International Journal of Medical Informatics (2004) 73, 599–606

ELSEVIER

International Journal of Medical Informatics

www.intl.elsevierhealth.com/journals/ijmi

Protecting patient privacy by quantifiable control of disclosures in disseminated databases

Lucila Ohno-Machado[a,*], Paulo Sérgio Panse Silveira[b], Staal Vinterbo[a]

[a] *Decision Systems Group, Division of Health Science and Technology, Brigham and Women's Hospital, Harvard/MIT, Boston, MA 02115, USA*
[b] *Medical Informatics, Department of Pathology, School of Medicine at the University of São Paulo, São Paulo, SP, LIM-01/HC-FMUSP, Brazil*

KEYWORDS
Consumer informatics;
Patient privacy;
Confidentiality;
Anonymization;
Predictive modeling

Summary One of the fundamental rights of patients is to have their privacy protected by health care organizations, so that information that can be used to identify a particular individual is not used to reveal sensitive patient data such as diagnoses, reasons for ordering tests, test results, etc. A common practice is to remove sensitive data from databases that are disseminated to the public, but this can make the disseminated database useless for important public health purposes. If the degree of anonymity of a disseminated data set could be measured, it would be possible to design algorithms that can assure that the desired level of confidentiality is achieved. Privacy protection in disseminated databases can be facilitated by the use of special ambiguation algorithms. Most of these algorithms are aimed at making one individual indistinguishable from one or more of his peers. However, even in databases considered ''anonymous'', it may still be possible to obtain sensitive information about some individuals or groups of individuals with the use of pattern recognition algorithms. In this article, we study the problem of determining the degree of ambiguation in disseminated databases and discuss its implications in the development and testing of ''anonymization'' algorithms.

1. Introduction

Privacy is a fundamental right and needs to be protected. For health care related information, there are regulations for disclosure [1–3]. These regulations were motivated by the public's concern of breaches of confidentiality that might result in discrimination embarrassment or economic harm. The recent progress in electronic medical record technology, the Internet, and the genomic revolution, together with media reports on violations of privacy, have generated increasing interest in this topic [4–7]. Since the use of electronic records has been steadily increasing, it makes sense to develop quantifiable tools to measure the potential for privacy violation in data sets that may be disclosed to third parties [8]. A common concern is that sensitive information is more easily available with the use of networked computers.

Protection of privacy of patients is not a new concern and it is possible to track references back to the 1970s. As stated by Annas, although ''the regulations under the Health Insurance Portability and Accountability Act of 1996 (HIPAA) regarding the privacy of medical records are new, the concept of using federal law to protect the privacy of medical records is not'' [9]. According this author, HIPAA

* Corresponding author. Tel.: +1 617 732 8543; fax: +1 617 739 3672.
E-mail address: machado@dsg.harvard.edu (L. Ohno-Machado).

1386-5056/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2004.05.002

regulations, although extensive, may be excessively complex and particularly designed to comply with necessities of large health insurance companies and to guarantee government access to patient information. However, it is important to have a way to control by whom and when patient information would be disclosed. In this article, we study the problem of quantifying the potential to uniquely identify an individual from a disclosed data set, with emphasis on data that are considered particularly sensitive.

It is useful to distinguish between access control and disclosure control. The former addresses access rights to the data, while the latter is concerned with what information is disclosed.

Unwanted disclosure of information can happen in many ways. Examples are breaches of security in institutional infrastructure, computer system security compromises, insecure transmission of information, and acts of disloyal employees. Unwanted disclosure can also be a part of wanted disclosure, in which information that should *not* be disclosed is mixed or embedded within information that should be. Biomedical research may be dependent on dissemination of data from health care sources. We will restrict our discussion to unwanted disclosure in such disseminated data. In this article, we will define a database to be a collection of records, each containing a set of fields called ''features'' that have associated ''values''.

Since total lack of disclosure is not realistic, current regulations require that ''minimal amount'' of information be given to a certain party. Practical application of such regulations requires anchoring in a well-defined context. Elicitation of (i) what ''privacy'' really means, (ii) the individual perspectives on different definitions of privacy, (iii) how to quantify both information needs and content, (iv) privacy standards, and (v) requirements that must be met in order to comply with these standards are prerequisites for such an anchoring.

