ELSEVIER

International Journal of Medical Informatics 60 (2000) 177–183

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

Anonymous statistical methods versus cryptographic methods in epidemiology

Catherine Quantin *, François-André Allaert, Liliane Dusserre

Medical Informatics Department, Dijon University Hospital, 1 Boulevard Jeanne d'Arc, BP 1542, 21034 Dijon Cedex, France

Abstract

Sensitive data are most often indirectly identifiable and so need to be rendered anonymous in order to ensure privacy. Statistical methods to provide anonymity require data perturbation and so generate data processing difficulties. Encryption methods, while preserving confidentiality, do not require data modification. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Epidemiological survey; Linkage of records; Hash-coding; Security; Data privacy

1. Introduction

In the last decade, the demand has increased considerably for microdata sets "*i.e. data sets containing data on individual respondents …*" [1] and for detailed information such as, for statistical studies, large multidimensional tables. However, both tables and "*microdata sets lead to considerable problems when trying to protect the privacy of the respondents*" [2].

The European directive 95/46/CE, from the European parliament and the Council on the protection of individuals, about processing and circulation of personal data, defines 'personal data' by "*any information relating to an identified or identifiable natural person; an identifiable person is one who can be identified, directly or indirectly, in particular, by reference to an identification number or to one or more factors specific to his physical, mental, economic, cultural or social identity*". With this definition, the elements identifying a person are quite varied and have different discriminant powers: for instance, the birth date is much more discriminant than the sex. Moreover, the preambles 14–17 of the Directive extend this definition to data issued by techniques "*used to capture, transmit, manipulate, record, store it or communicate sound and image data relating to natural persons*" so that, from the legal point of view, it appears often difficult to recognize personal data.

Sweeney [3] stresses that "*de-identifying data*" i.e. removing or replacing variables

* Corresponding author.
E-mail address: cquantin@u-bourgogne.fr (C. Quantin).

PII: S1386-5056(00)00118-0

that directly identify an individual (such as name and address, called direct identifiers) does not guarantee anonymity because "*the term anonymous implies that the data cannot be linked to identify any individual*". So, the data are not anonymous if a combination of variables allows a one-to-one relationship with the corresponding individuals in a public nominative register. These variables, even if they are not direct identifiers, are said to be 'sensitive' because they allow identifying individuals for which sensitive information may be revealed.

This paper aims at showing how the statistical point of view on identification meets the legal one, and the limits-and the dangers-of the techniques used to provide anonymity.

2. 'Indirectly identifiable' information from the statistical point of view

The statistical problem is to estimate the probability of identity disclosure through data linking, which implies the possibility of knowing all the public (or private) files that can provide linkage with the identity of the individuals. In the USA, "*population registers, such as local census data, voter registration lists, city directories, motor vehicle agencies, tax assessors, real estate agencies and the World Wide Web, are publicly available and often include a postal code and birth date along with the accompanying name and address*" [3]; they are maintained by federal agencies and protected by a complex web of federal statutes [4]. Similar files, particularly for commercial purposes, are used in Europe where the legislation on data protection is more restrictive.

The next point to define is a threshold for the risk of identifying a person [5] under which anonymity could be admitted.

The USA Social Security Administration, recognizing that unique combinations of characteristics can be linked with other data sources, defined as a rule that the data must contain at least five individuals with the same combination of characteristics. In fact, the risk of being identified could depend on the number of individuals matching the same characteristics not only in the file to be released, but also within the other sources used for re-identification. Greenberg and Voshell [6], for example, showed on files from different geographic regions a relationship between the percentage of individuals with a unique combination of characteristic values and the size of the region. That is why statisticians [7] suggested basing the threshold computation on the percentage of identifications obtained from linkage attempts between the data to be released and a nominative public register.

More precisely, statistical reasoning [8] about disclosure is based on the a priori probability of an individual, whose characteristics are in a file, to be identified. In an earlier paper, Biggeri and Zanella [9] considered the hypothesis that all individuals had the same a priori probability, which they split into three factors (a) probability of an individual in the file being protected, appearing either in a public nominative file or in data collected for illegitimate purposes (intruder's file); (b) probability of an individual having common data with the intruder's file and (c) probability of an identifiable individual being the object of a disclosure attempt.

A new difficulty for determining the threshold is then to define a level of information privacy: "*disclosure takes place if publication of statistical data makes it possible to determine the identity of even a single individual*" [6], because the damage caused

to an individual by revealing sensitive information will not be reduced by a low probability of disclosure. Conversely, for the community, a given individual's interest should not interfere with the possible benefit of data processing for epidemiological research, for example.

