Evaluating privacy of individuals in medical data

Health Informatics Journal
XX(X):1–13

DOI: 10.1177/1460458220983398
www.sagepub.com/
SAGE

Shannon K S Kroes[2,1,3], Mart P Janssen[2], Rolf H H Groenwold[3] and Matthijs van Leeuwen[1]

Abstract

Although data protection is compulsory when personal data is shared, there is no systematic method available to evaluate to what extent each individual is at risk of a privacy breach. We use a collection of measures that quantify how much information is needed to uncover sensitive information. Combined with visualization techniques, our approach can be used to perform a detailed privacy analysis of medical data. Because privacy is evaluated per variable, these adjustments can be made while incorporating how likely it is that these variables will be exploited to uncover sensitive information in practice, as is mandatory in the European Union. Additionally, the analysis of privacy can be used to evaluate to what extent knowledge on specific variables in the data can contribute to privacy breaches, which can subsequently guide the use of anonymization techniques, such as *generalization*.

Keywords

Anonymization, Data exchange, Generalization, Privacy, Uniqueness

Introduction

To be able to conduct research in the medical field, researchers often need to acquire and combine data collected by different institutions. For data to be published or exchanged, however, it is essential that the privacy of individuals can be guaranteed. Particularly medical data can contain very sensitive information that patients have supplied primarily for health care purposes and these patients may not even be aware that their records are used for medical research.[1]

The data controller has multiple responsibilities towards these patients, pertaining to the stage of data collection, storage and processing, according to the General Data Protection Regulation.[2] The principles of data protection specifically apply to data in which individuals could be identified or to data that could be used to identify individuals. To assess whether individuals are identifiable, it needs to be assessed which information can be exploited to uncover sensitive information and how accessible this information is.[2] For example, when a medical diagnosis can be uncovered using only a patient's age and gender, the patient may be more vulnerable to a privacy breach than when specific medical information is needed to uncover the diagnosis.

[1]Leiden University
[2]Sanquin Research, the Netherlands
[3]Leiden University Medical Center

Corresponding author:
Transfusion Technology Assessment Group, Donor Medicine Research Department, Sanquin Research, Plesmanlaan 125Y, 1066 CX, Amsterdam.

Email: s.kroes@sanquin.nl

Such individualized analyses of privacy have been performed in the literature, but primarily with case studies and no standardized procedure has been proposed.[3,4] A systematic and automatic approach is required, because studying the amount of background information needed to uncover an individual can be difficult to execute by hand. Many different combinations of background information have to be considered and without a systematic approach, some sensitivities may be overlooked or overestimated. For example, we will show that variables that can take on many values are generally less sensitive to privacy breaches than binary variables, though the contrary is often expected to be the case.

In the field of anonymization methodology, a number of measures have been proposed that could be used to analyze individual privacy risks.[5] Particularly, l-diversity, the (α, k)-anonymity framework, and (X, Y)-anonymity and -linkability can be used to evaluate the difficulty with which sensitive information can be uncovered when certain background information is known.[6–8] These measures are currently used on an aggregated scale, to define how privacy should be optimized by an anonymization method referred to as *generalization*. Generalization entails replacing observed values in the data by larger ranges of values, with the goal of creating overlap of information among different individuals, see Table 1 for an example. Applications can be found in many fields, including the medical field.[9]

In this work, we will show how existing definitions of privacy by Wong et al. (2006) and by Wang and Fung (2006) can be used by medical researchers who are required to evaluate the level of privacy of *individuals* in their data.[7,8] We make multiple adjustments to the measures and combine them with visualization techniques, resulting in a detailed representation of privacy.

This enables researchers to evaluate privacy such that they can carry out the responsibility that has been imposed by the General Data Protection Regulation in the European Union.[2]

Methods

In the following three subsections, we first introduce relevant definitions and notation of the setting that we consider. Next, we describe the measures that we will use to evaluate privacy, including examples. In the third and final subsection, we explain the materials and methods that will be used to demonstrate our approach in the Results section.

Setting, definitions and notation

We assume there is a data *owner*, e.g., a medical institution, that owns (original) data that they would like to share or publish. The data, denoted D, is presumed to be in matrix form, where each row represents one of n individuals and each column represents a variable v. We denote the domain of a variable v, i.e. the set of values the variable can take on, by $\text{dom}(v)$, and its size by $|\text{dom}(v)|$. Each variable is labeled as either *sensitive* or *auxiliary*. Sensitive variables contain private information that should not be revealed to third parties. All notation in our Methods is for a scenario with exactly one sensitive variable, but we will show that the analysis can be easily repeated with different variables labeled as sensitive.

An *adversary* is someone who is interested in revealing sensitive information from the data as shared by its owner. For this, we assume that an adversary 1) knows that an individual of interest is present in the data, and 2) may have *auxiliary information* to help uncover sensitive information. In its most general form, auxiliary information on an individual i, denoted a_i, is any set of constraints that limit the possible values that auxiliary variables can have for individual i. Unless mentioned otherwise, in this paper a_i refers to specific variable-value combinations, e.g., an adversary may know that a person is male and has age 48. Analogue to a_i, we define s_i to be the value of the sensitive variable for an individual i.

