ESGI77 REPORT

MATEMATYKA STOSOWANA TOM 12/53 2011

Mathematical techniques for the protection of patient's privacy in medical databases

Problem presented by

Marcin Węgrzyniak, Paweł Gaza

Centre of Health Information Systems (CSIOZ[1])

Executive Summary

In modern society, keeping the balance between privacy and public access to information is becoming a widespread problem more and more often. Valid data is crucial for many kinds of research, but the public good should not be achieved at the expense of individuals.

While creating a central database of patients, the CSIOZ wishes to provide statistical information for selected institutions. However, there are some plans to extend the access by providing the statistics to researchers or even to citizens. This might pose a significant risk of disclosure of some private, sensitive information about individuals.

This report proposes some methods to prevent data leaks. One category of suggestions is based on the idea of modifying statistics, so that they would maintain importance for statisticians and at the same time guarantee the protection of patient's privacy.

Another group of proposed mechanisms, though sometimes difficult to implement, enables one to obtain precise statistics, while restricting such queries which might reveal sensitive information.

[1] Acronym derived from the Polish name of the institution concerned.

Report author

Marcin Przybyłko (University of Warsaw)

Contributors

Orest Dorosh (Institute of Atomic Energy POLATOM)
Joanna Giemza (University of Warsaw)
Tomasz Górski (Military University of Technology)
Anna Kortyka (Polish Academy of Sciences)
Dorota Kowalska (Warsaw University of Technology)
Tomasz Kraśnicki (Wroclaw Medical University)
Paweł Szlendak (Warsaw University of Technology)
Przemysław Świtalski (University of Warsaw)
Michał Warchoł (Jagiellonian University in Krakow)

ESGI77 was organised jointly by
Systems Research Institute of the Polish Academy of Sciences
Institute of Mathematics of the Polish Academy of Sciences
Oxford Centre for Collaborative Applied Mathematics

and it was supported by
Sygnity S.A.
Industrial Development Agency Joint Stock Company

under the honorary patronage of
The British Embassy in Poland

Contents

1 Introduction

1.1 Background

(1.1.1) With the development of information technology, vast amount of data is being accumulated in various systems. Data gathered by some companies and organizations is not only stored, but also published. This leads to an increase in the volume of publicly available data, and consequently to the growing risk of privacy violation.

(1.1.2) Privacy is important, however, it may be understood in various ways, so that its meaning may not be the same for all people. This is particularly important in the case of medical databases as privacy of many groups is at risk.

(1.1.3) The database may be accessible either to the public or to the limited audience (e.g. medical doctors), about which we can make additional assumptions. Therefore, various target groups should be taken into consideration.

1.2 Problem statement

(1.2.1) The main idea is to develop techniques and methodology to assure privacy protection in publicly available statistical databases. A desired solution ought to reduce the risk of the private data leak from statistical databases. It should also help database owners to make appropriate decisions, such as:

- Which form should available data take (fixed statistical data sheet, dynamically generated statistics, etc.)?
- What should be the scope of available statistics?
- What additional parameters should be taken into account?

(1.2.2) Furthermore, the proposed solution should meet the following criteria:

- understandability (as a criterion of evaluation or as the information presented to the public),
- simplicity of implementation,
- verifiability.

(1.2.3) The project work will focus on the following aspects:

- possible measures of privacy in terms of medical databases, with emphasis laid on statistical databases,
- analysis of methods of data protection adequate for statistical databases,
- suggestions for further exploration.

2 Problem analysis

2.1 How to determine sensitivity of data

(2.1.1) It seems indispensable to establish which information is sensitive. According to the Polish law, every medical data about an individual is sensitive. Still, it pertains only to the exact information about a person. For example "he/she has a lung cancer" or less precisely - "he/she has a cancer" or maybe even more generally - "a person is sick". Probably nobody minds a statement: "this person earns between 2000 PLN and 20000 PLN", but what about: "this person earns between 4000 and 5000 PLN"? A suitable choice of the interval is crucial. Privacy is a very subjective issue, so this measure could be acquired in the same way as QALY([2]), with surveys for patients.

