Achieving k-anonymity in DataMarts used for gene expressions exploitation

Konrad Stark, [1] Johann Eder [1], Kurt Zatloukal [2]

[1]University of Vienna, Department of Knowledge and Business Engineering,
Rathausstrasse 19/9 A-1010 Wien
konrad.stark@univie.ac.at, johann.eder@univie.ac.at www.cs.univie.ac.at

[2]Medical University Graz, Institute of Pathology,
Auenbruggerplatz 25, A-8036 Graz
kurt.zatloukal@meduni-graz.at, www.meduni-graz.at, www.bioresource-med.com

Abstract

Gene expression profiling is a sophisticated method to discover differences in activation patterns of genes between different patient collectives. By reasonably defining patient groups from a medical point of view, subsequent gene expression analysis may reveal disease-related gene expression patterns that are applicable for tumor markers and pharmacological target identification. When releasing patient-specific data for medical studies privacy protection has to be guaranteed for ethical and legal reasons. k-anonymisation may be used to generate a sufficient number of k data twins in order to ensure that sensitive data used in analyses is protected from being linked to individuals. We use an adapted concept of k-anonymity for distributed data sources and include various customisation parameters in the anonymisation process to guarantee that the transformed data is still applicable for further processing. We present a real-world medical-relevant use case and show how the related data is materialised, anonymised, and released in a data mart for testing the related hypotheses.

1 Introduction

Gene expression profiling is a sophisticated method to discover differences in activation patterns of genes between healthy and diseased cells and tissues. Molecular signatures that correlate with diseases may be unveiled and used for prognoses. In cancer research tumour markers are derived from reliable classification patterns of gene expressions. Further, previously unknown subtypes of cancers may be detected. To support both gene expression analysis and data release for clinical studies we developed a virtual data warehouse solution where patient collectives are selectable and gene expression profiles are linked for further analysis. Patient data (from clinical databases) and gene expression data is integrated and materialised on demand in a data mart. Subsequent analyses may be used to detect connections (correlations) between patient related data (personal data, history, lifestyle data, clinical analysis data) and patterns in gene expression profiles.

When releasing patient-specific data for medical studies privacy protection has to be guaranteed for ethical and legal reasons. Even when immediately identifying attributes like name, address

or day of birth are eliminated, other attributes (quasi-identifying) may be used to link the released data with external data to re-identify individuals. The concept of k-anonymity requires that each distinct combination of quasi-identifying attributes occurs at least k times in a shared table. [PS98]. Hence, a sufficient number of k data twins is used to mantle the individuality of persons. The number of data twins may be increased by transforming attribute values to more general ones using predefined generalisation hierarchies or by simply suppressing attributes.

Several methods for achieving k-anonymity have been developed. Generally, we can distinguish between approaches using global recoding [WYC04, LDR05, FWY05] of attribute values, those using constrained local recoding [LDR06] and those using full local recoding ([XWP+06]). Global recoding of attribute values may be defined as a set of functions that transform attribute values to generalised or altered values. In a single-dimensional global recoding all values of a certain attribute domain are mapped to transformed values, while multi-dimensional global recoding (also considered as constrained local recoding) recode combinations of attribute domains. Full local recoding is used to recode non-distinctive attribute values, that is, two tuples with equivalent attribute values may be recoded differently. We are using full-domain generalisation which is a kind of single-dimensional global recoding. Each attribute is generalised independently from all the other attributes. Further, full-domain generalisation guarantees that every attribute value belongs to the same generalised domain [LDR05]. We prefer full-domain generalisations since the recoded data is useful for the context of our application. The transformed records are released to medical studies that investigate correlations between attributes, classify diseases in subtypes and make survival analyses of different courses of diseases. All those analyses need data transformed to equal generalisation levels. If the values for disease free survival are recoded to arbitrary intervals (e.g.: [0-5], [10-20] months) or not coded at all, a postprocessing of records is required. Additionally, after anonymisation any record may be easily transformed to a more general value but can not be recoded in a more specific one. Thus, in comparative analysis, all attribute values would have to be transformed up to the generalisation level of the most general one.

