the number of equivalence classes for each $k$ on different combinations of these attribute fields respectively. Especially the equivalence classes for $k = 1$ were regarded as they represent an indicator for population uniqueness [14]. By transforming the data with generalization and suppression, we adjusted the counts of equivalence classes to reduce population uniqueness and improve *k*-anonymity. To go even further, we then recoded the ICD-10 and OPS codes to corresponding higher levels in the classification hierarchy.

## 4. Results

### *4.1. Equivalence classes without data transformation*

In the period from 2004 to 2017, there are 423,087 different patients in the data repository. When considering only the age of the patients, there are no equivalence classes with $k = 1$ nor with $k = 2$. Actually, even with the highest age of 117 yrs. there are three individuals (see Table 1), followed by four individuals for the ages 114 yrs. and 116 yrs. respectively.

Looking at the attributes *diagnosis* (encoded in German ICD-10 GM) and *type* of diagnosis (primary or secondary), it turns out that 3,351 diagnosis codes are unique in the data repository. With a total number of 12,652 distinct diagnoses that makes up

26.5 % unique diagnoses. The combination of the attributes *sex, age, diagnosis* and diagnosis *type* shows a number of 287,733 equivalence classes with one individual ($k = 1$). Similarly, the combination of attributes *sex* and *age* with *procedure* (encoded in OPS) shows 206,260 individuals.

**Table 1**. Number of equivalence class for selected k for various attribute sets

| Attribute Set | k (in extracts) | # of Equivalence Classes |
|---|---|---|
| {*age*} | 1;2 | 0 |
| | 3 | 1 |
| | 4 | 2 |
| {*diagnosis*,*type*} | 1 | 3,351 |
| | 2 | 1,748 |
| | 3 | 1,142 |
| {*sex*,*age*,*diagnosis*,*type*} | 1 | 287,733 |
| | 2 | 110,616 |
| | 3 | 60,353 |
| {*sex*,*age*,*procedure*} | 1 | 206,260 |
| | 2 | 67,633 |
| | 3 | 34,526 |

### *4.2. Equivalence classes with transformations*

For the most research projects it would not be viable to cut out 287,733 records from the population in order to create a *k*-anonymity with $k = 2$. For this reason, we try to reduce uniqueness throughout the records in the data repository regarding the named quasi-identifying attributes by generalization. As a first possibility, records may be grouped in age groups by ten years. This results in a *k*-anonymity of the data set with $k = 105$ when considering only the age attribute (see Table 2).

Regarding the attributes diagnosis and diagnosis type, omitting all but the first three digits of the German ICD-10 code reduces uniqueness of the records from 3,351 to 97, i.e. by a factor of 35 for $k = 1$. Applying both, age grouping and ICD-10 obfuscation improves uniqueness of the combination *sex, age, diagnosis* and diagnosis *type* from 287,733 to 5,835, i.e. by a factor of 49 for $k = 1$. Depending on the use case, we can completely omit the diagnosis type, which could improve uniqueness once more by factor 2 for $k = 1$ (from 5,835 to 2,413).

Instead of cutting the ICD-10 codes to fixed numbers of digits, we then recoded them to higher levels of the classification hierarchy. In a first step, we substituted the diagnosis code with the corresponding group (e.g. "G44.311" was recoded to "G40-G47 Episodic and paroxysmal disorders"). A comparison of Table 1 and Table 2 shows that we were able to reduce uniqueness to only 206 individual records for $k = 1$ that way. We counted even less equivalence classes after recoding the ICD-10 codes to the corresponding chapter code, the highest level in the classification (e.g. "G44.311" to "G00-G99 Chapter VI"). It is noticeable that by cutting the ICD-10 code to only one digit (e.g. "G44.311" to "G"), a *k*-anonymity with $k = 202$ can be created when considering only the attribute *diagnosis*. Although this would grant little to no utility at all for research.

