*German Medical Data Sciences: Shaping Change – Creative Solutions for Innovative Medicine*
*R. Röhrig et al. (Eds.)*


*doi:10.3233/SHTI190829*

# Protecting Biomedical Data Against Attribute Disclosure

Helmut SPENGLER[a,1] and Fabian PRASSER[a]
[a] *Technical University of Munich, School of Medicine, Institute of Medical Informatics, Statistics and Epidemiology, Munich, Germany*

**Abstract.** Modern medical research requires access to patient-level data of significant detail and volume. In this context, privacy concerns and legal requirements demand careful consideration. Data anonymization, which means that data is transformed to reduce privacy risks, is an important building block of data protection concepts. However, common methods of data anonymization often fail to protect data against inference of sensitive attribute values (also called attribute disclosure). Measures against such attacks have been developed, but it has been argued that they are of little practical relevance, as they involve significant data transformations which reduce output data utility to an unacceptable degree. In this article, we present an experimental study of the degree of protection and impact on data utility provided by different approaches for protecting biomedical data from attribute disclosure. We quantified the utility and privacy risks of datasets that have been protected using different anonymization methods and parameterizations. We put the results into relation with trivial baseline approaches, visualized them in the form of risk-utility curves and analyzed basic statistical properties of the sensitive attributes (e.g. the skewness of their distribution). Our results confirm that it is difficult to protect data from attribute disclosure, but they also indicate that it can be possible to achieve reasonable degrees of protection when appropriate methods are chosen based on data characteristics. While it is hard to give general recommendations, the approach presented in this article and the tools that we have used can be helpful for deciding how a given dataset can best be protected in a specific usage scenario.

**Keywords.** data protection, data anonymization, inference attacks

## 1. Introduction

To be able to develop methods for individualized prevention, diagnosis and therapy, modern medical research requires data of comprehensive breadth and depth [1]. In order to create the required big datasets, patient-level data must be re-used for secondary purposes and shared across institutional boundaries [2,3]. In this context, an important recent project is the German Medical Informatics Initiative (MII) in which four consortia work together on creating a nation-wide infrastructure for data integration and sharing [4–7]. In such projects, data protection aspects and legal requirements, e.g. specified by the European General Data Protection Regulation (GDPR) [8] or the US Health Insurance Portability and Accountability Act

[1] Corresponding Author, Helmut Spengler, Technical University of Munich, School of Medicine, Klinikum rechts der Isar, Institute of Medical Informatics, Statistics and Epidemiology, Grillparzerstr. 18, 81675 Munich, Germany; E-mail: helmut.spengler@tum.de.

(HIPAA) [9], need careful consideration. An important building block of data protection concepts is anonymization, which means that data is transformed to reduce privacy risks. This approach has, for example, been used to perform the first large-scale cross-site analysis within the context of the MII (the so called "demonstrator study", which focused on multimorbidity and rare diseases).

In the area of data anonymization, bridging the gap between legal requirements and technical solutions is challenging and subject of ongoing research [10]. One important issue is that, technically, the risk of re-identification is a continuum, while the legal perspective is dichotomous. In this article, we refer to the threats outlined in the *Opinion 05/2014 on Anonymisation Techniques* by the Article 29 Working Party [11], an independent advisory body on data protection and privacy in the European Union, which has now been replaced by the European Data Protection Board. These threats are: (1) singling out, (2) linkage and (3) inference attacks.

Anonymization methods which reduce the risk of successful linkage or singling out are widely applied in the field, but these methods are often not sufficient to prevent attackers from inferring sensitive personal information (also called *attribute disclosure*). Various models can be used to quantify the risk of attribute disclosure and to transform data to make sure that risks fall below a specified threshold [11]. However, it has been argued that such models are of little practical relevance, because implementing them requires significant data transformations which may remove an unacceptable amount of information [12].

