Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-312

A Generic Method for Assessing the Quality of De-Identified Health Data

Fabian PRASSER[1], Raffael BILD, Klaus A. KUHN
Chair of Biomedical Informatics, Institute of Medical Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Germany

Abstract. Data sharing plays an important role in modern biomedical research. Due to the inherent sensitivity of health data, patient privacy must be protected. De-identification means to transform a dataset in such a way that it becomes extremely difficult for an attacker to link its records to identified individuals. This can be achieved with different types of data transformations. As transformation impacts the information content of a dataset, it is important to balance an increase in privacy with a decrease in data quality. To this end, models for measuring both aspects are needed. Non-Uniform Entropy is a model for data quality which is frequently recommended for de-identifying health data. In this work we show that it cannot be used in a meaningful way for measuring the quality of data which has been transformed with several important types of data transformation. We introduce a generic variant, which overcomes this limitation. We performed experiments with real-world datasets, which show that our method provides a unified framework in which the quality of differently transformed data can be compared to find a good or even optimal solution to a given data de-identification problem. We have implemented our method into ARX, an open source anonymization tool for biomedical data.

Keywords. security, privacy, biomedical data, de-identification, data quality

1. Introduction

Data sharing plays an important role in modern biomedical research, for example in efforts towards precision medicine [1]. Due to the inherent sensitivity of health data, patient privacy must be protected. t has been recommended to employ organizational and legal safeguards, such as data use agreements and data access committees, and to inform data subjects about risks of data sharing already in the informed consent [2]. Multiple layers of access to sensitive data should be used to create controlled environments in which it becomes possible to reason about privacy risks and to manage them with data *de-identification* [3].

De-identification means to transform data in such a way that it becomes extremly difficult for an attacker to link the dataset to identified or identifiable individuals [3]. This can be achieved in a variety of ways. There are different transformation models, such as attribute generalization or suppression, and even these individual types of transformation can be applied in different ways. As different transformations have

[1] Corresponding Author: Fabian Prasser, Chair of Biomedical Informatics, Institute of Medical Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany; E-mail: fabian.prasser@tum.de.

different impact on the information content of a dataset, it is important to balance an increase in privacy with a decrease in data quality. To this end, models for measuring both aspects are needed. In this article, we focus on data quality.

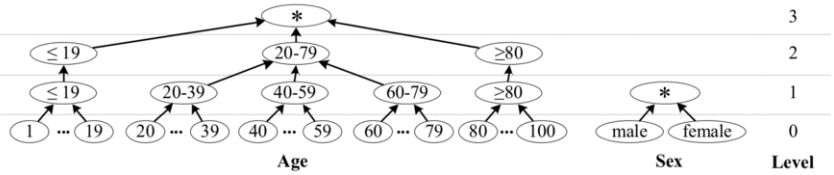

Figure 1. Generalization hierarchies for attributes *age* and *sex*.

2. Background

Health data is typically de-identified with user-defined generalization hierarchies [3, 4]. Here, the precision of values of attributes which are associated with high risks of re-identification is iteratively reduced. As a consequence, privacy risks decrease. Two simple examples are shown in Figure 1. Each hierarchy contains a set of increasing *levels*, which specify values with increasing coverage of the attribute's domain.

	Age	Sex
0	20	Male
1	65	Male
2	55	Male
3	40	Female
4	65	Female
5	65	Female

Input dataset (D)

Age	Sex
20-39	Male
60-79	Male
40-59	Male
40-59	Female
60-79	Female
60-79	Female

Full-domain generalization (D_g)

Age	Sex
20-79	Male
20-79	Male
40-59	*
40-59	*
65	Female
65	Female

Local recoding (D_l)

Figure 2. Example dataset and two generalizations with global and local recoding.

Different methods exist for transforming data with generalization. An example is illustrated in Figure 2, which uses the hierarchies from Figure 1. *Full-domain generalization* means that the same level of generalization is used to transform all values of an attribute. For example, D_g can be derived from D by transforming the first attribute (*age*) to level 1 and the second attribute (*sex*) to level 0 of the associated hierarchies. *Local recoding* means that identical values of an attribute can be generalized to different levels in different records. For example, in D_l values of the attribute *age* have been transformed to level 2 in the first two records, to level 1 in the next two records and to level 0 in the last two records. We note that with local recoding, value and record *suppression* (i.e. removal) can also be modelled: by considering a suppressed value to be generalized to the root node of the associated hierarchy. In the example, the attribute *sex* has been suppressed in records 2 and 3 of the dataset D_l.