Quantifying all aspects of a relevant context of data privacy is an impossible task, but it may be possible to anticipate certain contexts given prior experience and to develop algorithms to quantify the degree of ambiguation in a given data set. Current anonymization systems adapt the principle of minimal information loss while trying to guarantee against a predefined class of disclosures. Examples are hindering disclosure of sensitive cells in statistical data, and assuring that there is no possible way to uniquely identify an individual in a database. Methods to accomplish this goal utilize algorithms that include, among others, cell suppression, outlier removal, and generalization or ambiguation (see [10] for a brief review of anonymization systems).

Briefly, ambiguation systems concern themselves with three kind of values in relational databases, as classified by Chiang et al. [11]. They are (1) identifying, (2) easily-known and (3) unknown values. The first do not offer any challenge since they are unique to each patient, such as social security number, and should be deleted. Easily-known are values such as eye color or height which are shared among many people and can be obtained from other available public sources or by observation; however, by combining many of this kind of features it is not impossible identify a specific person. The third category represents data that need to be protected, such as those that include results from examinations or diagnoses of patients.

1.1. Privacy and anonymity

Two differing extreme perspectives on privacy can be envisioned: (1) institutional, in which the goal may be to disseminate as much information as possible, and (2) individual, in which no dissemination of information is wanted. In between these extremes lies the need to protect against the disclosure of specific data that can reveal sensitive information. This continuum of privacy perspectives is acknowledged in the 1991 (revised in 1997) Institute of Medicine (IOM) report ''The Computer Based Patient Record: An Essential Technology for Health Care'' [12] by addressing privacy from both provider and patient standpoints.

From a technical standpoint, we can identify two different issues: (1) the discovery of individual information items from disseminated data, and (2) the discovery of relations within the data. The former can be illustrated by the discovery of social security numbers that have been removed from the disclosed data, while the latter can be illustrated by the discovery of a relationship between a set of genotypes and predisposition to certain diseases. Relations in the data (as well as *which* data items were collected, as in the example of laboratory tests) may reveal information both about the health provider and the individual, hence protecting provider and patient privacy requires protection against disclosure of particular relations. A form of this protection has been investigated in [13]. No information about a particular individual is released if we cannot determine whether this individual is included in the data set of question. From this perspective, it is the anonymity of the individual that is the main concern raised on the issue of privacy.

There exists confusion about what is meant by ''anonymous''. This can be exemplified by the different entries in Webster's dictionary [14]:

Anonymous:

1. not named or identified;
2. of unknown authorship or origin;
3. lacking individuality, distinction, or recognizability.

The removal of names and other explicit identifiers from records such as social security numbers is often termed ''de-identification'', and as such is compatible with the first definition of anonymous, but not the third, which is also extremely important in the disclosure of medical data. Regulations for the protection of health related information have generally treated the ''de-identification'' of the medical record as a sufficient privacy protection mechanism. As demonstrated by Sweeney [15], this may not be sufficient to protect privacy. The reason is that the data may contain features that are available elsewhere, and when values from these features are combined, identification of a sufficiently small subgroup of individuals in the general population is possible and consequently jeopardizes privacy of an individual in that group. Some privacy protection mechanisms have been designed with these problems in mind [16].

1.2. Example

Imagine that, for all *n* records in a community database, the value for a certain disease of interest is disclosed, and that someone wants to find whether a certain neighbor or celebrity who lives is a neighborhood has that particular disease. The prevalence of that disease in the population at large is 8% and in that database in particular is 4%. Suppose that ''non-identifiable'' information such as ethnic origin, age range, weight, and height are available (''identifiable'' information such as name, SSN, date of birth, and address are removed), and that a simple query reveals that, from *n* entries in the database, *y* entries match the profile of the neighbor or celebrity, in terms of those four attributes. Among these *y* records, the prevalence of the disease of interest is 80% (10 times that of the population at large, 20 times more than that of the population in the database). It is easy to see that, although the database was ''de-identified'', one could still obtain valuable (albeit somewhat uncertain) information regarding a particular individual or group of individuals. In this example, the prior probability of disease to any record in the database was 4%, but after removing those records that did not match the profile, the posterior probability went up to 80%. This added information, quantified by the change in probabilities, can be used maliciously to infer that the neighbor or celebrity indeed has that disease. From an aggregate perspective, an insurance company could use this information as a basis to deny eligibility for certain profiles of patients, or a company might use it to make hiring/firing decisions. These decisions would not be based on the ''identification'' of a certain record, but on inferences that can be drawn from a ''de-identified'' database.