Although the influence of anonymization on data utility has been studied extensively (see e.g. [13–17]), the literature lacks guidance on the strengths and weaknesses of different approaches for protecting biomedical data from attribute disclosure and insights into factors influencing their performance. The objective of the work described in this article was to study indicators and tools that can help to decide when and how biomedical data can be protected from sensitive attribute disclosure without compromising its usefulness too much. We focused on truthful transformation methods, which maintain the plausibility of data, as this is an important requirement in the biomedical domain [18].

## 2. Methods

### *2.1. Background*

*Singling out* means that an attacker is able to isolate some or all records which identify an individual in a dataset [11]. This threat is also mentioned as an example of re-identification in the GDPR [8]. *Linkability* denotes the ability to link two (or more) records relating to the same individual or a group of individuals, either within the same dataset or in different datasets. Technically, the attributes that can be used for such attacks are called *quasi-identifiers*. *Attribute inference* (or disclosure) [19] occurs when specific individuals can be associated with attribute values representing sensitive information (e.g. a diagnosis indicating an HIV infection). An attribute which may take sensitive values is called a *sensitive attribute*. Table 1 shows an example dataset in which the risk of singling out and successful linkage has been reduced by generalizing the attribute "Age" and removing the values of the attribute "Sex".

**Table 1.** Example dataset which has been protected against re-identification only.

| Quasi-identifiers | | | Sensitive attribute |
|---|---|---|---|
| *Age* | *Sex* | *State* | *Diagnosis* |
| [60,80[ | * | NY | Colon cancer |
| [60,80[ | * | NY | Colon cancer |
| [60,80[ | * | NY | Colon cancer |
| [60,80[ | * | NY | Breast cancer |
| [20,50[ | * | NY | Hodgkin disease |
| [20,50[ | * | NY | Breast cancer |
| [20,50[ | * | NY | Colon cancer |

This transformation increases the indistinguishability of the quasi-identifiers while at the same time reducing data utility. Despite the resulting reduction of risks of successful singling out or linkage, the first group of records is susceptible to attribute disclosure: if an adversary targets the record about a 65-year-old male living in the state of New York – which matches the first set of four records with indistinguishable quasi-identifiers – it can be inferred with high probability that the diagnosis is "Colon cancer".

*Privacy models* can be used to estimate the degree of protection of a dataset against such attacks and they can hence be used to control the risks involved with sharing or re-using data by transforming the data in such a way that a given risk threshold is met. The best-known privacy model is *k*-anonymity [20], which guarantees indistinguishability regarding the quasi-identifiers and hence prevents singling out and reduces the risk of correct linkage. Privacy models for protecting data from attribute disclosure measure risks either based on a quantification of the *diversity* of sensitive data or on a quantification of the *distances* between the distributions of sensitive data in certain groups of records and in the overall dataset. Distance-based models have been developed more recently, with the aim to overcome limitations of diversity-based models, for example with respect to skewed data [21]. Well-known examples are $\ell$-diversity [22], which is diversity-based, and *t*-closeness [21] as well as $\beta$-likeness [23], which are distance-based.

### *2.2. Experimental Design*

The aim of our work was to gain insights into how well different privacy models for protecting data against attribute disclosure are suited to balance risks against output data utility and what factors need to be considered to achieve an optimal trade-off. To this, we studied the results obtained with different privacy models in relation to trivial baseline solutions. First, we generally protected the data from singling out and linkage using the *k*-anonymity privacy model. Next, we constructed baselines for risk and utility using the following trivial protection measures:

1. Protecting the datasets against attribute disclosure by completely removing all sensitive attribute values (*Full Protection*)

2. Not protecting the dataset from attribute disclosure by keeping the sensitive attribute values as-is (*No Protection*)

Within this context, we compared the trade-off between risks and utility of datasets that have been protected against attribute disclosure using different privacy models and parameterizations.

