Epidemiological and Statistical Secured Matching in France

Catherine Quantin[1-2*], Benoît Riandey[3]
1- CHRU Dijon, Service de Biostatistique et d'Informatique Médicale (DIM), Dijon, F-21000, France. Contact author : catherine.quantin@chu-dijon.fr
2- Inserm, U866, Univ de Bourgogne, Dijon, F-21000, France
3- Institut National d'Etudes Démographiques, Paris, France
* Corresponding author : Catherine Quantin, email : catherine.quantin@chu-dijon.fr

Abstract

When we need to match different files, we generally face the problem of preserving confidentiality and respecting private life, especially if the files come from various sources. This problem may, on some occasions, dramatically restrict the utilization of administrative data for social statistics and research. In order to overcome this problem, some organizations, statistical institutes and research centers use the Secured Matching technique.
Secured Matching is performed by using anonymous matching keys. One way to proceed is by using a technique for the irreversible encryption of the identifiers. The identifying elements common to several files are transformed by the same encryption process, allowing the matching to be done in an anonymous way by using the encrypted identifiers, with no possible link to the original identifiers
In the session, we will talk about the use of Secured Matching in France and describe some applications in epidemiology. We also show how these techniques of secured matching can be considered as a potential for innovation for official French statistics.

Keywords: Record linkage, hash-coding, Security, Non-reversible encryption, Epidemiological survey

1. Introduction

In 1969, in their extraordinarily innovative founding article, Ivan Felligi and Allan Sunter (Fellegi and Sunter 1969) opened the way for the inter-sectorial and inter-institutional processing of administrative files such as the various chapters of a survey questionnaire. With their probabilistic matching technique, they brought a solution to the knotty problems of the absence of identifiers or the legal restriction of their use.

Countries in Scandinavia and Northern Europe took advantage of the wealth of information in the social files of their social democracies as well as national confidence in their institutions to develop very effective statistics based on matching. This even allowed them to dispense with the collection of census data. The wealth of information in French administrative files gave rise to the hope that an equally effective statistics service could be created. However, in 1978, fears remaining from the period of Nazi occupation led to very restrictive legislation, which all but paralyzed efforts in this field. In any case, it prevented the matching of administrative files, and thus led to their analysis as individual items.

The ethical obligation of epidemiologists to cure and the need for the efficient management of healthcare expenditure led researchers to find an effective and secure solution based on the hashing of identifiers. There are many applications in epidemiology and in healthcare economics, whether it concerns the anonymous counting of patients with AIDS and other notifiable diseases, or matching care episodes for the same patient in different hospitals and healthcare establishments, or economic information related to healthcare costs and their management.

Such hashing techniques (SHA algorithm) (Quantin, Bouzelat et al. 1998) have enabled this progress to take place, but they do have a limit: fundamentally, they can only be used if matching was planned during the initial stages of the project. Any extension of the project following a change in the analysis plan or a new participation is impossible. We have thus imagined a secure matching technique that allows controlled enlargement of the scope of investigation without impinging on the level of confidentiality. The technique associates hashing using a public key and encryption (reversible) with a secret key. This overcomes the above-mentioned limit while maintaining a high level of security (Quantin, Bouzelat et al. 1998). On our suggestion, this technique was implemented for healthcare statistics in Switzerland (Borst, Allaert et al. 2001).

In this paper, we will describe some French applications of Secured Matching in epidemiology and show how these techniques can be considered as a potential for innovation for official French statistics.

2. Innovation for epidemiologists and healthcare economists

Confronted with the challenges of public health, doctors cannot remain powerless and passive. With the explosion of the HIV/AIDS epidemic in the absence of any treatment, epidemiologists first had to measure the expansion of the epidemic even though the free tests proposed to the population were anonymous. Permanent anonymous identifiers for patients were created to make it possible to link positive tests related to the same person and thus to avoid counting the same patient twice. Very quickly, the SHA hashing technique proved to be the most reliable.

While, for reasons of state security, encryption in France was restricted to the army and diplomacy until authorization for others was provided for in the French decree n°99-199 of the 17 March 1999, hashing was authorized without difficulty because, given its irreversible nature, it could not be used to communicate sensitive information confidentially.