When an adversary can deduce s_i from a shared dataset D and some auxiliary information a_i, this is called a *privacy breach*.

Our approach requires three additional assumptions on auxiliary information:

Assumption 1. Auxiliary information is *correct*, i.e., a_i never excludes values that i actually has.

Assumption 2. The size and content of auxiliary information is unknown to the data owner.

Assumption 3. An adversary has auxiliary information on one individual only.

The first assumption is necessary because it is impossible to reason about what an adversary might conclude based on wrong information. The second assumption implies that we will need to consider all possible instances of auxiliary information. The third assumption may seem strong, but is necessary to keep the evaluation of privacy on the level of individuals feasible; others have also implicitly made this assumption, in the sense that it is assumed that adversaries cannot combine information on different individuals to rule out sensitive values.[6–8,10]

We now define upward and downward privacy, which have both been previously proposed in the l-diversity framework.[6]

Definition 1. An individual i is said to have *upward privacy* in data D, when D and any auxiliary information a_i do not enable an adversary to deduce the sensitive value of this individual, i.e., s_i.

Definition 2. An individual i is said to have *downward privacy* in data D, when D and any auxiliary information a_i do not enable an adversary to deduce that s_i is *not* a certain value.

Note that upward privacy is a prerequisite for downward privacy: when data allow to deduce that an individual has a particular sensitive value, this also implies that this individual does not have any other sensitive value. Consequently, downward privacy is a stronger notion of privacy than upward privacy. For binary sensitive variables, however, the two notions concur.

When a variable of an original data set is *generalized*, a generalization algorithm is used to merge different values of the variable. For example, an age variable could be represented by intervals of five or ten years instead of one year; see Table 1 for an example.

Privacy measures

In this subsection we introduce the measures we use for upward and downward privacy. The measures bear similarities to those proposed by Wong et al. (2006) and by Wang and Fung (2006).[7,8] At the end of the next two subsections we will discuss the changes we made for our approach.

Quantifying upward privacy We start with the introduction of an example.

Example 1: Upward privacy. Consider the original data set in Table 1, in a situation where an adversary, Bob, is interested in the diagnosis of Alice. Bob has knowledge on both auxiliary variables *age* and *gender*, as he knows that Alice is a woman of 50 years old. Clearly, Bob's auxiliary information is sufficient to identify Alice, who corresponds to the first row, and thus to breach her upward privacy by deducing the diagnosis.

In the example, Alice's upward privacy is breached because only a single row corresponds to the information that Bob has on her. If we replace the age variable by a generalized version with intervals of five years (Option 1 in the table), there is still only one row that corresponds

Table 1. Example data set, with *diagnosis* as sensitive variable and two options for the generalization of *age*.

Original data				Generalizations of Age	
ID	*Diagnosis*	Gender	Age	Option 1	Option 2
1	*cancer*	female	50	50-54	45-54
2	*cancer*	male	40	40-44	35-44
3	*cancer*	female	35	35-39	35-44
4	*arthrosis*	male	64	60-64	55-64
5	*diabetes*	female	49	45-49	45-54

to a 50-year-old woman; i.e., the privacy breach remains. If, however, we generalize *age* to ten year intervals (Option 2), we obtain a data set in which two rows correspond to a 50-year-old woman. Moreover, these rows have different sensitive values. As a consequence, Bob cannot be certain about Alice's diagnosis given the information that he has, hence Alice is protected from an upward privacy breach.

The example shows that individuals can be protected from upward privacy breaches by the presence of other individuals with similar auxiliary variable characteristics. That is, privacy is strongly related to *uniqueness*, as has been observed previously.[6,10] It is this 'uniqueness' that we should measure and quantify. We quantify uniqueness as a measure of privacy by modeling a scenario where an adversary will search the data for all individuals corresponding to the adversary's auxiliary information. We call these individuals *peers*[*]. Following, for an individual i and auxiliary information a_i, we quantify upward privacy as the proportion of peers in the data that have a sensitive value different from s_i. Formally, we define the *Proportion of Protective Peers* for i and given a_i as

$$\text{PPP}_i(a_i) := 1 - \frac{\#\{s_i, a_i\}}{\#\{a_i\}},$$

where $\#\{x\}$ denotes the number of individuals (or rows) in the data that match x, and $\{a_i\}$ denotes the collection of peers. (Clearly, x may concern sensitive and/or auxiliary variables.)