Following the concept of k-anonymity the anonymity of patients may be preserved by generalising and/or suppressing attribute values in order to generate data twins in the released data. Our data integration and anonymisation concepts have been developed in the context of the Austrian Genome Programme GEN-AU [iA] and as preliminary work for the biobank initiative of the Medical University Graz [BIO]. In this paper we extend our anonymisation algorithm [SEZ06] in order to guarantee k-anonymity in a shared table that is built from a set of data sources. We prevent that none of the shared table records may be used to reidentify individuals by generating data twins for each data source separately. Instead of allowing access to the entire shared data we release data as the result of data requests. Users are able to select a subset of the shared table they want to use in further analysis by defining projection and selection criteria.

Generalisation of attribute values is always related with an information loss. Though, the acceptable information loss may differ as the case arises. Some attributes should be generalised as little as possible while others may be transformed to more general values. For instance, in some medical study it's sufficient to know the age of patients in steps of five (40-45, 46-50,..) but the size of tumours is to be specified as detailed as possible. Hence, we allow the users to weight the importance of attributes for a specific data request by defining priorities and generalisation limits. Those criteria are considered in the anonymisation process and have a strong impact on the quality of the transformed data.

Further, we take into account different information losses in generalisation hierarchies. We therefore attempt to integrate knowledge of medical domain experts by allowing them to assess the information loss within each attribute hierarchy. Medical experts assign information loss quantifiers to each level of a generalisatin hierarchy in order to estimate the data quality decrease for each transformation step. A more detailed description of different information losses is given in section 4.2.

The paper is structured as follows: We describe the type of privacy protection we concentrate on in section 2. Section 3 describes the data sources that are available at the Medical University Graz and are integrated in a shared table. The process of anonymisation is specified in section 4. Finally, a detailed example use case from the medical research domain is presented in section 5.

2 General threat scenario

The original work of Samarati [PS98] focuses on the reidentifaction risk of linking released private data with publicly available information (i.e. city's voter list). However, privacy protection through k-anonymity is also applicable to shared data that result from data integration of multiple parties. In the work of [WFD05] two parties share data to benefit from a common classification analysis. Since the data sources remain separated, until an integrated table is built and common *identifying* attributes are used to link the two sources, those attributes could be used to learn attribute values of the other party. Thus, the set of *quasi-identifying* attributes is extended to all shared attributes of both parties. However, in our approach a trusted third party is responsible for linking records from all data sources and eliminating *identifying* attributes in the released data. Each data source has its own set of *quasi-identifying* attributes which corresponds to the set of attributes that is shared globally.

In our scenario a few parties share their private data in a publicly available table. Each party performs analyses on the global table but no party is able to deduce additional infomation related to its stored individuals. In Fig.1 three parties A, B, C are sharing a set of attributes in an integrated table. Each party has identifying keys for the individuals that are not shared globally. Any party may access the shared table, execute queries and use the result sets for further analysis. Though, it should not be possible to infer the identity of individuals by matching the global result set with the local data source. Further, if any external intruder gains access to the global table and to any local source he is not able to link all shared attributes to one individual. Therefore, we impose the k-anonymity constraint on each data source separately. Any globally shared record has to have k data twins in every data source. If we want to provide k-anonymity of k=3 in our example each tuple $t_i = t(i, v_1^i, v_2^i, v_3^i, v_4^i, v_5^i, v_6^i, v_7^i)$ must have two other data twins for (v_1^i, v_2^i) in data source A, (v_3^i, v_4^i, v_5^i) in B and (v_6^i, v_7^i) in C.

Consider the following example: pathological findings, survival data and clinical data is integrated in a shared table (table 1). Although no immediately identifying attributes are shared, distinct attribute value combinations may be used to reidentify individuals in the local sources. We examine for every data source whether there exist data twins for the shared attributes. In our case records 1, 2, 5 and 6 have no data twin for attributes PT, PN and PM. Further, records 3 and 6 have unique values for cause of death and all records do not have any data twin for