2.2 Architecture sensitivity

(2.2.1) The architecture of the computer system is presented in the feasibility study provided by the CSIOZ. The following chief aspects of security are covered in this document: authorization, authentication and accountability. Furthermore, it is necessary to design the security of the system in accordance with Sherwood Applied Business Security Architecture (SABSA) [4]. This methodology should cover the majority of problems with security architecture. Therefore, we will restrict our further analysis only to security aspects of statistical databases.

(2.2.2) Security can be increased thanks to the distinction drawn between databases for medical and statistical purposes.. The Study Group has proposed the model of the hierarchical medical database with the limited access to every part of DB.

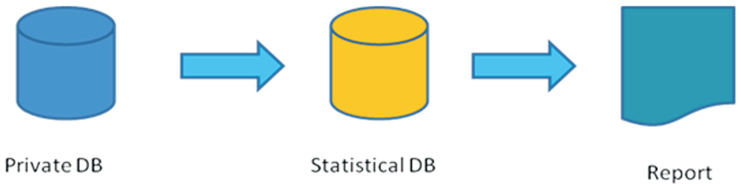

Figure 1: Model of creating statistical reports by CSIOZ.

(2.2.3) The model is presented in the Figure 1. The medical database consists of PrivateDB that contains whole information about patients

[2] Quality Adjusted Life Year a measure of quality of life, used in determining the cost-effectiveness of medical procedures.

(including personal data) and StatisticalDB comprising data from statistical reports and the collection of reports. Every part of medical database has a different level of accessibility: PrivateDB may be accessed by a personal data owner (a patient) and a doctor with direct permission given by the data owner. StatisticalDB can be queried by special permission. A part of the report collection can be available in the public domain.

2.3 Security of statistical databases (SDB)

(2.3.1) Another aim of security is to prevent the disclosure of sensitive data by means of inference. First, we have to specify when a statistical database is secure. It has to be indicated that disclosed data is almost always something more than data released, because additional information can be gained by the inference. Let us develop a model of data in SDB [1].

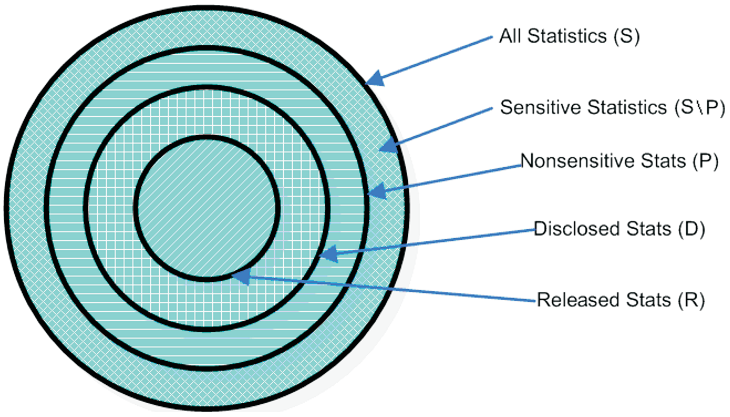

Figure 2: Fragmentation model of statistical information. S - set of all statistics, $P \subseteq S$ - set of statistics classified as nonsensitive, $R \subseteq S$ - set of released statistics, D set of statistics disclosed by R (including the statistics in R). Statistical database is secure if and only if $D \subseteq P$ (R discloses no sensitive statistics).

(2.3.2) There are some risks that could lead to the disclosure of sensitive information about an individual. These include:

- too narrow data set + additional knowledge,
- comparison of statistics with similar data set,
- diachronic comparison of changes in statistics.

(2.3.3) We have to take into account the following specific conditions:

- the specific set of statistics of the CSIOZ system needs to be provided to the defined institutions,

- data input for statistical data input by forms,
- the existing model of the CSIOZ system.

3 Security methods for statistical databases (SDB)

(3.0.1) In this chapter we present and discuss methods that aim to prevent the disclosure of sensitive data. These techniques are supposed to clear statistics of personal data and hinder extracting personal data by comparing database queries. The whole chapter is a synthesis of the available research in the subject, especially [1], [2] and [3].

3.1 Anonymizing data

(3.1.1) In order to develop methods for avoiding the disclosure it is necessary to define first in what form statistical data will be published. The most popular types include: tables of pre-defined statistics (frequency, magnitude), micro-data files or an on-line query system.