Technically, the risk of attribute disclosure corresponds to the ability of an attacker to infer values of the sensitive attribute from values of the quasi-identifiers. We modeled this by using logistic regression classifiers that have been trained on anonymized data to predict the values of the sensitive attribute using the quasi-identifiers as features [24]. We quantified the risk by determining the prediction accuracy using 3-fold cross-validation. For measuring the utility of output data, we used a general-purpose model that captures the granularity of output data [25].

All experiments were performed using the open source data anonymization tool ARX, which we configured to use local generalization [26], which is a truthful transformation method. Protection against singling out and linkage has been implemented using *5*-anonymity [20]. For additionally protecting the datasets against attribute disclosure, we used the following privacy models: distinct-$\ell$-diversity [22], *t*-closeness using the earth-movers distance based on value generalization hierarchies [21] and enhanced $\beta$-likeness [23]. For each model, we selected a representative set of parameterizations covering the complete spectrum of reasonable values. The source code of our experiments is available online [27].

**Table 2.** Statistical properties of the sensitive attributes in the evaluation datasets.

| Dataset | Sensitive attribute | Domain size | Min. frequency | Max. frequency | Dispersion index |
|---|---|---|---|---|---|
| Census | Marital status | 5 | 0.014 | 0.446 | 0.777 |
| Health interviews | Marital status | 10 | $1.3 \cdot 10^{-5}$ | 0.236 | 0.899 |
| Census | Education | 25 | 0.010 | 0.176 | 0.943 |
| Health interviews | Education | 26 | 0.001 | 0.192 | 0.952 |

To be able to study influencing factors with respect to data characteristics, we selected datasets with a similar schema but different statistical properties. The *Health interviews*[2] dataset consisted of 1,193,645 records from the U.S. National Health Interview Survey (NHIS). The *Census*[3] dataset contained 68,725 responses to the American Community Survey (ACS) from randomly selected people living in the state of Massachusetts in the U.S. We selected *Sex*, *Age* and *Race* as quasi-identifiers, since demographic parameters are typically considered to be associated with a high risk of re-identification [28] and to provide comparability with prior studies [16,20,25]. As examples for sensitive attributes we selected *Marital status* as well as *Education* and protected them from inference attacks.

Table 2 provides an overview of the statistical properties of the sensitive attributes in the evaluation datasets. We used the *dispersion index* [29], as an indicator for the

[2] https://nhis.ipums.org/nhis/

[3] http://www.census.gov/programs-surveys/acs/data/pums.html

skewness of the distribution of attribute values (*skewness of the sensitive data*, henceforth). A value of 1 represents minimum skewness with a uniform distribution of two or more distinct attribute values. A value of 0 represents maximum skewness where there is no distribution of values but only *one* uniform value.

## 3. Results

To visualize the results of our experiments, we computed risk-utility curves that summarize the results obtained for the different privacy models and parameterizations into one diagram per dataset and sensitive attribute. Each data point represents an anonymized output dataset for a specific model and parameter value. The x- and y-coordinates of the data points represent the according risk and utility values, which have been normalized in relation to the baseline approaches *"Full Protection"* (lowest risk, lowest utility) and "*No Protection"* (highest risk, highest utility). Table 3 contains an overview of the values obtained for the baseline approaches.

**Table 3.** Baseline values for the evaluation datasets.

| Dataset | Sensitive attribute | *Full Protection* | | *No Protection* | |
|---|---|---|---|---|---|
| | | Risk | Utility | Risk | Utility |
| Census | Marital status | 0.0 | 0.7481 | 0.7495 | 0.9981 |
| Health interviews | Marital status | 0.0 | 0.7498 | 0.6108 | 0.9998 |
| Census | Education | 0.0 | 0.7481 | 0.3609 | 0.9981 |
| Health interviews | Education | 0.0 | 0.7498 | 0.3954 | 0.9998 |

The risk-utility curves are shown in Figure 1. The results have been sorted such that the experiment with the highest skewness of sensitive data (*Census / Marital status*) is on the left and the experiment with the lowest skewness of sensitive data (*Health interviews /Education*) is on the right.