2. When are data anonymous enough?

What is needed to protect privacy at this level is a way to ensure that any given combination of feature values (comprising a partial record) in the data does match a large enough subgroup of the population, i.e., that the data are *ambiguous* enough. This matches the third definition of anonymous given above. A problem with defining what *ambiguous enough* is that, when different sources of data are combined, this requires omniscience. Achieving a pre-defined degree of ambiguity in the disclosed data, however, only requires the information given in the data, and is achievable. In fact, most anonymization methods rely on this.

In order to use a feature value to identify individuals in the population, the feature must be available outside the data itself. For example, consider the case in which the patient's date of birth is substituted by age range in the disseminated data listing all cases of infectious diseases in a certain region, but the zip code for the patient's home address is included. In the disseminated database, all of the people in the upper age ranges have some communicable disease, such as tuberculosis (TB). Names and unique identifiers such as social security numbers are obviously excluded. Such a database would be useful for several purposes, for example, in the exploration of whether clusters of disease are associated with socio-economic conditions. If there is a group of senior citizens can be linked to a particular zip code area, an obvious identification of these individuals and the fact that they have a communicable disease can be drawn from the matching of an ''anonymous'' data set and a publicly available one. Only features that are not available outside of the data cannot be used to tie information to any explicit identifier directly. By transforming a feature, potentially publicly available, into a feature that is not available outside the data, linking is prohibited. An example of this is given by Armstrong et al. [17] in the context of masking geographical coordinates such as zip code. These are transformed in such a way that the original location cannot be established, but properties such as proximity are preserved, allowing cluster-

ing. The problem with such transformations is that a particular transformation preserves only parts of the information contained in the feature, and if it is not a priori known what information is relevant for the application of the data, there is a risk that important information is lost. For example, if the goal of disseminating the database were to plan for staffing of local community centers, the transformed database would be useless.

In general, ensuring anonymity requires the removal of information. This loss is balanced against the wish to supply a means for research as efficiently as possible, i.e., containing as much exact relevant information as possible.

2.1. Different applications have different protection needs

Different applications of the same data may require different information, e.g., descriptive versus predictive applications. Anonymization is not unique. For example, it is necessary to take into consideration the reason why data are being disclosed to determine removal of which features will be necessary to render application of functional dependencies impossible [18] for non-intended use of the same data. It follows that some variants of anonymous tables are better for a given information need than another. This requires an analysis of applications' information needs and an analysis of information loss incurred by different approaches to anonymization in order to optimize data utility. Tailoring of anonymization to application needs might produce different versions of the same data. Even though each version preserves anonymity, the combination of these versions might not. It is therefore crucial that inferences using multiple versions of the same data be hindered.

Another issue refers to inferences based on ''meta knowledge''. Meta knowledge is knowledge about the database beyond its contents. An example would be the knowledge that an absolute minimum of information loss was incurred in order to meet given anonymity requirements. As examples given by several authors [10,13,19] show, lack of redundancy caused by such minimal loss can leave the data vulnerable to inferences about sensitive information.

2.2. There is potential for inference in disclosed data by statistical and machine learning methods

As more complex inference engines based on statistical or machine learning methods are being used to ''mine'' databases, it is reasonable to consider the issue of anonymity as a function of the potential for inference in a certain database. We therefore claim that ''de-identification'' of health related information is not a sufficient mechanism to protect privacy, and that this issue must be studied in conjunction with predictive models that can be ''learned'' from the data. These models can help in measuring the degree of anonymity in a database by determining the potential for inference. The current legislation addresses this issue subjectively, by stating that the disclosed data should not be re-constructable using reasonable statistical or other scientific methods [1]. Like Sweeney [15], we believe that it is necessary to address this issue objectively and quantitatively.

We propose the investigation of the use of machine learning methods not only to evaluate solutions produced by these heuristics (e.g., quantify the inferences that can be made using different databases, that have been ambiguated at different degrees), but also using their principles to develop anonymization algorithms. For example, we have created table ambiguation algorithms that can be used to hinder certain types of inferences in databases [13]. We can build prediction models and use them to evaluate the degree of information loss in databases containing different degrees of ambiguity.

3. Degrees of ambiguity

Quantitative assessment of ambiguity is necessary in order to be able to offer guarantees of privacy maintenance. From the above discussion, we can identify the following possible measures:

1. Ambiguity of a given set of feature values in the population (and also in the data at hand).
2. The effect of adding a particular feature value to a given set of feature values.
3. The availability of a given feature outside the data, i.e., the potential cost of making a feature value available.