In the example above, Bob has *maximum auxiliary information*, as he knows the values of *all* auxiliary variables, i.e., *age* and *gender*. We denote the maximum auxiliary information for individual i by a_i^{max}. PPP has a minimum of 0, in which case all individuals in the data with given auxiliary information have the same sensitive value. It can easily be shown that when $\text{PPP}_i(a_i^{max})$ does not equal 0, $\text{PPP}_i(a_i)$ will not equal 0 for any other auxiliary information $a_i \subset a_i^{max}$. This guarantee can alternatively be seen from the *number of protective peers*:

$$\text{NPP}_i(a_i) = \#\{a_i\} - \#\{s_i, a_i\}.$$

NPP denotes the number of peers with different sensitive information (again, for given i and a_i). It is clear that NPP cannot decrease as we remove variable-value pairs from a_i. For example, the number of 50-year-old women in a particular data set cannot be larger than either the number of 50 year-old's or the number of women. Although it is obvious that an individual is vulnerable to an upward privacy breach when $\text{PPP}_i(a_i) = 0$, and thus $\text{NPP}_i(a_i) = 0$, a threshold needs to be chosen to define when an individual's privacy is considered to be *protected*, i.e., when individuals are considered not to be vulnerable to upward privacy breaches. We impose a threshold p as follows.

Definition 3. An individual i has *p-upward privacy* iff $\text{PPP}_i(a_i) > p$ for every a_i, with $0 \leq p < 1$.

When PPP is larger than p, the data owner finds that the individual is sufficiently protected from upward privacy breaches by the presence of peers with different sensitive values.

The measures (X, Y)-linkability and α-deassociation also consider the number of protective peers relative to the total number of peers.[7,8] We have chosen a different parameterization, such that higher values on PPP correspond to higher privacy, which makes our parameterization easier to interpret. Furthermore, our threshold is implemented differently, so that a threshold of 0 can be used. This threshold indicates that there must be at least one protective peer for an individual in the data, which is a condition that is much more difficult to impose in the other two frameworks. Due to the fact that a PPP of 0 would already be difficult to attain for every individual in a typical medical data set,

[*]Groups of peers are often referred to as an *equivalence class* in the literature.

we expect that this threshold will be most frequently used and our formulation is therefore a much more practical implementation.

Quantifying downward privacy We continue the example with downward privacy.

Example 2. Downward privacy. Consider Table 1 again, with *age* generalized by ten years (Option 2). Although Alice is protected from an upward privacy breach according to Definition 1, Bob can still infer something about Alice's diagnosis: she does *not* have arthrosis. This is an example of a downward privacy breach, as the adversary does not consider all sensitive values to be possible. By ruling out sensitive values, it may also be possible to narrow it down to the true sensitive value.

We use a measure strongly related to PPP to model downward privacy. Let $\mathbf{v}_i$ denote the values in the domain of the sensitive variable that individual i does *not* have, i.e., $\mathbf{v}_i = \text{dom}(s) \setminus s_i$. For each $v \in \mathbf{v}_i$, we determine the proportion of peers having v, which can be interpreted as the extent to which the (false) sensitive value might be considered for individual i. We define (value-specific) *Peer Protection* for given individual i, sensitive value v, and auxiliary information a_i as

$$\text{PP}_{i,v}(a_i) := \frac{\#\{v, a_i\}}{\#\{a_i\}}.$$

When $\text{PP}_{i,v}(a_i) = 0$, sensitive value v can be ruled out for individual i and thus provides no protection. Note that PPP can be trivially and naturally redefined in terms of PP, i.e., $\text{PPP}_i(a_i) = \sum_{v \in \mathbf{v}_i} \text{PP}_{i,v}(a_i)$.

Similar to upward privacy, we need a threshold to decide when a possible value occurs frequently enough in peers to provide protection. Given a threshold $q \in [0, 1)$, we define the following indicator function to decide whether a value v is considered probable for an individual i:

$$I_{q,i,v}(a_i) = \begin{cases} 0, & \text{if } \text{PP}_{i,v}(a_i) \leq q \\ 1, & \text{if } \text{PP}_{i,v}(a_i) > q. \end{cases}$$

Based on this, we can quantify how many false sensitive values would be considered by an adversary (and thus provide protection against downward privacy breaches). For this we define the *Proportion of Alternatives Considered* (PoAC) for an individual i, threshold q, and auxiliary information a_i:

$$\text{PoAC}_{q,i}(a_i) = \frac{\sum_{v \in \mathbf{v}_i} I_{q,i,v}(a_i)}{|\text{dom}(s)| - 1}.$$

Finally, downward privacy is defined as follows:

Definition 4. An individual i has *q-downward privacy* iff $\text{PoAC}_{q,i}(a_i)$=1 for every a_i, with $0 \leq q < 1$.

This means an individual is considered safe from downward privacy breaches if they are protected by sufficient peers for every possible false sensitive value, i.e., $\text{PP}_{i,v}(a_i) > p$ for every $v \in \mathbf{v}_i$.

Although we could choose q independently from p, in practice it makes sense to choose them jointly, so that downward privacy remains a stronger notion of privacy than upward privacy. For example, when all false sensitive values are expected to be equally likely, we could first choose p based on the domain size and then set $q = \frac{p}{|\text{dom}(s)|-1}$, which would divide p uniformly over all values in $\mathbf{v}_i$. This also results in $p = q$ for binary sensitive variables, which is a logical choice as this would make upward and downward privacy coincide for that case.