(3.1.2) A micro-data file is a set of records, each concerning a specific individual or an incident. It can contain identifiers (personal data), such as name, address, Social Security number([3]) or National Identity Card number([4]). Such information must be removed (depersonalizing/anonymizing data) before publishing in order to protect privacy of individuals.

(3.1.3) In the case of the CSIOZ system most of the statistics are in the form of tables of frequency [5] and the table of magnitude data [8]. In our opinion, reports derived from these data can contain sensitive information, because micro-data contain values of attributes for individuals.

(3.1.4) One of the methods for improving security of statistical databases is to make an appropriate choice about the data which is to be stored and processed in statistical database. For instance, dates of patient's admission and discharge from hospital can be presented in a different way by indicating the number of days spent in hospital and the month of arrival. It is more general and exact dates are not necessary for the purpose of the majority of statistics. This applies also to other attributes – we may not need dates of birth or it could be rounded off to years or even decades.

([3]) PESEL

([4]) NIP

3.2 Methods of restricting queries

(3.2.1) **Maximum order control**
Maximum order control [1] is a method which restricts the number of attributes that can be used in the query. Allowed combinations of attributes can be obtained by means of the lattice model [6]. For every element you check, you should set a partial order of inclusions between the subsets of the attribute set, provided it is secure to share. One gets all combinations of attributes that could be used safely in queries, as well as such that should be forbidden. Furthermore, one gets the number of attributes that could be used in the query (if n is a minimum number of attributes in a non-safe query then $n-1$ is the number of attributes that could be safely used in the query).

(3.2.2) Theoretically, this method is perfect, but practically it has a very narrow application. The verification of safety of attributes is computationally complex (exponential time) - every combination of attribute values has to be checked if the result is an n-element set (where number n depends on a desired safety level). Furthermore, every combination must be verified on a real database containing millions of records.

(3.2.3) The method could be used while creating annual databases with data that would not be modified. Then, it is possible to do some computations (even if it takes a month) and share a secure database. Still, the problem arises when other databases contain information of the same type and time period (or they refer to such information), because it could reduce safety. It should be assumed that every database that has once been published (especially in a digital form) should be considered as constantly available (it might have been copied before the access was denied).

(3.2.4) **Minimum query-set size**
This is one of the methods in which an agency does not release statistics considered sensitive. In this case, a statistic is said to be sensitive if it is calculated on a set of cardinality smaller than some specified value. This value (called "threshold value") varies for different types of data. It is usually determined by the sensitivity of the information that the agency is allowed to publish and by the precision required to carry out disclosure (see Section 2.1).

(3.2.5) **Complementary suppressions method**
This method supports mechanisms proposed earlier by supplementing previously disregarded aspects. If only one cell is suppressed and it is allowed to ask for the marginal total then information is

not sufficiently protected (the exact value of the cell can be calculated by a simple subtraction). In this case, other, non-sensitive cells also have to be suppressed (complementary suppressions). In all cases except for the simplest ones it is nearly impossible to prove that preventing the specific set of cells has provided adequate security.

(3.2.6) **Granularization**
We propose this method as an effective rule of disclosure limitation. This solution, which might resemble "rolling-up categories", is related to the increase of statistical sample in the statistical query. For example, instead of giving an answer to the query: "how many people died of cancer in Warsaw last month" (it might be less than 5), we could respond: "last month 54 people died of cancer in the Masovian Voivodship". Obviously, if the number revealed is still dangerously small, the sample size should be increased further on. In this way, we can greatly reduce the risk of revealing too narrow data set.

(3.2.7) A disadvantage of granularization is the lack of the precise response for a given query, which becomes especially problematic, when answers to queries with both more global and more local sample size are the same (the same number of people suffered from the same disease in Warsaw and in Poland).

3.3 Methods of adding noise to the statistics

(3.3.1) There are some methods of modifying results of queries in a way that prevents inference, but does not change statistical importance:

- Adding random noise
- Controlled rounding method
- Random rounding

One may consider two of the possible methods of adding noise [1,2]: creating an artificial database (AD) (which will contain perturbed data) or modifying data on-line. The first method requires additional storage space and recalculating AD after each update of the original data, while the second one may cause greater time overhead (when queries are to be computed), but it is more flexible.