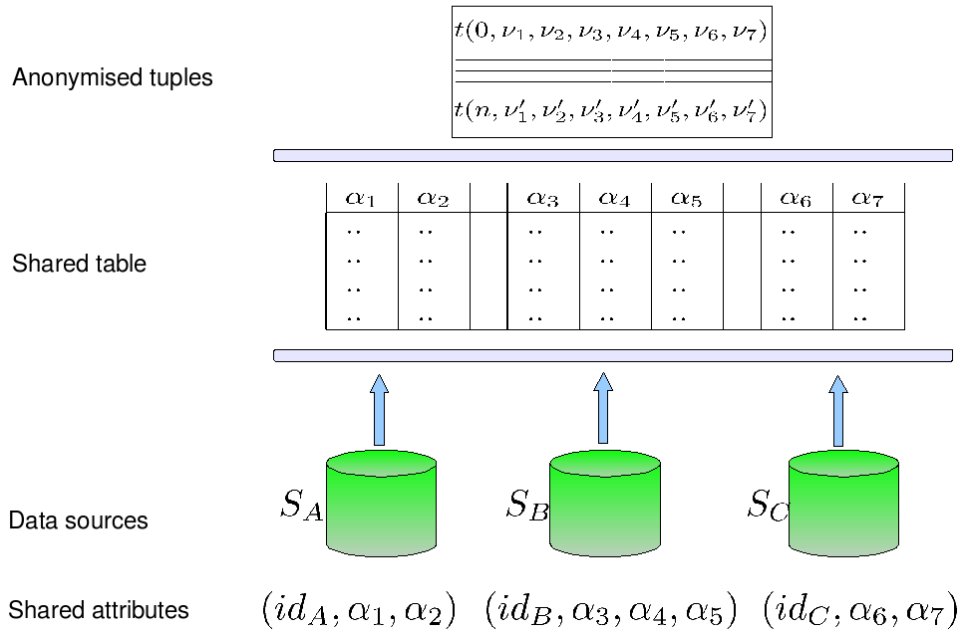

Figure 1: Shared table

columns weight and height. By applying recoding and generalisations data twins are created as shown in table 2. Staging PT is generalised by replacing 3a and 3b values with an aggregated 3 specification. The cause of death column is transformed by categorically recoding textual descriptions in ICD-N codes. Further, attributes weight and height are mapped to BMI (body mass index) categories. Hence, k-anonymity of 2 has been accomplished for attributes of findings (PT,PN,PM), survival data and BMI.

RecId		*PT*	*PN*	*PM*		*Cause of Death*		*Height(cm)*	*Weight(kg)*
1		3A	0	1		Atherosclerotic heart disease		185	85
2		3A	0	X		Sigmoid colon		173	80
3		3B	0	X		Colon, unspecified		175	69
4		3B	0	X		Sigmoid colon		172	70
5		3B	0	1		Atherosclerotic heart disease		170	62
6		3B	0	X		Caecum		183	80

Table 1: Integrated Table

3 Data sources

The pathology Graz has different databases for pathological findings, gene expression profiles and survival data. All three data sources contain sensitive information of patients like tumour diagnoses, gene expression profiles, survival periods and causes of death. Each data source access is limited to authorised persons. A virtual integration layer links data records from all

RecId		*PT*	*PN*	*PM*		*Cause of Death*		*BMI*
1		3	0	1		I25		[25.0,..,29.9] Overweight
2		3	0	X		C18		[25.0,..,29.9] Overweight
3		3	0	X		C18		[18.5,..,24.9] Normal
4		3	0	X		C18		[18.5,..,24.9] Normal
5		3	0	1		I25		[18.5,..,24.9] Normal
6		3	0	X		C18		[18.5,..,24.9] Normal

Table 2: Anonymised Table

sources, selects subsets of relevant attributes and allows export of records for further medical investigation. Gene expression profiles are related to tissue or blood samples and are not immediately linked to patients. The relationship from individuals to gene expression profiles must not be deduced from records of the virtual integration layer. A similar protection is to be guaranteed for survival data: survival period and cause of death are essential parameters for survival analyses, though, that information must not be explicitly linked to patients. In the following, we present the data sources in more detail.

Table 3 is a set of patient findings storing pathological diagnoses related to different types of preparations. Tissue samples are taken from biopsies, as preparations during operations or from autopsies. The majority of samples is tumour-related, thus, tumour-specific data such as staging, grading, textual diagnosis and tumour localisation is stored as well as the age of the patient at that moment the preparation was taken. SendDate gives information about the date the preparation was taken and Sender stores information about the hospital where the preparation originated. The column PatId links to a separate patient table where identifying attributes name, surname, day-of-birth are available. Each person having access to the findings table also has access to the patient table and is able to identify findings of individuals.