With respect to downward privacy, PoAC bears some similarities to α-rarity and (X, Y)-anonymity.[7,8] These alternatives deviate from our definition in that the former is independent from the true sensitive value while the latter uses an aggregated measure that implicitly assumes a threshold p=0. Therefore, PoAC reflects the risk that specific sensitive information can be extracted more accurately than α-rarity

and it is more flexible than (X, Y)-anonymity in that the user can specify a threshold.

Example 3. Interpretation of the measures To conclude this subsection, we briefly illustrate how the measures can be interpreted when applied to Example 1. With generalization Option 1, the number of women between 50 and 54 is equal to the number of women between 50 and 54 with cancer. Thus, given Bob's auxiliary information on Alice PPP, NPP and PP all equal 0. Specifically, in this case NPP can be interpreted as the number of women without cancer for whom age could be 50 years.

When *age* is generalized by ten year intervals (Option 2), half of the women between 45 and 54 have cancer and Alice would have a PPP of 0.5 and NPP of 1. Note that if Bob's auxiliary information states that Alice is between 40 and 60 years old, he would still consider all 50-year-old women without cancer and thus the NPP would not decrease. Her downward privacy could still be considered low, as only the false sensitive value diabetes is considered. With a threshold q of 0, her $\text{PoAC}_{i,0}(a_i)$ equals 0.5, as $\text{PP}_{i,diabetes}(a_i^{max}) = 0.5$ and $\text{PP}_{i,arthrosis}(a_i^{max}) = 0$.

Scalability of the measures In a naive approach, the number of times the number of peers has to be counted increases linearly with the number of sensitive variables (k_s) and with the number of individuals (n) and factorially with the number of auxiliary variables (k_a). This amounts to $n \sum_{j=1}^{k_s} |dom(s_j)| \sum_{a=1}^{k_a} \binom{k_a}{a}$ search operations in the worst-case scenario, where every combination of auxiliary variables is tested one at a time for every sensitive value. A more efficient implementation could account for the fact that peers with the same sensitive value will have the same level of privacy. Moreover, when a set of auxiliary values occurs x times in the data, any set that contains this set of values, will also occur at most x times (e.g. if there is only one fifty-year old woman in the data, there is also at most one fifty-year-old woman with cancer). Another possibility is to perform the privacy analysis for a random sample of the individuals in the data, for example using finite population statistics to evaluate the generalizability of these results. If a data set is too large, even for a more efficient implementation, one can also consider using the NPP or evaluating only a subset of the combinations of auxiliary information.[4] On a i5-8265U CPU 1.60GHz with 16GB RAM, the computation of all PPP values for Figure 2 took approximately 4000 minutes CPU time when testing for all combinations of auxiliary information and 10 minutes when only assessing the PPP with maximum auxiliary information.

Materials and methods for demonstration

In the next section we will demonstrate the use of the measures for the systematic evaluation of privacy of individuals. Here, we describe the three publicly available data sets that will be used. We also explain how we use the measures to quantify and subsequently visualize privacy.

Data sets To demonstrate our approach, we use three data sets from the University of California Irvine (UCI) Machine Learning Repository: the Adult data set[11], a data set on diabetes from US hospitals[12], and a cervical cancer risk factor data set[13]. Details on the selected variables can be found in Table 2. We have included the variables of the adult data set that are most frequently used in articles on generalization algorithms.[6,14,15] Individuals with missing values on one of the selected variables were excluded from the data set, which concerned less than 10% of all rows in each of the data sets. We excluded individuals with missing values to avoid having to make additional decisions that potentially influence the results.

Using the measures for privacy evaluation To evaluate privacy, we use the previously detailed measures to assess whether auxiliary information can reveal sensitive information. First, we use the PPP on each variable to gain insight into which variables are vulnerable to being

uncovered. Second, to investigate which variable is most likely to provide a privacy increase when generalized, we assess whether there is an increase in the PoAC when each variable is left out of the auxiliary information, i.e., how much of an increase in privacy is to be expected if information on this variable were unknown to the adversary.

Implementation and visualization We performed all analyses and subsequent visualization in R Studio version 1.0.136. We used the *heatmap.2* function from the *gplots* package to create heatmaps for visualization of the privacy measurements, and used the function's feature that allows columns to be ordered such that the plot is easiest to interpret. We set the colors, such that all values below or equal to the chosen threshold (such as p) are dark red. Because two data sets contain an excessive number of individuals relative to the amount of pixels, we used subsampling to smooth the plots. We developed a basic online tool that implements our approach: https://skskroes.shinyapps.io/Evaluating_and_visualizing_privacy/. Additionally, our code is made available on the Github repository https://github.com/ShannonKroes/Evaluating_and_visualizing_privacy. If users want to apply the code to their (potentially sensitive) data, we recommend downloading the R code and running it on their own device.