Each of these can ideally be assessed on an ordinal scale, and can be combined to produce composite measures. In the discussion below, we present some possible measures.

3.1. Partial ambiguation: linkable and unlinkable attributes

We state that a database can be considered ''x-ambiguated'' if every entry is indistinguishable

from at least *x* different entries. We can relax this requirement to the parts of the record that are presumed to be ''not linkable to other databases (i.e., not available outside the data to be disseminated). For example, assume that a certain institution is the only one able to perform a certain genetic test. The attributes for this test can be made available together with an ''*x*-ambiguated'' clinical record, so the ''ambiguation'' is only relative to the clinical record (and not the genetic record, which may be unique). In other words, an individual entry would be indistinguishable from at least *x* other entries in terms of the clinical record, but not in terms of the genetic information. Since there would be no way to ''link'' the record using the genetic portion (as no one else in the world has that type of information), for all practical purposes the database is ''ambiguous''. Note that this relaxation of the assumptions makes the process of ambiguation somewhat simpler, as not all features need to be considered in order to determine whether functional dependencies exist in the database. It does not destroy keys in a database, however, as the entries may still be perfectly distinguishable when all attributes are considered.

3.2. Relative ambiguation: sensitive and non-sensitive attributes

We also need to define relative ambiguation. Relative ambiguation is concerned with the protection of sensitive information, or the process of decreasing the recognizability of an individual or group of individuals with regards to sensitive information. A key point is that relative ambiguation might be necessary and sufficient to protect privacy. For example, preventing inference on a certain disease may be desirable, but preventing inference among findings may not be necessary, and might even be wanted (i.e., we want to prevent the inference that most middle-aged females have a certain disease, but we are not worried if we can infer that most of these individuals have brown eyes), so complete ambiguation may be unnecessary. In this case, the ambiguation will be driven by the possible inferences on the sensitive attribute. The usual practice is to completely remove ''sensitive attributes'' or generalize its values (e.g., substitution of every entry value by the average). There is significant loss of information with these procedures.

It is possible to construct indices of relative ambiguation and have them ''drive'' an ambiguation