As a solution, statisticians proposed to estimate the risk of disclosure not as an average risk for a file but separately for each individual. The principle is to take into account the characteristics of the individuals and to determine the frequency of each combination of values of the sensitive variables. So, Franconi and Benedetti [7] estimate an individual probability (which may vary between individuals): they count the individuals corresponding with each combination of values of the sensitive variables, then group the combinations according to the number, h, of individuals they identify. This easily locates identifiable combinations since the risk of identifying an individual decreases when the corresponding value of h increases: a birth year and the name of a small city may be associated to $h = 1$ (one individual only), while a birth year and the name of a big city may be associated to $h = 5$ (five individuals), so that the risk of disclosure is obviously higher in the first case. Reasoning this way, Biggeri and Zanella [9] have only considered the simplest case of combinations each describing a unique individual ($h = 1$).

Sweeney [3] also stressed that the threshold should take into account the degree of anonymity required for the released data, for example, medical databases may contain more sensitive information than databases dealt with by travel agencies. Supposing that valuable data (e.g. medical data) merits more effort (visiting or phoning the individuals, direct mailing) for identification, the threshold should be decreased to render these efforts useless.

3. Statistical methods to provide anonymity

Statisticians focused on aggregation of individuals in a data file by enforcing them to share the same characteristics, through data perturbation such as "*changing singletons to median values*, *inserting complementary records*, *generalizing codes*, *swapping entries*, *scrambling records*, *suppressing information*" [3], for example, to reduce a birth date to the year, to remove either the age or the full record of a centenarian.

For generalizing codes, controlled rounding (i.e. replacing the true value by an appropriately rounded value) has been the most studied method [10–12]. Other authors proposed to masking microdata by adding noise, generated by bootstrapping from the original empirical distribution of the data [13].

Sweeney [3] introduced the Scrub System, for removing or 'scrubbing' identifying information in text documents [14], and the Datafly System, devoted to field-structured records [15], where the user must specify a level of anonymity for each variable field: this level is a number between 0 and 1 determining the minimum bin size for the field, so that every value in the field occurring less than this minimum bin size is recoded (for example birth date is recoded in months or years), or even removed, if after this generalization the bin size still does not meet the requirement.

The other approach proposed by Statistics Netherlands [16] is considered as the most usable confidentiality software in the European Community. Like the Datafly system, the program μ-Argus generalizes values or even removes outlier information in order to achieve a minimal bin size requirement but, unlike Datafly, does not suppress entire records.

The problem then is to evaluate the effect of these modifications on the data processing

and the resulting statistical biases. More generally, it raises the question of the loss of quality that we can accept in exchange for anonymity.

4. What loss of data quality can we accept in exchange for anonymity?

Data perturbation implies a statistical bias since it is not randomly applied. Moreover, removing or substituting outlier values may strongly affect mean and standard deviation or hide rare events, such as serious complications, whose consequences upon the studied disease should not be ignored.

Cox demonstrated [17] for two-dimensional tables that, among the different methods for data suppression or perturbation, only controlled rounding met the objectives of non-bias and reduced data distortion, but this technique appears too complex for higher dimensional tables [18]. Fuller proposed to reduce biases by treating as a measurement error [18] the masking error due to data perturbation. Some statisticians [19] proposed algorithms minimizing the information loss due to data perturbations or suppressions. But many methodological problems must be solved before applying these techniques appropriately.

Anonymous data, by definition, prevents any possibility of finding the original document, e.g. a medical record, for checking: this implies a strict validation of all data recorded in the base before providing anonymity whereas, in practice, it is rare to eliminate all errors before processing.

Another problem appears with research in evidence based medicine, where data perturbation disagrees with auditability and 'good clinical practice' which state that the database used for the statistical analysis must reproduce exactly the medical files. It is moreover unacceptable that data perturbation should lead to unfair assessments by choosing the method to provide anonymity, which insures the 'best' results.

5. Advantages and disadvantages of cryptographic methods

Statistical methods for anonymity, implying data perturbation and thereby loss of data quality, make appealing the methods proposed by computer scientists [20] to achieve data protection through security systems based on user identification, passwords, cryptology, output control, operating system safeguards and review of both job requests and outputs [21].

For many years, the use of cryptographic methods was restricted by the law in many countries like France or the United States of America. Cryptographic methods were considered as a weapon, and therefore, their implementation was subject to specific authorization and regular controls by the Security Agencies. As a consequence, it was difficult to use a really efficient cryptographic algorithm i.e. an algorithm using a key with a length of more than 40 bytes.

In Europe, the legislation has been recently modified in order to allow the appearance and the recognition of the electronic signature. A European directive 1999/93/CE on electronic signature has been voted by the parliament in November 1999. As a consequence, the different member States must modify now their national legislation in order to allow a general admittance of cryptographic methods to protect data and to sign it.