A perfect risk utility curve would contain a data point with zero risk and 100% utility. The line between (0,0) and (1,1) represents all solutions where reductions in risk are directly proportional to reductions in utility. The further a solution is above this line, the higher the increase in protection relative to the reduction in utility. Solutions below the line can be considered suboptimal, as they represent trade-offs that are worse than those provided by the baseline approaches. However, it must be noted that these solutions are not necessarily worthless, particularly in scenarios where neither of the two baseline approaches can be used.

As can be seen, the privacy models studied in this article enable a balancing of risks and utility within the spectrum of removing all sensitive data (0,0) and not protecting a dataset at all (1,1). However, when protecting datasets with skewed sensitive attributes (*Census / Marital status* and *Health interviews / Marital status*) no solutions could be found that provide a better risk-utility trade-off than the baseline approaches.

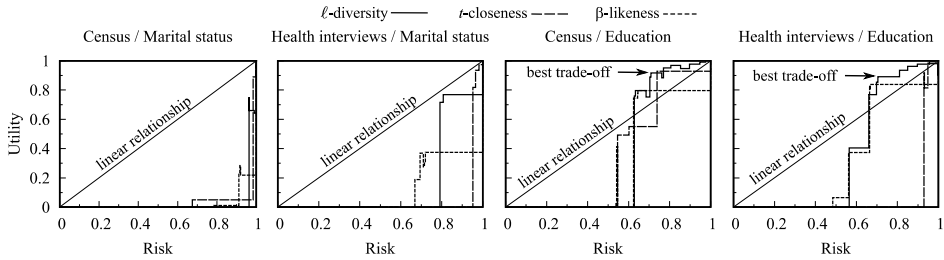


**Figure 1.** Risk-utility curves for the privacy models $\ell$-diversity, $t$-closeness, and $\beta$-likeness. The solution with the best risk-utility trade-off above the linear relationship is marked with an arrow.

For datasets with low skewness of the sensitive data (*Census / Education* and *Health interviews / Education*), particularly when using $\ell$-diversity and $\beta$-likeness, there were quite a few solutions that provided significant reductions in risk without affecting data utility too much. Using more sophisticated distance-based models did not lead to better results compared to outputs produced using the very basic $\ell$-diversity model, which simply restricts the number of distinct sensitive attribute values per group of indistinguishable records [22]. This was true even for data sets with highly skewed sensitive data. In both experiments with the *Health interviews* datasets, $t$-closeness provided the worst trade-offs of all models.

## 4. Discussion

Our work is not the first to study the risk-utility trade-off provided by methods for protecting data against attribute disclosure. Notably, Brickell and Shmatikov thoroughly studied the effects of implementing common privacy models on the performance of statistical classification models trained on output data [16]. However, as Li and Li have pointed out, the methodology used in [16] is unsound [30]. Moreover, when studying the effect of methods for protecting data from attribute disclosure, statistical classification is typically used for measuring data utility while we used it to estimate the risk of sensitive attribute inference. Furthermore, we also considered the properties of the distribution of sensitive attribute values. Finally, we have put a specific emphasis on truthful data transformation methods which have been recommended for biomedical data [31,32].

Our results indicate that it is indeed hard to achieve a reasonable trade-off for skewed data, even with distance-based models that have been developed specifically for this purpose [21,23]. When data is only moderately skewed, both $\beta$-likeness and $\ell$-diversity can yield significantly better risk-utility trade-offs than the baseline approaches. Interestingly, using $\ell$-diversity, which is the simplest model with the most intuitive semantics, often provided better results than more recent and more sophisticated models like $\beta$-likeness.