Survival data is stored as shown in table 4. Cause of death is specified by two ICD-10 codes

PREPID	PATID	LOC.	AGE	PT	PN	PM	G	R	SENDDATE	SENDER
8	19	Colon	67	3	2	1	2	X	23.09.2003	Surgery
11	19	Mamma	69	1C	1BII	0	3	0	15.02.1998	Graz East
14	22	Mamma	51	1A	1A	0	3	0	07.03.2005	Surgery
...	...	...	...	...	...	...	...	...	...	...

Table 3: Findings

(ICD-E and ICD-N). The survival period is calculated by subtracting send-date (findings table) from the day-of-death column. Pat-Ident is a combination of name, surname and day-of-birth. Patients that are still alive have empty columns for day-of-death and ICD-E/ICD-N. Finally, gene expression profiles are stored separately in table 5 and are linked by preparation ids (PrepId). In the use case in section 5 we focus on analyses of the survival period from survival data and the gene expression profiles from the genes table. Therefore, linking gene expression profiles and survival data data to individuals may only be accomplished by matching attributes of the findings table with the shared ones. For that reason we just have to transform attributes from the findings table to fulfil the k-anonymity constraint.

PAT-IDENT	DAY-OF-DEATH	ICD-E	ICD-N
Ident1	15.08.2005	S224	V486
Ident2	17.03.2000	S069	V849
Ident3	19.01.2006	I219	U50
...	...	...	...

Table 4: Survival data

PREP-ID	GENE EXPRESSION PROFILE
1	$(e_1^1, e_2^1, .., e_n^1)$
3	$(e_1^3, e_2^3, .., e_n^3)$
4	$(e_1^4, e_2^4, .., e_n^4)$
...	...

Table 5: Gene Expression Profiles

4 Anonymisation

Given a set of source tables $T_1, .., T_M$ that are joined in a shared table. Identifying attributes are used to link records, but are not shared. Let the set of shared attributes be QI and a single attribute be denoted as $\alpha_i \in QI$. Each source table T_j provides a set of attributes AS_j, whereas $QI = \bigcup_j^M AS_j$. Considering the threat scenario of section 2, we investigate the number of data twins for each source table separately. Therefore, we will have to determine for every T_j the set of its distinct attribute value combinations. Each table T_j stores a set of tuples $t(id, \nu_1, .., \nu_n)$ where id is the unique tuple id and n is the number of table attributes. An equivalence class $[class_i]$ may be generated for each attribute value combination $(\nu_1, .., \nu_n)$ and has an associated cardinality c_i counting the number of its elements. Equivalence classes are needed to check the k-anonymity constraint for a certain attribute value combination. Further details are given in section 4.5.
Attribute values are transformed to more general values to increase the number of data twins. Each attribute has an associated generalisation hierarchy used for recoding. We distinguish between a **dimension hierarchy** and a **member hierarchy** [SEZ06]. While the member hierarchy is used to transform record values to generalised values, the dimension hierarchy is used for evaluation of generalisation steps.

4.1 Dimension hierarchy and member hierarchy

A dimensional hierarchy is composed of generalisation levels. Let the lowest generalisation level be L_0, at that level all attribute values remain unchanged. Let L_{max} be the maximal generalisation level for a certain α_i. At this level, all values of an attribute α_i are combined to a single generalised value. The highest generalisation level corresponds to the suppression of the attribute. Generalsations are always associated with an **information loss**, since an attribute value is transformed to a less specific value. While the dimensional hierarchy measures the information loss, the member hierarchy models the transformation from specific values to generalised values for each attribute. Each attribute α_i ($\alpha_i \in$ *QI*) takes its values from a certain domain, whereas an attribute domain may be *nominal*, *ordinal* or *numeric*. Attributes with nominal or ordinal data types are typically generalised using a generalisation hierarchy of parent-child re-

lations. Attributes of numeric domain type may be generalised along interval-based hierarchies.