Entropy is a well-known measure for information content. In the context of data de-identification, it was originally introduced as a model for measuring the loss of information by De Waal and Willenborg [5]. Gionis and Tassa introduced a slight variation, called *Non-Uniform Entropy*, which in contrast to the original proposal increases monotonically with increasing degrees of generalization [6]. Non-Uniform Entropy is frequently used in scientific works, e.g. [4] and [7], and it has been recommended as a quality model for health data de-identification [3].

3. Objective

Different transformation models have different advantages and drawbacks. Attribute generalization has been recommended for de-identifying health data, because it is intuitive and truthful, i.e. non-pertubative [3]. Full-domain generalization has been recommended because it results in datasets which are easy to analyze, as all values have been transformed to the same generalization level [4]. However, full-domain generalization is not very flexible and, depending on the distribution of the data, more information may be removed than required [8]. With local recoding, transformations are also truthful, although the results may be difficult to analyze [9].

Already this short discussion shows that the suitability of different transformation methods depends on the use case, for example, whether a dataset will be analyzed by epidemiologists or used for machine learning. In ARX, which is an open source anonymization tool for biomedical data [10], we have therefore implemented all of the methods described previously. As we have also explained already, users as well as de-identification algorithms need to assess the quality of de-identified data. For this purpose, we have implemented Non-Uniform Entropy. However, we will show in this article that Non-Uniform Entropy is not suited well for evaluating the quality of locally recoded data. As the model is frequently recommended for biomedical data, we have developed a generic variant which can be used to assess the information loss induced by transforming data with arbitrary combinations of full-domain generalization, local recoding and record or value suppression.

4. Methods

The Non-Uniform Entropy of a transformed dataset is defined as the sum of the Non-Uniform Entropy of each column. Without loss of generality we will therefore focus on datasets with a single attribute in the remainder of this article. An example using the attribute *age* of the datasets from Figure 2 is shown in Figure 3.

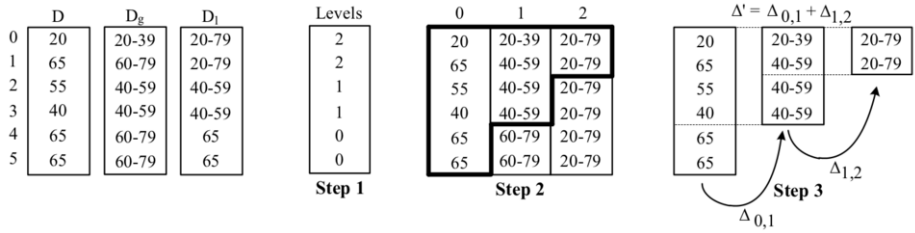

Figure 3. Attribute *age* of D, D_g and D_l as well as an evaluation of Non-Uniform Entropy for D_l.

The basic idea of Non-Uniform Entropy is to compare the frequencies of attribute values in the transformed dataset with the according frequencies in the input dataset. The function *f(D, x)* returns the frequency of the value of row x in dataset D. For example, *f(D, 2) = 1* as one record of D has a value of *"55"* and $f(D_g, 2) = 2$ as two records of D_g have a value of *"40-59"*. When a dataset D is transformed to another dataset D', loss of information is defined as:

$$\Delta(D, D') = \sum_{x \in D} -log \frac{f(D, x)}{f(D', x)}$$ [6].

In the formula it is assumed that the quotient is always ≤ 1 as the frequency of attribute values can only increase with full-domain generalization. Consequently, the negative logarithm of the quotient is always a positive number and the sum of all of this numbers defines the overall loss of information. However, already our simple example from Figure 3 shows that this does not work well with local recoding, for example when transforming D to D_l. Let us consider the value in row 4, which is *"65"*. The frequency of this value in D is three, but the frequency of the value in D_l is two. Hence, the value of the quotient becomes larger than one, and the negative logarithm of the quotient, becomes negative. Non-Uniform Entropy therefore measures an *information gain* for this row. To overcome this limitation, we employ a three-step process, as is shown in Figure 3. **First**, we calculate the generalization level for each record by matching the values of the input dataset against the generalization hierarchies. **Second**, for each generalization level n used in the dataset, we determine the set of all records that are *affected by n*, which means that they are generalized to a level higher than or equal to n. For example, in Figure 3 the records 0, 1, 2 and 3 of D_l are affected by generalization level 1. As can also be seen in the figure, the set of records affected by a generalization level n is always a subset of or equal to the set of records affected by any other generalization level $n' < n$. This follows from the fact that the data is transformed with hierarchies. **Third**, we iterate over each generalization level $n > 0$ and calculate the information loss according to Non-Uniform Entropy for transforming the set of records affected by generalization level n from level $n - 1$ to level n. The information loss for the whole dataset is defined as the sum of these individual losses. We note that by focussing on the records affected by a given level, the frequencies of values in the same cells may be different in different summands. In our example, the frequency of the value *"40-59"* is 3 when calculating $\Delta_{0,1}$ and 1 when calculating $\Delta_{1,2}$.