So rather than applying statistical methods which appear to decrease data quality application of cryptographic methods may be useful. If you require a strict anonymity without

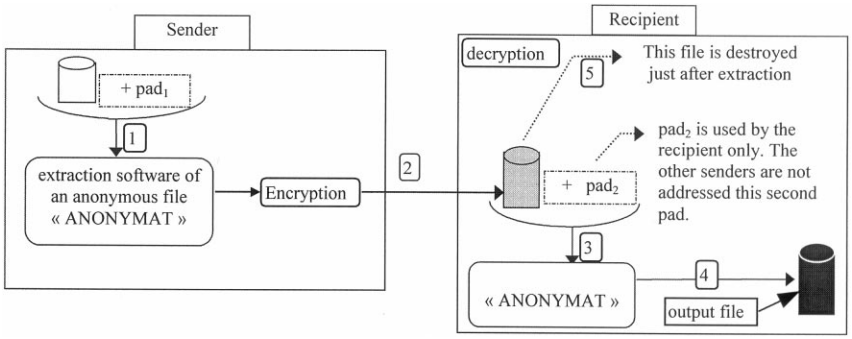

Fig. 1. Transmission of an anonymous file.

any possibility of getting *back* to the original data, the best solution is to use a hash algorithm [22]. If you need to ensure data confidentiality during data transmission or during storage periods but the data manager needs to be able to access personal data directly for the purpose of his work and is duly authorized for that, the use of symmetric cryptographic algorithms is recommended. But as these methods are reversible, the access to the data and their integrity must be protected. This can be handled through asymmetric algorithms which allow a strong authentication of all people having access to the database, the signature of all data consulted or modified, and therefore the auditability of any data processing.

However, these two methods have also disadvantages that people must take into account before implementing them. If the hash algorithm is chosen, no return will be possible to the patient's identity. Therefore, the quality control of the data has to be performed before the hash coding. The sender may keep the corresponding table between the identity and the hash code and thus may provide some rectification to specific requests sent by the data manager. But the process will introduce important delay and postpone the statistical analysis. Moreover because of the aspects of the hashcode, it is more difficult to detect any integrity default of the data sent than on an uncoded identity.

Although mathematically irreversible, the hash computation does not completely guarantee information security. Thus, one could apply the algorithm to a large file of identities and compare the resulting codes to the codes to be deciphered. To prevent this dictionary attack [23], two pads[1] must have been introduced before the application of the hash algorithm [24]. The first pad (pad1), is used by all senders of information and the second one (pad2) is applied by the recipient, which anonymously ensures the file linkage. Nominal information is therefore hashed twice with these two pads. The aim of pad1 (resp. pad2) is to prevent a dictionary attack from the recipient (resp. sender). As a consequence, the person in charge of the pad management should not be involved in the study and should not be given the hashed files (Fig. 1)

The advantage of the symmetric algorithm is to protect the data from anybody who doesn't belong to the authorized persons using the network. But any autho-

[1] A pad is a large random file

rized person may read or modify the data. Therefore, each access to data must be protected which implies a clear definition of the access rights and the auditability of each access to data. That can be implemented by the use of passwords but the use of an asymmetric algorithm allows strong authentification and non-repudiation through an electronic signature. In France, the Professional Smart Card, which is the key of the medical network's security, contains both a symmetric and an asymmetric algorithm.

6. Conclusion

Although the idea of using cryptology in communication to maintain good levels of security and privacy is not new [25], encryption techniques are under-utilized in official statistics, likely due to previous legal restrictions. For example, the French legislation concerning encryption has only been recently updated in order to simplify the use of encryption techniques for personal data security.

An extreme solution could be systematically to demand, through legislation, the encryption of all data, and not some as suggested by Dalenius [26], without specifying any degree of anonymity: this would suppress the problem of defining "*indirectly identifiable*" information and would facilitate recognition of the sender of a file and record matching while preserving confidentiality [27], which would be very useful for multicentre studies. Moreover, encryption could be performed at the stage of data capture to reinforce security. However, to ensure data privacy, data processing requiring description resulting in "*nominative or indirectly nominative*" data should be placed under the responsibility of a certified authority.

Acknowledgements

This research has been sponsored by the Burgundy Regional Council and the French 'Ligue Bourguignonne' against Cancer.

References

[1] A.G. de Waal, L.C.R.J. Willenborg, Global Recoding and Local Suppession in Microdata Sets. Statistics Netherlands, Department of Statistical Methods 12th International Symposium on Methodology Issues: From data to information, Statistics Canada Ottawa, 1–3 November 1995.