We present a small example of our visualization in Figure 1 that depicts downward privacy (q=0) for Table 1, with the second generalization option[†]. Each row represents a variable and each column represents an individual. We depict the corresponding variable names and for this example we also show the ID numbers from Table 1 for the columns. The color in a cell represents the level of downward privacy given that the other variables in that column are known. For example, for the third individual, the diagnosis can be uncovered if age and gender are known, because all women in this age category have the diagnosis *cancer*. Therefore, the first cell in the third column is red, indicating the worst level of privacy. Gender is more difficult to uncover, since both a male and a female have cancer and fall in the age category 35-44 and thus the value on *gender* is protected for both of these individuals, shown by a green second cell in the third column. On the other hand, only two out of three of the age categories (35-44 and 45-54) are associated with a woman with cancer, thus resulting in a PoAC of $\frac{2}{3}$ and a yellow last cell in the third column.

Results

In the following, we will visualize the previously detailed measures with heatmaps to provide an intuitive representation of the risks of privacy breaches. First, we will use our approach to evaluate the privacy of the three data sets in Table 2. Next, we highlight how this can reveal where vulnerabilities to privacy breaches occur, and where these may originate from. Finally, we demonstrate how the approach can be used as an asset for generalization.

Evaluating upward privacy in the data sets

Figure 2 visualizes upward privacy for each variable for each data set. The level of privacy is visualized in a heatmap, such that a row closer to red indicates that that variable is more vulnerable to privacy breaches. The heatmaps illustrate that the level of privacy is generally low in all three data sets, despite the fact that two of the data sets are comprised of a relatively large number of individuals. This relates to the large number of unique rows in the data, as shown in Table 2. Using the PPP, we can also make between-variable comparisons. One important observation is that variables with larger domains tend to be more difficult to uncover, and are thus associated with lower privacy risks. This is due to the fact that these variables are also most

[†]For the example we left the columns in the same order as in Table 1.

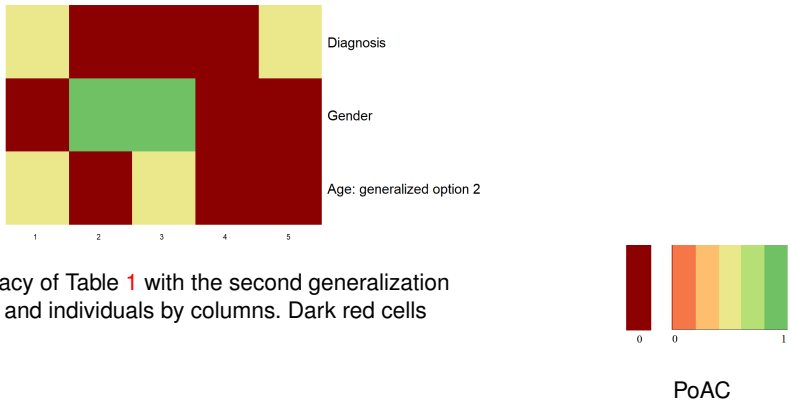


Figure 1. Visualization of downward privacy of Table 1 with the second generalization (q=0). Variables are represented by rows and individuals by columns. Dark red cells indicate PoAC equals 0.

Table 2. Properties of data used.

	n	n^{unique}	Variables (domain size)
Adult	30 162	19 502	Gender (2), Age (72), Race (5), Marital status (7), Education (16), Native country (41), Work class (7), *Salary* (2), Occupation (14)
Diabetes	99 493	77 748	Gender (3), Age (10), Race (5), Number of lab procedures (118), Number of medications (75), Change in medications (2), Diabetes medications (2), *Readmitted* (3)
Cervical cancer	789	254	Number of pregnancies (11), Smoking (2), Age (43), *Biopsy result* (2)

n denotes the total number of rows, n^{unique} the number of unique rows. Sensitive variables are given in *italic*, the domain size of each variable is given between brackets.

(a) Adult **(b)** Diabetes **(c)** Cervical cancer

Figure 2. Upward privacy in the three data sets. Variables are represented by rows and individuals by columns. Each value depicts the minimum PPP over all possible combinations of auxiliary information for that particular individual (column), under the assumption that that variable (row) is labeled as sensitive. Colours more towards green indicate high privacy and dark red colours indicate privacy insufficient according to the chosen threshold.

informative when part of the auxiliary information. That is, when labeling such a variable as sensitive, it follows that the values for this variable are unknown to the adversary, which means that the adversary loses valuable information. A noteworthy example is the variable *Occupation* in the adult data set, which is difficult to uncover, despite the fact that this variable is very frequently selected as a target to be protected by generalization algorithms.[6,10,15] In fact, researchers tend to explicitly choose to generalize data such that variables with a larger domain are protected, whereas Figure 2 shows that binary variables can be much more vulnerable to privacy breaches.[15] For example, the variables *change in medication* and *diabetes medications* in the diabetes data set are easy to uncover. This shows the importance of assessing the distribution of privacy before using or testing a generalization algorithm on the data.