The results for the attribute combination {*sex,age,procedure*} behave similar and are also listed in Table 2. The number of individuals is reduced from 206,260 to 19,413 individuals by grouping the population by age and cutting the OPS code after the first

four digits. This means a reduction of population uniqueness by a factor of about 13. When we recoded the OPS code to category level of the classification hierarchy, we were able to count only 50 unique records for $k = 1$.

**Table 2**. Number of equivalence class for selected k for various attribute sets with transformed attribute values

| Attribute Set | Transformation | k (in extracts) | # of Equivalence Classes |
|---|---|---|---|
| {*age*} | Generalization (age grouping by 10 yrs.) | <105 | 0 |
| | | 105 | 1 |
| | | 2982 | 1 |
| {*diagnosis*,*type*} | Generalization (cutting ICD-10 code after 3 digits) | 1 | 97 |
| | | 2 | 66 |
| | | 3 | 56 |
| {*sex*,*age*,*diagnosis*,*type*} | Both above-mentioned | 1 | 5,835 |
| | | 2 | 3,453 |
| {*sex*,*age*,*diagnosis*} | As above, without diagnosis type | 1 | 2,413 |
| | | 2 | 1,504 |
| {*sex*,*age*,*diagnosis*} | Recoding ICD-10 code to group level | 1 | 206 |
| | | 2 | 104 |
| {*sex*,*age*,*diagnosis*} | Recoding ICD-10 code to chapter level | 1 | 14 |
| | | 2 | 10 |
| {diagnosis} | Cutting ICD-10 code after 1 digit | <202 | 0 |
| | | 202 | 1 |
| {*sex*,*age*,*procedure*} | Recoding OPS to category level | 1 | 50 |
| | | 2 | 28 |

## 5. Discussion

The combination of quasi-identifying attributes *sex, age, diagnosis* plus diagnosis *type* as well as *procedure* has unique records in the population of the Erlangen University Hospital's research data repository. By grouping individuals in age groups of 10 yrs. and cutting ICD-10 as well as OPS codes to a fixed number of digits, we are able to reduce the population uniqueness. Despite the above-mentioned transformations, we were not able to achieve a certain level of *k*-anonymity with $k > 1$ while still maintaining a particular degree of utility of this data. The information loss when cutting the codes to a fixed number of digits will not be feasible in the most cases. The population uniqueness of OPS-301 codes in combination with age and sex is even higher.

By recoding the ICD-10 and OPS codes to higher levels in the classification hierarchy we were able reduce the unique records in the population to a minimum, while still maintaining a particular utility for research. It will be the task of the researcher to estimate the impact of information loss, depending on the particular use case and anonymization strategy. To achieve a k-anonymity with $k > 1$, further anonymization strategies have to be examined, like e.g. global/local recoding [15], anonymization based on utility constraints [16] and others. For particular research projects, only subcohorts from the total population will be examined. Thus, there is a certain probability that unique records are not part of the subcohort.

This work examined the present data set regarding *k*-anonymity and record uniqueness. The risk model of *k*-anonymity is just one of many risk models besides *l*-diversity or *t*-closeness, for example. To further assess the risks of re-identification, additional models [13] also need to be considered and are going to be applied in future research which shall then be expanded across the whole MIRACUM consortium.

## Acknowledgement

This study has been conducted within the MIRACUM consortium. MIRACUM is funded by the German Ministry for Education and Research (BMBF) under the Funding Number FKZ 01ZZ1606H. The present work was performed in fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg.

## References

[1] M. Langarizadeh, A. Orooji, and A. Sheikhtaheri, Effectiveness of Anonymization Methods in Preserving Patients ' Privacy : A Systematic Literature Review, (2018) 80–87. doi:10.3233/978-1-61499-858-7-80.

[2] S.C. Semler, F. Wissing, and R. Heyder, German Medical Informatics Initiative., *Methods Inf. Med.* **57** (2018) e50–e56. doi:10.3414/ME18-03-0003.