(3.3.2) **Random noise**
Adding random noise on-line is a method of data perturbation which provides accurate statistics computed on the slightly modified data. One may substitute real values stored in database using Back's formula (1):

$$x'_i = x_i + z_{1i}(x_i - \bar{x}) + z_{2i}, \tag{1}$$

$$\bar{x}_c = \frac{\sum x_i}{|C|}, \tag{2}$$

$$E(z_{1i}) = 0,\ Var(z_{1i}) = 2a^2,\ E(z_{2i}) = 0,\ Var(z_{2i}) = \frac{2a^2(\bar{x}_c - \bar{x})}{|C|}, \tag{3}$$

where:
x_i is a single value in a database,
C is a set on which we compute a query,
z_i is an independent random variable.

(3.3.3) This method provides additive and multiplicative noise, which increases security. Parameter a tells how large the dispersion of the noise is. The larger a, the more questions are required to disclose private information (estimate original value with small error). Unfortunately, an increase of parameter a causes a decrease in the accuracy of computed statistics. This may be observed in Figure 3.

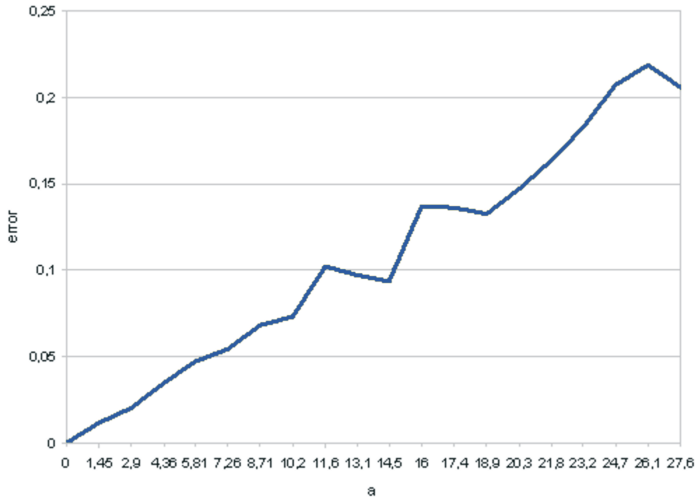

Figure 3: Mean value of relative error is plotted for different value of parameter "a". Values of parameter "a" are in range $[0, v]$, where $v = 27.59$ is the variance of test data.

(3.3.4) Artificial data creation

Assume that our micro-data (data in statistical database) has some specific probability distribution P. If we were able to find the approximation of this distribution, we could use it to distort the queried statistics in the following way:

- estimate the number of records k contributing to the queried statistic,

- sample k samples from the approximation of P to obtain a new set of synthetic data,
- compute the queried statistic on the synthetic data.

(3.3.5) The distortion of the resultant statistic depends on the approximation of P. The better the approximation, the more accurate the statistics are. However, the approximation should neither be too exact nor too inaccurate. The computed statistic should remain meaningful and at the same time contain the sufficient level of distortion so its exact value is not disclosed.

(3.3.6) Finding the approximation to P may be hard. The complexity of this problem depends on the complexity of data i.e. the number of attributes, the statistical dependencies of attributes in data, the set of values the attributes takes on and their type (continuous and discrete). For further research we propose Bayesian Networks [7] as the framework for modeling of complex probability distributions.

(3.3.7) **Controlled rounding method**

In this method, the answer for a query q is rounded with a specific function $r(q)$, which preserves the following property: "If $C_1, ..., C_m$ are disjoint query sets, $C_{m+1} = C_1 \cup ... \cup C_m$, and $q_i \in C_i$ then $r(q_1) + ... + r(q_m) = r(q_{m+1})$". To achieve that property for a given integer $p \geq 1$, the method finds an optimal controlled rounding $r(q)$ that minimizes the following objective function:

$$z_p = \sum |q - r(q)|^p. \tag{4}$$

(3.3.8) The problem of finding an optimal controlled rounding can be expressed as a capacity-constrained transportation problem and thereby solved using standard algorithms. The technique is particularly well-suited for protecting tables of relatively small frequency counts.