[2] L.C.R.J. Willenborg, A.G. de Well, W.J. Keller, Some Methodological Issues in Statistical Disclosure Control Statistics Netherlands, Department of Statistical Methods, Second Cathy Marsh Memorial Seminar, 7 November, 1995 London.

[3] L. Sweeney, in: B. Cesnik, A. McCray, J.-R. Scherres (Eds.), Three Computational Systems for Disclosing Medical Data in the Year MEDINFO 98, IMIA, IOS Press ,1998, Amsterdam, 1999, pp. 1124–1129.

[4] J.S. Cecil, Confidentiality legislation and the United States federal statistical system, J. Official Stat., 9;(2)519–35.

[5] G.T. Duncan, Inferential disclosure-limited microdata dissemination, in: Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, 1990, pp. 440–445.

[6] B. Greenberg, L. Voshell, Relating risk of disclosure for microdata and geographic area size, in: Proceedings of the Section on Survery Research Methods, American Statistical Association, Alexandria, VA, 1990, pp. 450–455.

[7] L. Franconi, R. Benedetti, Some aspects of disclosure avoidance in complex microdata files. http://europa. Eu. Int./en/comm/ eurostat/ research/ publications/ros/a3. htm.

[8] L.C.R.J. Willenborg, Remarks on disclosure control of microdata statistica, Neerlandica, 1992, 46 (1).

[9] L. Biggeri, F. Zannella, Release of microdata and statistical disclosure control in the new national system of Italy : main problems, some technical solutions, experiments. Proceedings of the 48th ISI Session, Cairo, 1991, pp. 307–310.

[10] B. Cox, L.H. Causey, R. Ernst, Application for transportation theory of statistical problems, J. Am. Stat. Assoc. 80 (392) (1985) 903–909.
[11] L.H. Cox, A constructive procedure for unviased controlled rounding, J. Am. Stat. Assoc. 82 (1987) 520–524.
[12] T. Dalenius, A simple procedure for controlled, Rounding. Statistisk Tidskrift 3 (1981) 202–208.
[13] R.J. Bowden, A.B. Sim, The privacy bootstrap, J. Bus. Econ. Stat. 10 (1992) 337–345.
[14] L. Sweeney, Replacing personally-indentifying information in medical records, the Scrub system, in: Proceedings of the American Medical Informatics Association, Hanley and Belfus, Washington, 1996, pp. 333–337.
[15] L. Sweeney, Guaranteeing anonymity when sharing medical data, the datafly system, in: Proceedings of the American Medical Informatics Association, Hanley and Belfus, Nashville, 1997, pp. 51–55.
[16] A. Hundepool, L Willenborg, μ and ι-argus: software for statistical disclosure control. Third Internatinal Seminar on Statistical Confidentiality Bled: 1996 www.cbs.nl/sdc/argus1.htm
[17] L.H. Cox, Disclosure Control Methods for Frequency Count Data Presented at the Census Advisory committee Meeting of the American Statistical Association, Bureau of the Census, 1984.
[18] W.A. Fuller, Masking Procedures for Disclosure Limitation, J. Official Stat. 9 (1991) 383–406.
[19] L.H. Cox, Suppression methodology and statistical disclosure control, J. Am. Stat. Assoc. 75 (1980) 377–385.
[20] S. Keller-McNulty, E. Unger, Database systems: inferential security, J. Official Stat. 9 (1993) 475–499.
[21] M. Cigrang, L. Rainwater, Balancing Data Access and Data Protection: the Luxembourg Income Study Experience. Proceedings of the Statistical Computing Section American Statistical Association, Alexandria, VA, 1990, pp. 24–26.
[22] C. Quantin, H. Bouzelat, F.A. Allaert, et al., How to ensure data security of an epidemiological follow-up assessment of an anonymous record linkage procedure, In. J. Med. Inf. 49 (1998) 117–122.
[23] E. Meux, Encrypting personal identifiers, Health Services Res. 29 (1994) 247–256.
[24] C. Quantin, H Bouzelat, F.A. Allaert, et al., Automatic record hash coding and linkage for epidemiological follow-up data confidentiality, Meth. Inf. Med. 37 (1998) 271–277.
[25] D.E. Denning, Cryptography and Data Security, Addison-Wesly, Reading, MA, 1982.
[26] T. Dalenius, Finding a needle in a haystack or indentifying anonymous census records, J. Official Stat. 2 (1986) 329–336.
[27] J.W.P.F. Kardaun, L.C.R.J. Willenborg, Cryptological applications in official statistics, statistics Netherlands, Department of Statistical Methods, Seminar on New Techniques and Technologies for Statistics, Bonn, 20–22 November 1995.