Detecting the origin of vulnerabilities

In Figure 3, we depict the contribution of each auxiliary variable to the level of privacy on the sensitive variable, as specified in Table 2. We show the PoAC for the sensitive variables selected for the adult and diabetes data sets (Figure 3a and 3b), and the PPP for the sensitive variable in the cervical cancer data set (Figure 3c). Each

row represents what the level of privacy would be if the variable corresponding to that row were unknown to the adversary. In this scenario, the values of all other auxiliary variables would be known to the adversary. Rows with colors closer to green colors correspond to variables that provide valuable information to uncover the sensitive value. Conform our expectations, particularly variables with larger domains provide information that has the potential to result in privacy breaches, because knowing the values on these variables will most likely enable the adversary to rule out a large number of individuals.

After analyzing which variables contribute to privacy breaches, the user can make choices regarding which variables should be generalized. In doing so, the *accessibility* of variable information needs to be taken into account. For example, in the diabetes data set, *age*, *number of medications* and *number of lab procedures* are all valuable pieces of information in uncovering readmission, but the difficulty with which information on these variables can be acquired may differ significantly. Incorporating how likely it is that certain information will be used to identify individuals is a compulsory part of privacy evaluation in the European Union, which is possible with our approach.[2]

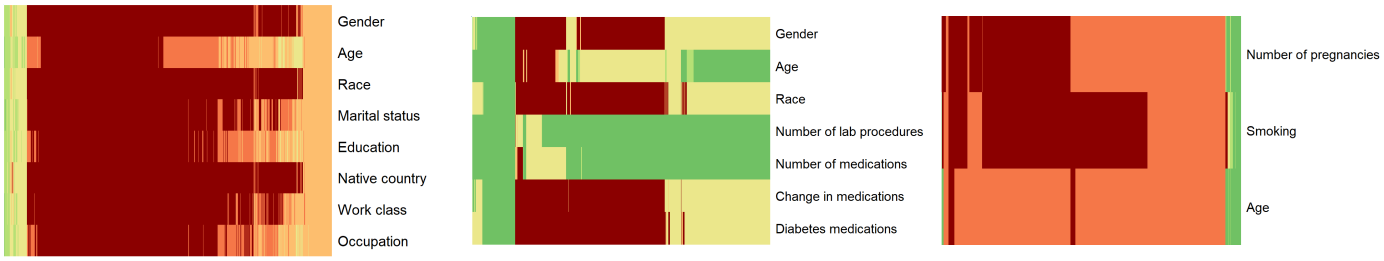
Assessing the potential for privacy increase resulting from generalization

In this subsection we show how Figure 3 can be used to assess the potential increase in privacy that can be gained from generalizing each variable. We discuss upward privacy of the biopsy result in the cervical cancer data, with p=0. Figure 3c shows that the variable with the largest domain, *age*, is most frequently associated with a vulnerability to privacy breaches, compared to the other two auxiliary variables. Therefore, we would expect that generalizing *age* would be more effective than generalizing *number of pregnancies*[‡]. To illustrate this difference, we show the effect of generalizing these variables in Figure 4. We generalize both variables separately, choosing multiple category sizes, with the constraint that each category must contain an equal number of values, except for the lowest and/or highest category. Figure 4 indeed shows that *age* has more potential to decrease uniqueness. With a threshold p of 0 the proportion of individuals with upward privacy can increase to up to 97%, whereas the maximum proportion of protected individuals due to generalizing the number of pregnancies is only 59%. Another important observation is that further generalization does not necessarily result in an increase in privacy and generalizations with the same category size can have very different effects. This can also be seen in the example data set in Table 1, where Alice is still unique when she falls into the age category 50-54, whereas she would have had increased protection if the age category had been 46-50.

Discussion

Summary Evaluating privacy risks has become a compulsory part of sharing individual patient data for scientific purposes. Researchers need to evaluate how much background information is needed to uncover sensitive information and how easily this background information can be accessed.[2] In this work, we have presented an approach to privacy analysis that is detailed enough to perform this task, by evaluating privacy per variable and per individual. As detailed in our Methods section, we make critical modifications to measures presented by Wong et al. (2006) and by Wang and Fung (2006), which have so far primarily been used on an aggregated level to optimize generalization algorithms.[7,8] Combined with visualization techniques, this reparameterization results in an intuitive representation of privacy risks in the data on an individualized level. This can provide insight into which variables cause these risks. In turn, these variables can

[‡]Generalizing *smoking* would result in everyone having the same value; this has been shown to be an ineffective generalization as depicted in the row corresponding to *smoking* in Figure 3.