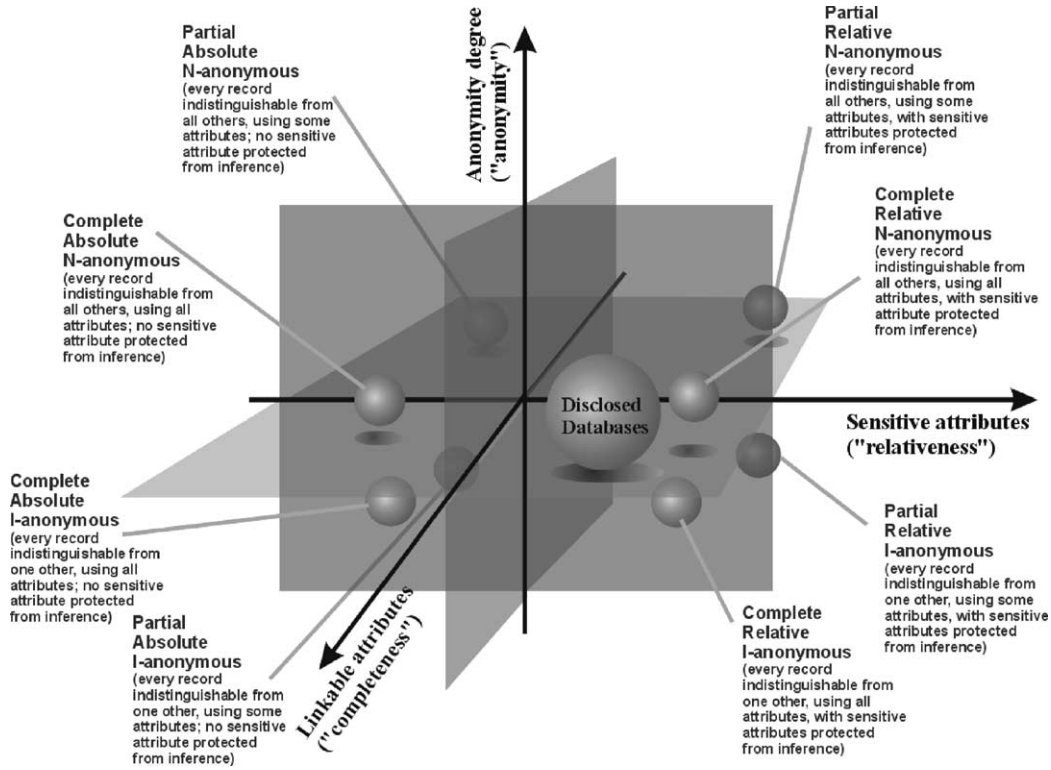

Fig. 1 Partial and relative anonymity degrees. Disclosed databases in intermediate stages of partial and relative anonymity, and different *x*-anonymity are illustrated.

algorithm. In relative ambiguation, one wants to make sure a certain patient is indistinguishable from a number of other individuals who have the same value for the sensitive attribute, and also indistinguishable from some individuals who have different values for the sensitive attribute, so that no inference can be made that would reveal the value for the sensitive attribute with a certain degree of certainty.

Additionally, there are situations where anonymity may be not enough as to ensure confidentiality. It can happen that a cluster of indistinguishable records have the same value for a given feature. Therefore, one cannot say what the record of a given patient is, but if it is possible to infer that this patient belong to this cluster (for instance, if the cluster contains all elder patients), it will be possible to infer about a specific value of a feature to the patient under discussion. For this reason, some algorithms can add other steps in order to avoid ''uniqueness'' of a feature in the same cluster [11].

Fig. 1 shows an illustration of degrees of ambiguation, as related to the concepts of partial and relative ambiguity. Disclosed databases are usually in intermediate stages of partial ambiguity, and vary in terms of *x*-ambiguity. Relative ambiguity, or protection against internal inferences, is sometimes overlooked.

Fig. 2 is a toy example designed to illustrate how deletions of select data can produce *x*-ambiguity. Assume that in this database there are 10 patients and five features were registered: hair type, eye color, T-shirt pattern, pants pattern, and skin color. Each feature can have a given number of values; for instance, hair may be curly, smooth or sticky, T-shirt pattern may be in balls, diagonals, stripes or squares, and so on. These individuals are represented in Fig. 2a. A level of three-ambiguity is