In summary we conclude that – contrary to popular opinion – it can be possible to significantly reduce the risk of successful inference attacks using well-known and truthful anonymization methods. However, it is important to consider the specifics of the dataset that is to be protected and the context of data usage. For this purpose, an in-

depth analysis must be performed, and the approach used in this paper can serve as a blueprint. To calculate the necessary statistics and analyze risk-utility trade-offs, common anonymization tools such as ARX [26] and sdcMicro [33] can be used. The source code of our analysis is available online [27]. However, in future work it should be considered to extend the tools mentioned with additional methods to more directly support the processes described (e.g. through the graphical user interface). Our results indicate that particular attention should be directed towards simple methods with intuitive semantics, such as $\ell$-diversity, and weak parameterizations, such as $\ell \leq 5$.

One of the main limitations of the present work is the use of specific methods for quantifying risks and data utility. In fact, different measurement methods may be relevant in different data use scenarios. This is particularly true for methods that quantify data utility.

## References


[1] T.B. Murdoch, and A.S. Detsky, The Inevitable Application of Big Data to Health Care, *J. Am. Med. Assoc.* **309** (2013) 1351. doi:10.1001/jama.2013.393.

[2] J. Christoph, L. Griebel, I. Leb, I. Engel, F. Köpcke, D. Toddenroth, H.-U. Prokosch, J. Laufer, K. Marquardt, and M. Sedlmayr, Secure Secondary Use of Clinical Data with Cloud-based NLP Services, *Methods Inf. Med.* **54** (2015) 276–282. doi:10.3414/ME13-01-0133.

[3] S. Schneeweiss, Learning from Big Health Care Data, *N Engl J Med*. **370** (2014) 2161–2163. doi:10.1056/NEJMp1401111.

[4] F. Prasser, O. Kohlbacher, U. Mansmann, B. Bauer, and K. Kuhn, Data Integration for Future Medicine (DIFUTURE), *Methods Inf. Med.* **57** (2018) e57–e65. doi:10.3414/ME17-02-0022.

[5] B. Haarbrandt, B. Schreiweis, S. Rey, U. Sax, S. Scheithauer, O. Rienhoff, P. Knaup-Gregori, U. Bavendiek, C. Dieterich, B. Brors, and others, HiGHmed – An Open Platform Approach to Enhance Care and Research across Institutional Boundaries, *Methods Inf. Med.* **57** (2018) e66–e81. doi:10.3414/ME18-02-0002.

[6] H.-U. Prokosch, T. Acker, J. Bernarding, H. Binder, M. Boeker, M. Boerries, P. Daumke, T. Ganslandt, J. Hesser, G. Höning, and others, MIRACUM: Medical Informatics in Research and Care in University Medicine, *Methods Inf. Med.* **57** (2018) e82–e91. doi:10.3414/ME17-02-0025.

[7] A. Winter, S. Stäubert, D. Ammon, S. Aiche, O. Beyan, V. Bischoff, P. Daumke, S. Decker, G. Funkat, J.E. Gewehr, and others, Smart Medical Information Technology for Healthcare (SMITH), *Methods Inf. Med.* **57** (2018) e92–e105. doi:10.3414/ME18-02-0004.

[8] The European Parliament and the Council of the European Union. Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), *Off J Eur Commun*. **L 119** (2016) 1–88.

[9] U.S. Department of Health and Human Services, Office for Civil Rights. HIPAA administrative simplification regulation, *45 CFR Parts 160, 162, 164*. (2013).

[10] M. Karg, Anonymität, Pseudonyme und Personenbezug revisited?, *Datenschutz Und Datensicherheit - DuD*. (2015). doi:10.1007/s11623-015-0463-z.

[11] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, (2014). http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed May 13, 2018).

[12] K. El Emam, and C. Álvarez, A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques, *Int. Data Priv. Law*. **5** (2015) 73–87. doi:10.1093/idpl/ipu033.

[13] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, A systematic review of re-identification attacks on health data, *PLoS One*. **6** (2011).