4.2 Information loss

By applying a sequence of consecutive generalisation steps, the information loss is accumulated. However, the data quality decrease is not necessarily equal for each generalisation step along the hierarchy. For instance, consider the generalisation of patient age values 10, 12, 16, 20. If 10 and 12 are generalised to [10-15] and 16 and 20 are generalised to [16-20], less information is lost than by generalising all four values to [10-20]. Another example from the medical research domain is the generalisation of the tumour staging PT that specifies the size of the tumour and its local spread. Stage PT is a numeric value in the range from 1-4 combined with prefixes for subclassification, such as 'a', 'b', 'c' or 'mic'. The set of transformation rules given in table 6 represents the member hierarchy for attribute PT.

VALUES		AGGREGATED VALUE
$1mic, 1a, 1b, 1c$	$\mapsto$	1
$1, 2$	$\mapsto$	$[1-2]$
$4a, 4b, 4c, 4d$	$\mapsto$	4
$3, 4$	$\mapsto$	$[3-4]$
$[1-2], [3-4]$	$\mapsto$	$*$

Table 6: Member hierarchy

By applying two generalisation steps value classification $1a$ is transformed into $[1-2]$. Staging PT is determined by the size of the primary tumour. While for instance in breast cancer, tumours with a greatest dimension of 2 cm are classified as T1, tumours with a greatest dimension of minimal 2 and maximal 5 cm are classified as T2. Thus, transformation from $(1mic, 1a, 1b, 1c) \mapsto 1$ is of different quality than transformation $(1, 2) \mapsto [1-2]$, since the difference among the subclassifications $1mic, 1a, 1b$ and $1c$ is smaller than the one among the classifications 1 and 2.
We allow to specify customised **information loss quantifiers** along the generalisation hierarchy. The information loss is 0 at L_0 and 1 at L_{max}. Information loss quantifiers $\varphi_{(j)}$ are assigned to all intermediate generalisation levels L_j, $0 < j < max$. A dimension hierarchy for attribute α_i is denoted as set of edges, whereas an edge models the information loss up to generalisation level k. $edge(\alpha_i, L_{k+1}, L_k, \varphi)$.

4.3 Priorities and Limits

Priorities are used to specify the relative importance for attributes and their granularity. In some cases exact values for a specific attribute may be favoured while the generalisation degree of others is negligible. Attributes with lower priorities are generalised first while attributes with higher priorities are only generalised when no other solution may be found. In some cases, attributes should be generalised only up to a certain degree or not transformed at all. Otherwise,

their values become useless for an application domain. Therefore, it's crucial to allow definition of limits for information loss, someone is willing to cope with.

4.4 Request

For a particular analysis, a request is made on the shared table. A request is a query consisting of a set of selections and a projection, specifying the table attributes which have to be derived from the shared table to perform the analysis. It may be defined as $R(\Pi_A, \sigma_{(\alpha_i\ op\ cri)}*)$, where $A \subseteq QI$, op is a comparison operator $(=, \neq, <, >, \leq, \geq)$, and cri is a value of the domain of attribute α_i. Selections restrict the range of values and projection allows to dismiss attributes which are not relevant. Clearly dropping attributes increases the likelihood of data twins. The result of the request query is then checked for k-anonymity and, if necessary, generalised by the anonymisation algorithm outlined below. If it is not possible to achieve the specified anonymity, no data will be released.

4.5 Anonymisation steps

Our algorithm attempts to find an appropriate anonymisation for a specific data request. The goal is to find a solution that is as close to data quality demands as possible. Thus, a depth-first search looks for a solution along the top-ranked search path. We want to emphasize that alternative solutions could be found by backtracking and ranking of alternative paths. For simplicity, we focus on the search along the top-ranked path in this work. Before starting with anonymisation the data request is examined. Since k-anonymity is to be accomplished in each source table T_j separately, anonymisation is composed of subsolutions for every data source. If a source table has none of the projected attributes, anonymisation is trivially fulfilled. The following three steps are executed for every involved data source.