(3.3.9) **Random rounding method**

In this method cell values are rounded, but instead of using standard rounding conventions a random decision is made as to whether they will be rounded up or down according to:

$$r(q) = \begin{cases} q & \text{if } d = 0 \\ q - d & \text{with probability } 1 - p \text{ (round down)} \\ q + (b - d) & \text{with probability } p \text{ (round up)} \end{cases} \tag{5}$$

A table prepared using random rounding of statistics q when $p = \frac{d}{b}$ is vulnerable to attack in query processing systems. If a query q is asked many times, its true value can be deduced by averaging the rounded values. The sum of the rounded statistics for disjoint query sets can differ from the rounded statistic for the union of the sets, which can be overcome by controlled rounding method.

3.4 Accuracy of reports

(3.4.1) Transformations of statistical data described in the previous subsections can dramatically distort information. In this subsection we present methods to estimate the accuracy of reports.

(3.4.2) Average absolute distance per cell for queries in a perturbed statistical database [9]:

$$AAD(D_{orig}, D_{pert}) = \frac{\sum_i |D_{pert}(X_i) - D_{orig}(X_i)|}{n_i}, \tag{6}$$

where:
D - frequency distribution,
X_i - cell (value of attribute),
n_i - number of cells.
It can be used to check the accuracy of statistics before releasing a report.
One may consider defining some threshold t such that if
$AAD(D_{orig}, D_{pert}) > t$ then create another perturbation with smaller variance.

(3.4.3) In order to measure information loss, caused by query restriction methods, one may use *Censorship* measure given by equation (7):

$$\textit{Censorship} = \frac{|B_s \cap P|}{|P|}, \tag{7}$$

where:
B_s - set of blocked statistics,
P - set of non-sensitive statistics.
When *Censorship* $= 0$ then database provides access to every non-sensitive statistic and there is no data loss. When *Censorship* $= 1$ then we do not provide any non-sensitive data. This measure does not say anything about a private data leak.

(3.4.4) In order to measure size of private data leak one can use *Recall* measure, given by equation (8):

$$\textit{Recall} = \frac{|B_s \cap (S \backslash P)|}{|S \backslash P|}, \tag{8}$$

where:
B_s - set of blocked statistics,
P - set of insensitive statistics,
S - set of all statistics.
When *Recall* $= 1$ then every sensitive statistic is blocked and the base is secure. This measure does not say anything about the information value of the restricted database.

(3.4.5) Another possible measure is the number of queries/questions necessary to disclose sensitive data. This constitutes a simple way to describe the usefulness of a particular security method. If time([5]) of disclosing data, protected by particular method is higher than time of its availability, then this data may be assumed protected, subsequently the method can be perceived as appropriate.

4. Conclusions

(4.1.1) According to the documents provided by the CSIOZ representatives (feasibility study), architecture aspects of database security are covered using the SABSA model [4]. If developed according to current plans, the architecture will comply with the standards. Furthermore, security of the system should be increased by creating a special statistical database that contains statistics acquired from the main database.

(4.1.2) On the other hand, security of static data ought to be investigated. Some of the forms used to gather information (e.g. mz/szp-11, mz/szp-11b, mz/N-1A [11]) contain detailed personal data which allows to reveal personal identity. Such information should be secured by means of at least one of the methods proposed in this report: adding noise to the statistics, restricting queries that are possibly dangerous (methods of specification are proposed in the report).

(4.1.3) Should the functionality of the CSIOZ systems be broadened (more complex statistics, easier forms of requesting statistics), mathematical methods can help to secure sensitive information. We recommend blurring (adding noise) of the released data and the strict control of the flow of precise data (analysing queries from institutions requiring precise information).

5. Further research

(5.1.1) In order to share precise information in a safe manner, one can compare answers to the requested query with all information that is public, and subsequently adjust them to the appropriate protection method. This technique has one big disadvantage, namely the computational complexity. For this reason, computational costs are acceptable only for a small amount of information. This drawback may be overcome if this method is applied in such institutions that

[5] Time is estimated here from number of necessary queries and delivery time

require precise data, in which case one can track every piece of information that has been released. The practical use of this method for the CSIOZ medical database may be a subject of further research.