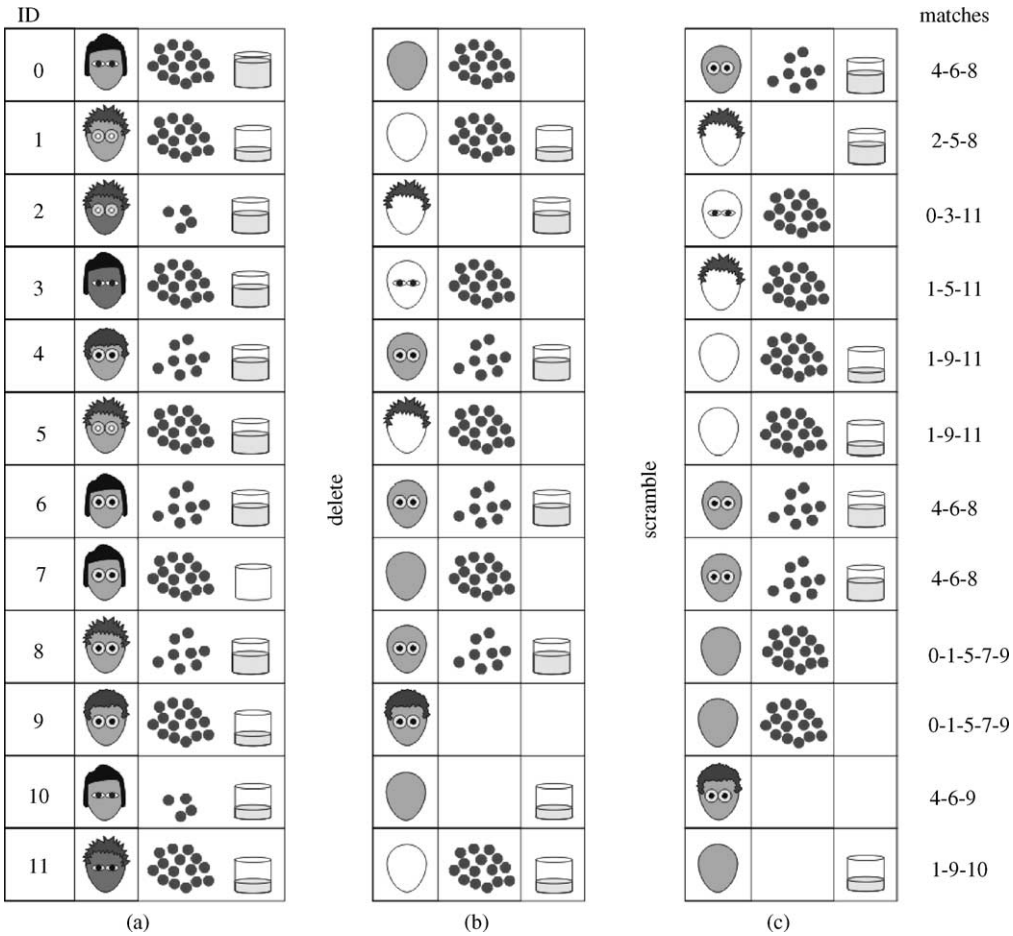

Fig. 2 Toy illustration of a database with five features (hair, eye color, skin color, hemoglobin and urine results). (a) Original data: it is possible to observe that each patient is distinguishable. (b) After some data deletion (represented by removed shapes) the resulting database adheres to three-ambiguity level. (c) Record order is scrambled and identification numbers are suppressed to produce the final output database. For instance, the first record in the right panel matches patients 4, 6 and 8 in left panel; the second record matches patients 2, 5 and 8, and so on.

required in the example and an algorithm should be able to choose deletions in strategic values, producing the output represented in Fig. 2b, which must minimize the number of deletions to achieve three-ambiguity level. It is obvious that the record numbers must be also suppressed since they are identification features and the order of records should be scrambled (Fig. 2c). If one picks up any individual from Fig. 2c, it will be possible to verify that it matches at least three patients in the original database in Fig. 2a.

4. Requirements for ambiguation algorithms

We anticipate that different utilities may be involved in decision theoretic models designed to assess the value of information of databases with varying degrees of ambiguation. This diversity will result in different requirements for ambiguation, depending on the intended users. Lists of linkable and sensitive attributes, desired degree of anonymity, together with the database itself, should serve as inputs to the ambiguation algorithms. For example, a potential user should know that a certain database should receive partial ambiguation involving linkable attributes A, B, and D, and that it should be ambiguated relative to sensitive attributes E and G. The database should be ambiguated to prevent identification of patients with a given threshold of a given ambiguity measure, thus allowing the user to know that the probability of obtaining the correct record is below some computable threshold, and that attributes E and G cannot be inferred given the input information.

We have so far concentrated on a very abstract description of ambiguation and the requirements of a solution. An optimal anonymization would imply minimal data loss and assurance of ambiguity at a desired level (if assumptions of linkable and sensitive attributes hold true). Certain instances of this problem yield hard optimization problems [10,13,18,20]. Certain heuristics can be used to obtain non-optimal but still adequate solutions, however.

We have studied the problem of anonymizing a data set from a Boolean algebra perspective, using cell supression. We have developed indices of absolute and relative ambiguity and algorithms that can be applied in individual tables [10], and have formalized the problem from a theoretical perspective [21]. In the case of non-relative ambiguity, the index reflects the minimum number of individuals that are indistinguishable. In the case of relative ambiguity, the index is proportional to the entropy (relative to the sensitive attribute) of each group of indistinguishable individuals, which is related to inference ability.

Recent follow-up on this topic has provided us with insight on the limitations of the algorithm proposed, as well as some new directions [13]. It became clear that the type of table ''ambiguation'' proposed is a powerful tool for hindering certain types of inferences based on the data, but that it is not sufficient for several applications. In particular, we have found that preventing conclusions with total certainty is not enough to avoid malicious use of datasets (see example above), and that a mechanism to monitor the degree of uncertainty in inferences that can still be made from a disclosed database has to be added. The types and quality of inferences that can be made depend, evidently, on the learning algorithm.

5. Discussion and summary

The main challenges in privacy protection, from a technical standpoint, are summarized below:

- Agreement on what constitutes an ''anonymized'' data set, and how the degree of anonymity can be quantified (formalization of the problem).
- Assessment of the availability of features outside of the data.
- Development of efficient algorithms to ambiguate a data set to a desired degree, considering linkable and sensitive attributes (implementation of a solution).
- Development of measures of data quality for different applications.
- A method for selecting an appropriate ambiguation strategy for a given application.
- Testing of ''anonymous'' data sets for potential breaches in privacy (verifying current solutions).
- Other practical challenges are represented by changes of status for attributes (e.g., a previously unlinkable attribute is now linkable, or a non-sensitive attribute is now sensitive), and strategies to enforce the use of anonymization algorithms. These challenges are easy to anticipate, but hard to solve.