These permanent anonymous identifiers thus allowed epidemiologists to match medical records from different hospitals in total respect of medical secrecy. This was the case for the Réseau Périnatal de Bourgogne (RPB), (Cornet, Gouyon et al. 2001), which includes all 18 public and private sector establishments of the region that provide care for pregnant women and newborn infants (approximately 18,000 births per year). Forty-two indicators have been collected since 1998. The information is extracted from Programme de Médicalisation de Systèmes d'Information (PMSI) abstracts, which are created for every hospitalization. Indeed, an abstract is created for every hospital stay in a public and private sector healthcare establishment, in order to determine the budget of hospitals according to their activity. Indicators that are not in the PMSI abstract, such as psychosocial risk factors, are recorded on a file attached to the PMSI abstract, thus constituting an «enlarged abstract». For the processing of medical data, these «enlarged abstracts» are linked at two different levels. On the one hand the «enlarged abstract» of the same person, mother or newborn infant, must be linkable for successive hospitalizations (several different units or establishments). On the other hand the «enlarged abstracts» of the mother must be linkable to those of the infant to allow evaluation of the postnatal impact of maternal diseases and risk factors. The linkage of anonymous data was made possible by the use of Anonymat software on the basis of six variables, recorded for the mother and her infant: the maiden name of the mother, her first name and date of birth, the first name of the infant and its date of birth, the post code of the mother's place of residence. Before transmission, the files are validated at each establishment.

In addition, the exhaustiveness and the quality of data collection are systematically verified by the RPB coordinating team, which ensures mother-infant linkage (for 99.9% of newborns) according to the Felligi and Sunter method.

In the absence of an identifier, epidemiologists have successfully used the Felligi and Sunter probabilistic matching technique on many occasions, as was the case, for example, in a study that aimed to evaluate the performance of determinations of vital status of patients (Fournel, Schwarzinger et al. 2009) by crossing hospital data with mortality data of the Institut National de la Statistique et des Etudes Economiques (INSEE), after rendering the information anonymous.
All of the patients (10,489), residing in metropolitan France or in overseas territories, hospitalized for the first time for a malignant tumor between 1998 and 2000 at the Institut Gustave-Roussy were included. The INSEE mortality data for the years 1998 to 2004 (approximately 3.5 million deaths) were used. After anonymization by hashing, the files for in-hospital mortality and morbidity were linked for family name, first name, date of birth and post code for the area the patient was born.
The results of the linkage showed the interest of using probabilistic linkage to obtain information on the vital status of a large number of patients at a low cost, since the proportion of correct links was 97.2%, sensitivity was 94.8% and specificity was 99.5%. These results were better for patients born in France: sensitivity 96.8% and specificity 99.8% than for patients born outside France (sensitivity 82.8% and specificity 97.7%), but the performance of the method has been improved by introducing a manual validation step.
This method is also used to cross hospital data with data from other sources, notably for validation purposes (Couris, Foret-Dodelin et al. 2004; Sagot, Mourtialon et al. 2012; Quantin, Benzenine et al. 2012(a); Quantin, Benzenine et al. 2012(b)).
Concerns about managing public healthcare expenditure more effectively and harmonizing the rights of all of the patients whatever their professional background have led healthcare economists to adopt a similar approach. First of all, it was necessary to link in an anonymous manner the PMSI abstracts relative to all of the hospitalizations in France for a given patient during a given year. This was done as early as 2001 thanks to hashing of the social security number, the date of birth and the sex, according to the recommendation of the Commission Nationale de l'Informatique et des Libertés. A similar system was implemented in Switzerland, which developed, in collaboration with a member of our team, a method that combined hashing and encryption techniques for hospital abstracts (Borst, Allaert et al. 2001).
Then, for all persons residing in France, all reimbursements of expenditure on healthcare were matched in a single database. Indeed, because of the different health insurance schemes and their separate databases, it was impossible to conduct studies on the whole population. A general database was thus created to gather all of these data: le Système national d'information inter régimes de l'assurance maladie (SNIIR-AM) (Goldberg, Jougla et al. 2012). The data collected in the SNIIR-AM include those from high-street practices (detailed codes for acts and drugs) as well as from hospitals (that is to say PMSI abstracts) and information on the pathology treated for patients with chronic or occupational diseases. In this database, data concerning the same person are linked thanks to an anonymous identifier (hashing of the social security number, the date of birth and the sex), using the technique described above. This database, which now includes more than a billion recorded items, is a fantastic tool for guiding public health policy.

Administrative information has a number of inherent limitations that only surveys can remediate. The IRDES (Institute for research and documentation in health economics) has for a long time collected health and social protection data for the health insurance organization via questionnaires The data from these questionnaires are linked to individual data for healthcare reimbursements or hospitalizations. This third organization, as a trusted party, is thus involved

in the process. IRDES manages the tables that link health insurance identifiers to survey identifiers and thus ensure anonymous matching. Hashing of the social security number could be another solution: rather than conducting surveys by selecting individuals and their reimbursement data from health insurance files and then engaging a survey institute to contact them by telephone, the opposite could be done: households could be selected from the census files, then the social security number could be collected so that the reimbursement data could be subsequently gathered. The social security number could be entered by the individuals surveyed, immediately hashed and then encrypted, which would avoid administrative complications related to knowing this number, while ensuring the confidentiality of the medico-administrative data.