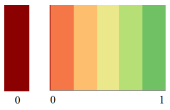


(a) Adult: PPP for the sensitive variable *salary*

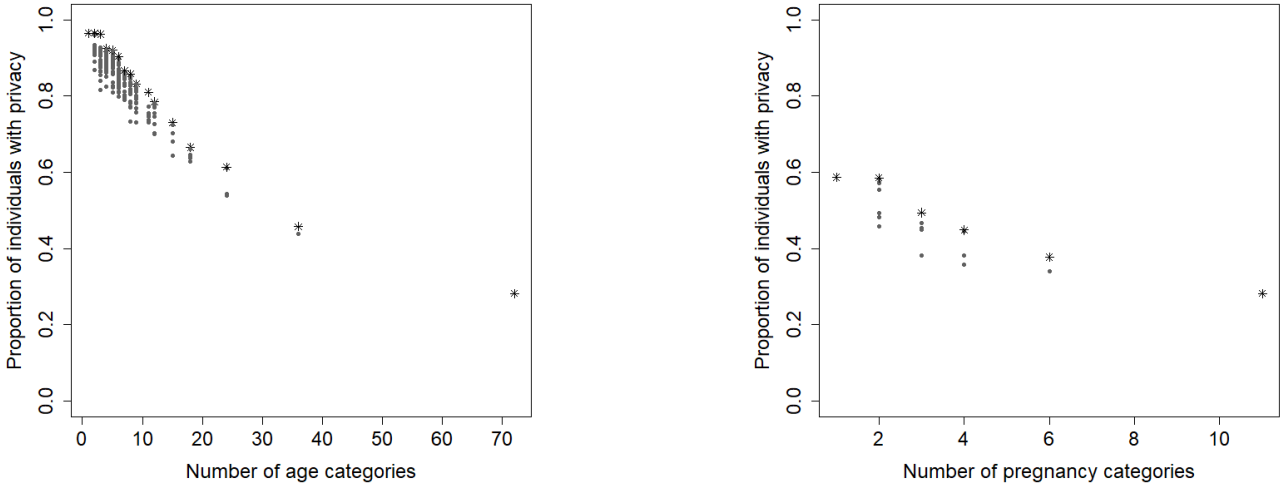
(b) Diabetes: PoAC for the sensitive variable *readmission*, q=0

(c) Cervical cancer: PPP for the sensitive variable *biopsy result*

Figure 3. Privacy increase when an auxiliary variable is unknown to the adversary for the three data sets. Every row represents an auxiliary variable and every column represents an individual. The auxiliary information that can be used to uncover the specified sensitive variable contains the true values on every auxiliary variable in the data, except for the variable corresponding to that row. Colors closer to green represent higher PPP or PoAC, indicating that knowing the value for that individual on that variable is more informative to the adversary. Dark red colors indicate that the privacy measure equals zero.



PPP or PoAC



(a) Generalization of *age* (p=0)

(b) Generalization of *number of pregnancies* (p=0)

Figure 4. Privacy of the sensitive variable *biopsy result* as a function of generalization. The x-axes depict the number of categories the variables are partitioned into, as an indication of the extent of generalization. The y-axes show the number of individuals who are deemed protected with $p = 0$, taking the minimum PPP over all combinations of auxiliary information. Because a variable can be partitioned into a certain number of categories in multiple ways, the same number of categories can correspond to different proportions of individuals with privacy. The generalization with the highest privacy gain per number of categories is marked with an asterisk.

be adjusted to minimize opportunities to misuse background information, e.g., by using the anonymization technique *generalization*.

Related work To our knowledge, this work is the first to present a method that enables data owners to study the level of privacy of individuals in their data that can also show the contribution of specific variables to privacy breaches. Though software is available to measure privacy, most of them use definitions that do not specifically assess whether sensitive information can be extracted, but only whether entire records of individuals are unique in the data. This disregards groups of peers with similar sensitive characteristics that are at risk of a privacy breach. This includes both the k-anonymity framework and the body of literature dedicated to measuring re-identification risk, as well as the corresponding software that implements these approaches (such as sdcMicro[16], the ARX data anonymization tool[17], Amnesia[18] and μ-ARGUS[19]), some of which have been applied to medical data.[10,20–22]

Another popular approach is to summarize the level of privacy with an integer l using l-diversity, which has been implemented in the UTD Anonymization toolbox.[6,23] This is very similar to the PoAC with a threshold $q = 0$, assuming maximum auxiliary information, without controlling for the number of possible sensitive values. Our approach is more flexible in setting thresholds and modeling auxiliary information. Further, *individual* risks can be evaluated and explored, as opposed to summarizing the level of privacy with the *worst-case* individual.

The UTD Anonymization toolbox also implements t-closeness, which compares the distribution of a sensitive attribute over the entire data set to the distribution of the same sensitive attribute within a group of peers.[14] Other approaches with a similar aim include β-likeness[24] and δ-disclosure privacy[25]. These thresholds and measures are on the scale of a chosen distance measure, such as the Kullback-Leibler divergence, but they are difficult to interpret, as they are often on a logarithmic scale. Additionally, they cannot be directly related to real-life situations where an adversary will try to uncover a sensitive value using certain background information.