1. **Preprocessing:**
   A data source T_j is sharing attribute set AS_j. We filter all projection and selection criteria of request $R = (\Pi_A, \sigma_{(\alpha_i\ op\ cri)}*)$ that are related to attributes of T_j. Hence, we deduce the following subrequest $R_j = (\Pi_{Sub}, \sigma_{(\alpha_l\ op\ cri)}*)$, $Sub = A \cap AS_j$ and $\forall\ \alpha_l \bullet \alpha_l \in A \cap AS_j$. We apply the filtered projection and selection criteria immediately on the shared table and retrieve a result set of tuples RS. For each distinct value combination of RS, an equivalence class $[class_j]$ with associated cardinality c_j is created. Finally, we store those equivalence classes in set T_{Anon}.

2. **Initialisation of algorithm:**
   Selection criteria do have to be considered in the anonymisation process. Any selection criteria limits the generalisation potential of an attribute. Consider the following example: A set of patients having colon cancer that is staged up to 1 should be released. The localisation (colon) has been specified exactly, that is, it cannot be generalised any further, since the requesting users knows the exact value. Staging 1 is a generalised value of (1a, 1b, 1c, 1mic). Hence, staging t is generalisable up to value 1. Selection criteria may determine the maximal limits of generalisations. If a user-defined limit and a selection criteria limit is specified for the same attribute the smaller one is taken.

Algorithm *Anonymisation Algorithm*
Input: $T_{Anon},\ priorV\,[n]\,,\ limV\,[n]\,,\ levelV\,[n]\,,\ k$
Output: $Sequence\ of\ generalisation\ steps \rightarrow GS$

```
1.  GS = [ ]
2.  while T_Anon not fulfils k-anonymity
3.     do minimum = ∞
4.        α_gen = null
5.        for all α_i ∈ QI
6.           do cl = level [α_i]
7.              edge(α_i, L_pl, L_cl, φ)
8.              if pl < limV [α_i])
9.                 then cost = φ * priorV [α_i]
10.                      if cost < minimum
11.                        then minimum = cost
12.                             α_gen = α_i
13.       T_Anon = generalise(T_Anon, α_gen)
14.       append(GS, (α_gen, pl))
15.       level [α_gen] = level [α_gen] + 1
16.       T_Anon = merge(T_Anon)
17. return GS
```

A pseudo-code description of the anonymisation algorithm is given below (*Anonymisation Algorithm*). The following input parameters are required: T_{Anon} is the set of equivalence classes that violates the k-anonymity constraint. The k-Anonymity parameters is specifying the number of requested data twins. However, that parameter has to be greater than the minimal value claimed by a general security policy. A priority vector for all α_i is specified as $priorV\,[n]$. The priority values are in the range $[0, .., 1]$, whereas the most import attribute has the highest priority value and all differences between any two consecutive priorities values are equal. The generalisation limits are stored in vector $limV\,[n]$ and a level vector ($levelV\,[n]$) stores the current generalisation levels for all attributes - initially, all level values are set to 0.

3. **Anonymisation of records:**
An anonymisation is accomplished by a sequence of n generalisation steps ($GS_1 \rightarrow GS_2 \rightarrow .. \rightarrow GS_n$). Each generalisation step ($GS_i, 1 \leq i \leq n$) transforms all values of a certain attribute α_g to more general values using the member hierarchy structure described in section 4.1. We are searching in a multidimensional space that is built from the generalisation hierarchies of all attributes. Let $Lmax_{(i)}$ be the maximal generalisation level of attribute α_i, then the number of all possible anonymisation solutions is $\prod_{i=1}^{n} Lmax_{(i)}$. We prune the search space by including user-defined limits and we evaluate search paths by taking information loss quantifiers and user-defined priorities into account. We calculate a quality measure by multiplying information loss quantifiers that are specified a priori in relations along the dimension hierarchy ($edge(\alpha_i, L_{pl}, L_{cl}, \varphi)$) with priorities that are specified by the user in context of his current data request. Attribute α_g is chosen by evaluating all possible generalisations of all attributes. α_g is the attribute that has the minimal weighted information loss after being transformed to the next generalisation level. Function $generalise(T_{Anon},\ \alpha_{gen})$ transforms all records of

T_{Anon} by replacing all attribute values of α_g with generalised values. In the next processing step, $merge(T_{Anon})$, a new equivalence class set is created by grouping distinct attribute values of T_{Anon} to equivalence classes. If each equivalence class has at least an associated cardinality of k, a solution has been found and the sequence of necessary generalisation steps is returned.