(5.1.2) Another area of research is the usability of statistical databases that after creation would not be modified and would be secured by the maximum-order control. In this way, almost everyone might be able to conduct a research on medical statistics without the risk of disclosure. It should be also investigated, how this would affect future information policy, because each further statistic published by the CSIOZ could lead to leaks of sensitive information.

Bibliography

[1] Denning, D.E., *Cryptography and data security*, Eddison-Wesley, 1982.

[2] Anderson, R.J., *Security Engineering: A guide to Building Dependable Distributed Systems*, Wiley, 2001.

[3] Federal Committee on Statistical Methodology, *Report on Statistical Disclosure Limitation Methodology*, 2005,
http://www.fcsm.gov/working-papers/SPWP22_rev.pdf,
(available 2010.10.01).

[4]
http://www.sabsa-institute.org/the-sabsa-method/the-sabsa-model.aspx
and
http://www.sabsa-institute.org/the-sabsa-method/the-sabsa-matrix.aspx.
(available 2010.10.01).

[5] CSIOZ Biuletyn Statystyczny Ministerstwa Zdrowia, Warszawa, 2009.

[6] Denning, D.E., A Lattice Model of Secure Information Flow, *Communication of the ACM.* Volume 19 Number 5, pp. 236–243, 1976.

[7] Castillo, E., Gutirrez, J.M., Hadi, A.S., *Expert Systems and Probabilistic Network Models.* Monographs in Computer Science. New York: Springer-Verlag, 1997.

[8] http://csioz.gov.pl/statystyka.php forms used to gather information for statistical database (available 2010.10.01).

[9] Shlomo, N., Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility. *Journal of Privacy and Confidentiality 2*, pp. 73–91, 2010.

[10] Garey, M.R., Johnson, D.S., *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman and co., New York, 1979.

[11] http://isap.sejm.gov.pl/Download?id=WDU20030510444&type=2,
Rozporządzenie Prezesa Rady Ministrów z dn. 20 lutego 2003 r., Dziennik Ustaw Nr 51, poz. 444, pp. 3365–3381 (available 2010.10.01).

Akronim: HEALTH

Tytuł: Matematyczne techniki ochrony prywatności pacjentów w medycznych bazach danych

Opis tematu (sformułowanie firmy/instytucji zlecającej):

Wraz z rozwojem technik informacyjnych w ramach różnych systemów gromadzone są coraz większe ilości danych. Dane te nie tylko są gromadzone wewnątrz poszczególnych organizacji, ale także udostępniane. Prowadzi to do coraz większej ilości danych dostępnych publicznie, a w konsekwencji coraz większych zagrożeń dla prywatności.

Prywatność staje się z każdym dniem coraz istotniejsza. Okazuje się jednak, że może być ona różnie rozumiana i dla różnych osób oznaczać co innego. Jest to szczególnie istotne w przypadku medycznych baz danych, gdzie podmiotów, których prywatność może być zagrożona, jest wiele.

Bazy danych mogą być udostępniane zarówno publicznie jak i ograniczonemu gronu (np. lekarzom), co do którego być może można przyjąć pewne założenia. W rozważaniach należy uwzględnić potencjalnie różne grupy docelowe informacji.

Prace projektu powinny umożliwić zleceniodawcy podjęcie decyzji co do planowanych udostępnień danych/baz danych statystycznych, zarówno co do formy (np. gotowe zestawienia statystyczne, formularz umożliwiający dynamiczne generowanie statystyk), zakresu udostępnianych statystyk, jak i ew. zdefiniowanych w trakcie projektu parametrów.

Zaproponowane metody powinny spełniać poniższe kryteria.

1. Być proste w interpretacji (zarówno jako kryterium oceny jak i w przedstawieniu dla opinii publicznej).
2. Być łatwe w implementacji.
3. Być sprawdzone.

W związku z powyższym prace projektu powinny skoncentrować się wokół poniższych punktów.

1. Zdefiniowania pojęcia i możliwych miar prywatności w kontekście medycznych baz danych ze szczególnym uwzględnieniem statystycznych.
2. Zaprezentowania matematycznych modeli problemu adekwatnych dla udostępniania medycznych danych.
3. Analizy i zaprezentowania adekwatnych sposobów ochrony danych.
4. Wskazania ew. dalszych pól eksploracji.