All of the mentioned measures and corresponding software evaluate privacy on an aggregate level. One approach that does measure personalized privacy is that by Xiao and Tao (2006), but in this work assumptions are made about how an external data set will be used to extract sensitive information and about what this data set looks like.[26] Specifically, when measuring privacy for a certain group of peers, it is assumed that the external data set contains all of these peers and the level of privacy increases factorially with the size of the external data set. For an extensive overview of technical privacy metrics, see Wagner and Eckhoff (2018).[5]

Strengths, limitations and future work As publishing data is becoming more common in scientific literature, our approach can be used to evaluate the privacy risks for the individuals included. Additionally, privacy can be evaluated when a hospital is considering supplying their data for the purpose of multi-center clinical research. Another possible application is in the process of developing or testing anonymization techniques, when the privacy of the input data needs to be evaluated in order to be able to interpret the performance of the methodology.

In our approach, upward and downward privacy is investigated by reviewing the corresponding measures for all combinations of auxiliary variable information. This is a very time-consuming task with our naive implementation and a more efficient implementation should be developed. Because many of the needed computations are independent, the code could be run in parallel to speed up the computation.

Another limitation of our current implementation is that continuous variables are treated as discrete. Continuous variables could be modeled more accurately by evaluating the proximity to auxiliary or sensitive values.[27] This entails that one can specify a range of values for the auxiliary information and that the threshold for a privacy breach could include a range of values that lie close to the true value.

Conclusion

We have proposed an approach that enables medical researchers to evaluate the level of privacy of individuals in their data. Specifically, our approach quantifies and visualizes which variable information can be exploited to breach privacy and thus which variables should be targeted with anonymization techniques. Considering that evaluating privacy and using anonymization methods is likely to be a responsibility that comes with the exchange of patient information, our approach can be a valuable asset in the process of sharing individual patient data.

Acknowledgements

We acknowledge contributions made by R.A. Middelburg during the early stages of conceptualization.

Declaration of conflicting interests

The authors declare that they have no competing interests.

Funding

This project is funded by the Sanquin Blood Supply Foundation (PPOC-16-27).

References

1. Cios KJ and Moore GW. Uniqueness of medical data mining. *Artificial intelligence in medicine* 2002; 26(1-2): 1–24.
2. European Parliament and Council of European Union. Regulation (eu), 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN. Accessed: 7 October 2020.
3. Solomon A, Hill R, Janssen E et al. Uniqueness and how it impacts privacy in health-related social science datasets. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, pp. 523–532.
4. Manfredi V. *Privacy Implications of New York City's Stop-and-Frisk Data*. PhD Thesis, Wellesley College, US, 2015.
5. Wagner I and Eckhoff D. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)* 2018; 51(3): 1–38.
6. Machanavajjhala A, Gehrke J, Kifer D et al. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*. IEEE, p. 24.
7. Wong RCW, Li J, Fu AWC et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 754–759.
8. Wang K and Fung B. Anonymizing sequential releases. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 414–423.
9. El Emam K, Dankar FK, Issa R et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association* 2009; 16(5): 670–682.
10. Sweeney L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002; 10(5): 557–570.
11. Dua D and Graff C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.
12. Strack B, DeShazo JP, Gennings C et al. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international* 2014; 2014.
13. Dheeru D and Karra Taniskidou E. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.
14. Li N, Li T and Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *23rd International Conference on Data Engineering*. IEEE, pp. 106–115.
15. Kohlmayer F, Prasser F and Kuhn KA. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. *Journal of biomedical informatics* 2015; 58: 37–48.
16. Templ M. Statistical disclosure control for microdata using the r-package sdcmicro. *Transactions on Data Privacy* 2008; 1(2): 67–85.
17. Prasser F and Kohlmayer F. Putting statistical disclosure control into practice: The arx data anonymization tool. In *Medical Data Privacy Handbook*. Springer, 2015. pp. 111–148.
18. Institute for the Management of Information Systems. Amnesia. a data anonymization tool supported by the institute for the management of information systems, 2016. URL https://

amnesia.openaire.eu/installation.html. Accessed: 7 October 2020.

19. Franconi L and Polettini S. Individual risk estimation in μ-argus: A review. In *International Workshop on Privacy in Statistical Databases*. Springer, pp. 262–272.
20. Skinner C and Holmes DJ. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics* 1998; 14(4): 361.
21. Dankar FK and El Emam K. A method for evaluating marketer re-identification risk. In *Proceedings of the 2010 EDBT/ICDT Workshops*. pp. 1–10.
22. Taneja H, Singh AK et al. Preserving privacy of patients based on re-identification risk. *Procedia Computer Science* 2015; 70: 448–454.
23. Kantarciouglu M, Inan A and Kuzu M. UTD anonymization toolbox, 2012. URL http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php. Accessed: 4 October 2020.
24. Cao J and Karras P. Publishing microdata with a robust privacy guarantee, 2012. 1208.0220.
25. Brickell J and Shmatikov V. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 70–78.
26. Xiao X and Tao Y. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. pp. 229–240.
27. Li J, Tao Y and Xiao X. Preservation of proximity privacy in publishing numerical sensitive data. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, pp. 473–486.