4. **Release of records:**
After the required generalisation steps have been determined for each source table, releasing of records may be initiated. We apply the projection and selection criteria on the shared table and transform the result set according to the generalisation steps.

5 Use Case

We have tested the following use case on pathological findings of the Institute of Pathology Graz. We focused on a staged mamma carcinoma data set of 16,417 cases.

5.1 Attribute Hierarchies

Table 7 represents an extract of the generalisation hierarchies for attributes staging PT, staging PN, staging PM and residual tumour description R. Columns 'Values' and 'Agg.Value' (aggregated Value) specify relationships of the member hierarchy. The dimension hierarchy levels are given in columns L_i and L_{i+1} and the information loss between the levels is stored in φ. Information loss quantifiers strongly depend on the semantics of attribute values and the context of use and have to be evaluated by users. The TNM cancer staging scheme may include different classification values for various tumour types [Fle97]. For simplicity, we focus on classification of mamma carcinoma in our examples. Consider the classification of tumour size by staging PT: a staging value of 1 classifies tumours up to the size of 2 cm, while tumours with a size between 2 and 5 cm are staged with value 2. The subcategories for staging 1 are $(1a, 1b, 1c)$ used for tumours of sizes (in cm) $(> 0.5, [0.5 - 1.0], [1.0 - 2.0])$. A generalisation from all subcategories to 1 has a different data quality decrease than a generalisation from values 1 and 2 to the next aggregated value $[1 - 2]$. Hence, the estimated information loss is smaller in the former case than in the latter. We estimate the information loss for attribute PT between level L_0 and L_1 with 30 %. Attributes R and PM are not generalisable hierarchically, but they may be suppressed.

5.2 Detailed example use case

Although tumours may be staged identically, they may show different behaviour in the following course of disease. We want to analyse gene expression profiles of mamma carcinoma of size 1 (staging PT) having a well-differentiated grading (G=1). Additionally, the survival periods should be exported to allow further grouping. Altogether, the following attributes are needed: PT, PN, PM, G, R from findings table, survival period in days from survival data and finally the corresponding gene expression profiles of the selected cases should be linked. We may derive the following request from the textual description:
$R_{(Findings)} = (\pi_{(PT,PN,PM,G,R)}, \{\sigma_{(Localisation='Mamma')}, \sigma_{(PT=1)}, \sigma_{(G=1)}\})$. We do not have to

ATTRIBUTE	VALUES	AGG.VALUE	φ	L_i	L_{i+1}
PT	1a, 1b, 1c, 1mic	1	0.3	0	1
R	X, 0, 1, 2	*	1.0	0	1
PN	1mi, 1a, 1b	1	0.5	0	1
PM	X, 0, 1	*	1.0	0	1
...	...	...	...	...	...

Table 7: Attribute hierarchies

anonymise the corresponding records of survival data table and gene expression table as mentioned in 3. Now we apply all steps described in section 4.5:

1. **Preprocessing:**
   By examining the values of the selection criteria, we discover that 'PT=1' is a generalised value. Hence, our search pattern for PT includes all child values of 1 → [1a,1b,1c,1mic]. We search for entries in the shared table that may be matched with the following pattern (Localisation=[Mamma], PT=[1a,1b,1c,1mic], N=[*], M=[*], G=[1], R=[*]). After applying projection and selection criteria, a set of 107 equivalence classes for 313 different entries is identified, whereas 67 classes have a cardinality of 1.

2. **Initialisation of algorithm:**
   Three attributes (Localisation, PT, G) have been used in selection criteria. Localisation is not generalisable, since 'Mamma' is not a generalised value, as well as grading value 1. Staging PT may be generalised up to generalised value '1'. The attribute priorities are specified as follows: R → 0.20, PM → 0.40, PN → 0.60, PT → 0.80. Further, a minimal k-anonymity of 2 is to be achieved.