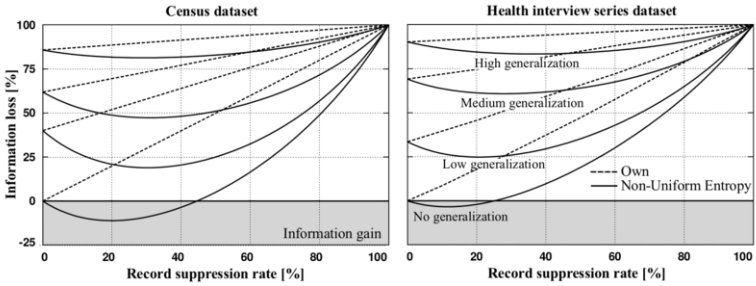

Figure 4. Information loss calculated with conventional Non-Uniform Entropy and the generic approach.

5. Results

We have evaluated our method with two well-known benchmark datasets [11]: *1)* an excerpt of 30,162 records from the 1994 US census database, *2)* 1,193,504 records from the Integrated Health Interview Series. Both datasets contained nine attributes. We have transformed the attributes by applying low, medium and high levels of full-domain generalization. Next, we suppressed increasing subsets of their records, which is a local recoding procedure. We calculated the information loss using conventional Non-Uniform Entropy and our generic approach. We have normalized these measures into the range [0,1]: 0% represents the original input dataset and 100% represents a variant of the dataset from which all information has been removed.

As can be seen in Figure 4, both conventional Non-Uniform Entropy as well as our generic variant measured the same loss of information for datasets which have been transformed with full-domain generalization only (the lines' starting points) and for datasets from which all data has been removed (the lines' endpoints).

However, the results of our approach reflected the linear increase in information loss which would be expected when an increasing fraction of records is removed from a dataset. In contrast, Non-Uniform Entropy first measured an increasing gain of information which then slowly turned into loss of information.

6. Conclusions

In this article, we have present a generic method for using Non-Uniform Entropy to assess the quality of data which has been de-identified with a wide variety of different transformation methods. Due to its sound information-theoretic foundation, Non-Uniform Entropy is of high practical relevance for health data de-identification [3, 4]. Our method provides a unified framework in which this model can be used to assess and compare the quality of differently transformed data to find a good or even optimal solution to a given de-identification problem. Non-Uniform Entropy has been designed for measuring the information loss induced by transforming values of discrete variables or variables which are discretized via generalization. For measuring the loss of information introduced by transforming continuous variables, which are also often used in biomedical data, we have implemented appropriate methods, e.g. mean squared error (MSE), into the ARX data anonymization tool as well [10].

References

[1] S. Schneeweiss, Learning from big health care data, *New England Journal of Medicine* **370** (2014), 2161–2163.

[2] B. Malin, D. Karp and R. H. Scheuermann, Technical and policy approaches to balancing patient privacy and data sharing in translational research, *Journal of Investigative Medicine* **58** (2010), 11–18.

[3] K. El Emam and L. Arbuckle, *Anonymizing health data: case studies and methods to get you started*, O'Reilly and Associates, Sebastopol, 1st ed., 2014.

[4] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt et al., A globally optimal k-anonymity method for the de-identification of health data, *Journal of the American Medical Informatics Association* **16** (2009), 670–682.

[5] A. De Waal and L. Willenborg, Information loss through global recoding and local suppression, *Netherlands Official Statistics* **14** (1999), 17–20.

[6] A. Gionis and T. Tassa, k-Anonymization with minimal loss of information, *Transactions on Knowledge and Data Engineering* **21** (2009), 206–219.

[7] V. Torra and J. Domingo-Ferrer, Disclosure control methods and information loss for microdata, P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, editors, *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies*, Elsevier, 2001, 91–110.

[8] J. Domingo-Ferrer and V. Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation, *Data Mining and Knowledge Discovery* **11** (2005), 195–212.

[9] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez and S. Martinez, t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation, *Transactions on Knowledge and Data Engineering* **27** (2015), 3098–3110.

[10] F. Prasser, F. Kohlmayer, R. Lautenschlaeger, C. Eckert and K. A. Kuhn, ARX – A comprehensive tool for anonymizing biomedical data, *AMIA Annual Symposium Proceedings*, 2014, 984–993.

[11] F. Prasser, F. Kohlmayer and K. A. Kuhn, A benchmark of globally-optimal anonymization methods for biomedical data, *International Symposium on Computer-Based Medical Systems*, 2014, 66–71.