[3] Medical Informatics Initiative, MII Core Data Set, (2017). http://www.medizininformatik-initiative.de/en/core-data-set (accessed October 27, 2018).

[4] T. Ganslandt, M. Boeker, M. Löbe, F. Prasser, J. Schepers, S.C. Semler, S. Thun, and U. Sax, Der Kerndatensatz der Medizininformatik-Initiative: Ein Schritt zur Sekundärnutzung von Versorgungsdaten auf nationaler Ebene, *Forum Der Medizin-Dokumentation Und Medizin-Informatik*. **20** (2018) 17–21.

[5] C. Haverkamp et al., Regional Differences in Thrombectomy Rates: Secondary use of Billing Codes in the MIRACUM (Medical Informatics for Research and Care in University Medicine) Consortium, *Clin. Neuroradiol.* **28** (2018) 225–234. doi:10.1007/s00062-017-0656-y.

[6] H.-U. Prokosch, T. Acker, J. Bernarding, H. Binder, M. Boeker, M. Boerries, P. Daumke, T. Ganslandt, J. Hesser, G. Höning, M. Neumaier, K. Marquardt, H. Renz, H.-J. Rothkötter, C. Schade-Brittinger, P. Schmücker, J. Schüttler, M. Sedlmayr, H. Serve, K. Sohrabi, and H. Storf, MIRACUM: Medical Informatics in Research and Care in University Medicine, *Methods Inf. Med.* **57** (2018) e82–e91. doi:10.3414/ME17-02-0025.

[7] T. Ganslandt, S. Mate, K. Helbing, U. Sax, and H.-U. Prokosch, Unlocking Data for Clinical Research – The German i2b2 Experience, *Appl. Clin. Inform.* **2** (2011) 116–127. doi:10.4338/ACI-2010-09-CR-0051.

[8] C. Maier, L. Lang, H. Storf, P. Vormstein, R. Bieber, J. Bernarding, T. Herrmann, C. Haverkamp, P. Horki, J. Laufer, F. Berger, G. Höning, H.W. Fritsch, J. Schüttler, T. Ganslandt, H.-U. Prokosch, and M. Sedlmayr, Towards Implementation of OMOP in a German University Hospital Consortium, *Appl. Clin. Inform.* **9** (2018) 54–61. doi:10.1055/s-0037-1617452.

[9] P. Samarati, and L. Sweeney, Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion., *Proc. IEEE Symp. Res. Secur. Priv.* (1998) 384–393. doi:http://dx.doi.org/10.1145/1150402.1150499.

[10] L. Sweeney, k-Anonymity: a Model for Protecting Privacy, *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* **10** (2002) 557–570. doi:10.1142/S0218488502001648.

[11] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* **10** (2002) 571–588. doi:10.1142/S021848850200165X.

[12] K. El Emam, Risk-based de-identification of health data, *IEEE Secur. Priv.* **8** (2010) 64–67. doi:10.1109/MSP.2010.103.

[13] F. Prasser, F. Kohlmayer, and K.A. Kuhn, The importance of context: Risk-based de-identification of biomedical data, *Methods Inf. Med.* **55** (2016) 347–355. doi:10.3414/ME16-01-0012.

[14] F.K. Dankar, K. El Emam, A. Neisa, and T. Roffey, Estimating the re-identification risk of clinical data sets, *BMC Med. Inform. Decis. Mak.* **12** (2012). doi:10.1186/1472-6947-12-66.

[15] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu, Utility-based anonymization using local recoding, in: Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '06, 2006: p. 785. doi:10.1145/1150402.1150504.

[16] G. Poulis, G. Loukides, S. Skiadopoulos, and A. Gkoulalas-Divanis, Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints, *J. Biomed. Inform.* **65** (2017) 76–96. doi:10.1016/j.jbi.2016.11.001.