3. **Anonymisation of records:**
   We calculate the weighted information loss for each generalisable attribute when being transformed to its next generalisation level. It is 0.20 (= 0.20 * 1.0) for R , 0.24 (= 0.80 * 0.3) for PT, 0.30 (= 0.60 * 0.5) for PN and 0.40 (= 0.40 * 1.0) for PM. Attribute R is chosen to be generalised ($GS_1 = (R, 1)$). As R values are only generalisable up to a general all value (*), the attribute is suppressed. The resulting new set of equivalence classes has 98 members and 48 classes have a cardinality of 1. In the next anonymisation step attribute PT is selected and all PT values are generalised ($GS_2 = (T, 1)$). We retrieve a set of 16 equivalence classes, whereas each class has a cardinality greater than 1.

4. **Release of records:**
   Finally we apply the selection and projection criteria on the shared table, execute generalisation steps GS_1 and GS_2 on the result set, and release the anonymised data together with the related gene expression profiles and survival periods.

Through these four steps, the data stemming from different sources of our virtual data warehouse have been integrated, anonymised and are released to the requester in form of a data mart. The users now can perform analytical queries or data mining procedures on this data mart, or

use the table as input to statistical tools. Tracking down the the released data to an individual person is not possible through this method.

5.3 Impact of parameters

The result set produced by the algorithm is strongly influenced by the parameters the user specified. Thus, different data quality demands of medical studies may be considered. If the input parameters (priorities and limits) of our use case example are changed a different anonymisation is created. In table 8 a summary of the input and output parameters of the original use case settings is given. Attribute R is suppressed, attribute PT is generalised up to generalisation level 1 and the transformed result set contains 16 equivalence classes satisfying the k-anonymity constraint of 2. If the priorities and limits are changed as shown in table 9 the same data request would be anonymised slightly differently: attributes PN and PT are generalised up to level 1, but no attribute is suppressed and 20 equivalence classes are created.

ATTRIBUTE	PRIORITY	LIMIT	TRANSFORMATION
PT	0.80	1	$0 \rightarrow 1$
PN	0.60	No	No
PM	0.40	No	No
R	0.20	No	Suppression
Equivalence classes:	16		

Table 8: Setting A

ATTRIBUTE	PRIORITY	LIMIT	TRANSFORMATION
PN	0.80	No	$0 \rightarrow 1$
R	0.60	No	No
PM	0.40	0	No
PT	0.20	1	$0 \rightarrow 1$
Equivalence classes:	20		

Table 9: Setting B

6 Conclusion

We presented an approach for flexible anonymisation of shared data stemming from different sources with the aim of minimizing information loss and maximizing use of data for research without compromising the privacy of patients. The major contribution of this approach is the individualzation of the concept of an information loss, when data are generalised. This allows at least the attempt to provide the data of main interest for a particular analysis in great detail while achieving the deired anonymity by sacrificing other granularity of the values of other attributes. We demonstrated the effectiveness of the approach in a use case, applying the algorithm to real world data. Various ways of improving the technical implementation of our strategy as well as an exhaustive evaluation of the performance of the algorithm are subject of ongoing work.

References

[BIO] A biobank for the advancement of medicine. *http://www.bioresource-med.com*.

[Fle97] Cooper J. S. Henson D. E. Hutter R. V. P. Kennedy B. J. Murphy G. P. O'Sullivan B. Yarbro J. W Fleming, I. D. Cancer staging manual, sixth edition. 1997.

[FWY05] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, pages 205–216, 2005.

[iA] GEN-AU Genomeresearch in Austria. *http://www.gen-au.at/english/content.jsp*.

[LDR05] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD Conference*, 2005.

[LDR06] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering*. IEEE Computer Society, 2006.

[PS98] L. Sweeney P. Samarati. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, 1998.

[SEZ06] Konrad Stark, Johann Eder, and Kurt Zatloukal. Priority-based k-anonymity accomplished by weighted generalisation structures. In *DaWaK 2006: Proc. of the 8th International Conference on Data Warehousing and Knowledge Discovery*, Lecture Notes in Computer Science, Volume 4081. Springer Verlag, 2006.

[WFD05] K. Wang, B. C. M. Fung, and G. Dong. Integrating private databases for data analysis. In *ISI 2005: Proc. of the 2005 IEEE International Conference on Intelligence and Security*, Lecture Notes in Computer Science, Volume 3498. Springer Verlag, 2005.

[WYC04] Ke Wang, Philip S. Yu, and Sourav Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, 2004.

[XWP+06] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 2006.