[14] A. Gionis, and T. Tassa, Using k-anonymization with minimal loss of information, *IEEE Trans. Knowl. Data Eng.* **21** (2009) 206–219. doi:10.1109/TKDE.2008.129.

[15] B.C.M. Fung, K. Wang, A.W.-C. Fu, and P.S. Yu, Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, 2010.

[16] J. Brickell, and V. Shmatikov, The cost of privacy: destruction of data-mining utility in anonymized data publishing, in: 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2008: pp. 70–78. doi:http://doi.acm.org/10.1145/1401890.1401904.

[17] F. Prasser, F. Kohlmayer, and K.A. Kuhn, Efficient and effective pruning strategies for health data de-

identification, *BMC Med. Inform. Decis. Mak.* (2016). doi:10.1186/s12911-016-0287-2.
[18] H. Lee, S. Kim, J.W. Kim, and Y.D. Chung, Utility-preserving anonymization for health data publishing, *BMC Med. Inform. Decis. Mak.* **17** (2017) 104.
[19] D. Lambert, Measures of disclosure risk and harm, *J. Off. Stat.* **9** (1993) 313–331.
[20] 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.
[21] N. Li, T. Li, and S. Venkatasubramanian, t-Closeness: Privacy beyond k-anonymity and l-diversity, in: 23rd IEEE Int. Conf. Data Eng., 2007: pp. 106–115. doi:10.1109/ICDE.2007.367856.
[22] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, *ACM Trans. Knowl. Discov. Data*. **1** (2007) 3. doi:10.1145/1217299.1217302.
[23] J. Cao, and P. Karras, Publishing Microdata with a Robust Privacy Guarantee, *Proc. VLDB Endow.* **5** (2012) 1388–1399. doi:10.14778/2350229.2350255.
[24] F. Prasser, J. Eicher, R. Bild, H. Spengler, and K.A. Kuhn, A tool for optimizing de-identified health data for use in statistical classification, in: 2017 IEEE 30th Int. Symp. Comput. Med. Syst., 2017: pp. 169–174.
[25] V.S. Iyengar, Transforming data to satisfy privacy constraints, in: 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2002: pp. 279–288. doi:10.1145/775047.775089.
[26] F. Prasser, F. Kohlmayer, R. Lautenschläger, and K.A. Kuhn, ARX - A Comprehensive Tool for Anonymizing Biomedical Data, in: AMIA Annu. Symp. Proc., 2014: pp. 984–993.
[27] arx-deidentifier/attribute-disclosure-benchmark, (2019). https://github.com/arx-deidentifier/attribute-disclosure-benchmark (accessed February 25, 2019).
[28] L. Sweeney, Simple demographics often identify people uniquely, *Carnegie Mellon Univ. Data Priv. Work. Pap. 3*. **671** (2000) 1–34.
[29] W.D. Schafer, Assessment of dispersion in categorical data, *Educ. Psychol. Meas.* **40** (1980) 879–883.
[30] T. Li, and N. Li, On the tradeoff between privacy and utility in data publishing, in: 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2009: pp. 517–526. doi:10.1145/1557019.1557079.
[31] K. El Emam, and L. Arbuckle, Anonymizing health data: case studies and methods to get you started, O'Reilly, 2013.
[32] K. El Emam, and B. Malin, Appendix B: Concepts and methods for de-identifying clinical trial data, in: Comm. Strateg. Responsible Shar. Clin. Trial Data; Board Heal. Sci. Policy; Inst. Med. Ed. Shar. Clin. Trial Data Maximizing Benefits, Minimizing Risk, 2015: pp. 1–290.
[33] M. Templ, A. Kowarik, and B. Meindl, Statistical disclosure control for microdata using the R-package sdcMicro, *Trans. Data Priv.* **1** (2008) 67–85. doi:10.18637/jss.v067.i04.