We have presented the problem of anonymity in the context of privacy protection, describing the difficulties in the development of metrics that indicate the degree of relative anonymity and the degree of protection for sensitive information. These metrics are a necessary but not sufficient step in the development of new anonymization algorithms.

Acknowledgements

This work was funded in part by grants R01LM06538-01 and R01LM07273 from the National Library of Medicine, National Institutes of Health. Paulo S.P. Silveira was a postdoctoral fellow at DSG funded by CAPES from October 2002 to September 2003 (process BEX0659/02-9).

References

[1] Rules and Regulations. Federal Register 65 (250) (December 28) (2000).
[2] J. Kulynych, D. Korn, The effect of the new federal medical privacy rule on research, N. Engl. J. Med. 346 (3) (2002) 201—204.
[3] J. Kulynych, D. Korn, The new federal medical-privacy rule, N. Engl. J. Med. 347 (15) (2002) 1133—1134.
[4] B. Barber, Patient data and security: an overview, Int. J. Med. Inform. 49 (1998) 19—30.
[5] E.B. Andrews, Data privacy, medical record confdentiality, and research in the interest of public health, Pharmacoepidemiol. Drug Safety 8 (1999) 247—260.
[6] S. Batami, Patient data confidentiality and patient rights, Int. J. Med. Inform. 62 (2001) 41—49.
[7] J.G. Anderson, Electronic patient records and the impact of the internet, Int. J. Med. Inform. 60 (2000) 111—118.
[8] C. Safran, H. Goldberg, Electronic patient records and the impact of the Internet, Int. J. Med. Inform. 60 (2000) 77—83.
[9] G.J. Annas, HIPAA regulations: a new era of medical-record privacy, N. Engl. J. Med. 348 (15) (2003) 1486—1490.
[10] A. Øhrn, L. Ohno-Machado, Using Boolean reasoning to anonymize databases, Artif. Intell. Med. 15 (1999) 235—254.
[11] W.C. Chiang, T.S. Hsu, S. Kuo, C.J. Liau, D.W. Wang, Preserving confidentiality when sharing medical database with Cellsecu system, Int. J. Med. Inform. 71 (2003) 17—23.
[12] R.S. Dick, E.B. Steen, D.E. Detmer, The Computer Based Patient Record: An Essential Technology for Health Care, revised ed., Institute of Medicine, Washington, DC, 1997, p. 234.
[13] L. Ohno-Machado, S. Vinterbo, Dreiseitl. Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance, J. Am. Med. Inform. Assoc. 9 (90061) (2002) S115—S119.
[14] Merriam-Webster Online Collegiate Dictionary. Available online at http://www.m-w.com/.
[15] L. Sweeney, Weaving technology and policy together to maintain confidentiality, Summer-Fall, J. Law. Med. Ethics. 25 (2—3) (98—110) (1997) 82.
[16] M.P. Armstrong, G. Rushton, D.L. Zimmerman, Geographically masking health data to preserve confidentiality. Stat. Med. 18 (1999) 497—525.
[17] S. Dawson, S.C. Vimercati, P. Lincoln, P. Samarati, Maximizing sharing of protected information, J. Comput. Sys. Sci. 64 (2002) 496—541.
[18] T-A. Su, G. Ozsoyoglu, Controlling FD and MVD inferences in multilevel relational database systems, IEEE Trans. Knowl. Data Eng. 3 (1991) 474—485.
[19] L. Sweeney, Guaranteeing anonymity when sharing medical data, the DataFly system, in: D.R. Masys (Ed.), Proceeding of the AMIA Annual Fall Symposium, Hanley and Belfus, Philadelphia, 1997, pp. 51—5.
[20] M. Fischetti, J.J. Salazar, Models and algorithms for the 2-dimensional cell suppression problem in statistical disclosure control, Math. Program. 84 (2) (1999) 283—312.
[21] S.A. Vinterbo, Privacy: a machine learning approach. IEEE Trans. Knowl. Data Eng. 16 (8) (2004).