The classical use of hashing makes it possible to match data according to the initial analysis plan thanks to a public hashing key, and precludes any other matching, even from the same identifier. This is the level of protection aimed for. However, it rules out any changes in the analysis plan. Let us suppose that two epidemiologists, both of whom used hashing techniques, conducted two independent studies, one on diabetes and the other on blindness, and conscious of cases of blindness caused by diabetes, they later decide to share their data. The strategy described above would preclude this collaboration, unless they manage to use Felligi and Sunter's probabilistic methods with success.

Another secure strategy would authorize this change in the analysis plan: it could be done if the same healthcare identifier were hashed in an identical manner to ensure anonymity, and then specifically encrypted. For a scientifically and ethically validated collaboration, a protocol could be established to use two encryption keys, to decrypt the identically hashed identifiers so that the data can be linked, and then make the resulting file secure by encryption with a third key (Quantin, Fassa et al. 2008).

3. A potential for innovation for official French statistics

For historical reasons, official statistics in France have lagged far behind in this domain. At the time the civil register was to be computerized, there was strong resistance against matching administrative files, considered an act of « big brother ». Politicians and militant lawyers, in their ignorance of the guarantees provided by statistical confidentiality, set up legislative barriers against matching administrative data. From 1978 to 2004, statistical confidentiality as incorporated in the 1951 public statistics law was no longer recognized by the new French law as a guarantee of the right of privacy of French citizens.

Public statisticians generally abandoned the synergetic analysis of individual data from several administrative files. The matching of files thanks to the civil status identifier had to be authorized via a very heavy decree procedure at the Conseil d'Etat or a specific law. A few brilliant studies were conducted, notably concerning pension files and pension contributions, but the statistical analysis of administrative files concentrated on the exploitation of data on a file-by file basis.

One example of this restriction concerned counting the number of doctors in each specialty. The statistics from the medical profession and the Ministry of Health did not agree. As the law forbade institutions to share their data, it was necessary to use employment data, which was particularly unsuitable for this work, to carry out the study. It was not possible to convince the authorities that since the hashing technique made it possible to count people with AIDS without duplications, it could also be used to count the number of doctors in each specialty.

The decree procedure at the Conseil d'Etat gave rise to the widespread idea that hashing was not useful for official statistics, and as the procedure was particularly heavy, interested parties avoided embarking on it.

The great wealth of information in French administrative files created for the purposes of the highly active welfare state could give rise to exciting projects. Let us consider the following original example: emigration outside France is a black hole in demographic statistics. It is calculated from differences in census data, which have more measurement errors than measurements. A new approach could be to hash the SSN (social security number), the identifier common to the health insurance system, civil status, employment status and welfare benefits. The health insurance files could be used to find the hashed SSN of people who have durably stopped consuming medical services, and from there create an age pyramid. From differences between this pyramid and the age pyramid of deaths in France, it would be possible to create a pyramid for people who have left France. Some may object that young people are most likely to leave France and are least likely to consume medical services. This difficulty in methodology could be compensated for by bringing in other sources, notably those related to employment or unemployment, which also reveal in an anonymous manner those who have left the system.

The first innovative studies have made it possible to match in an anonymous manner data for those who receive minimum benefits (revenu minimum d'insertion) and thus to follow the inter-university trajectories of students (Quantin, Gouyon et al. 2006).

Administrative files for employment and unemployment are so rich that they provide countless opportunities:

- Prolong studies on employment and the labor force using a long-term panel;
- Follow transitions between employment and unemployment in France;
- Measure the employment of ex-students.

The third opportunity has a limitation: with the concerns underlying the creation of the law Informatique et libertés, the SSN identifier was excluded from education files, which makes it impossible to follow the careers of a cohort of ex-students. However, at the age of 20 years, students have to personally subscribe to a health insurance scheme, and their SSN is included in the social file of the university they attend. This makes it possible to follow entry into the job market for students who followed Master's Degree and PhD courses. Thanks to the administrative files, we can therefore follow in an anonymous manner year-by-year the professional outcomes for ex-students or continue in the long term the longitudinal collection of data for the survey on the labor force.

4. Conclusion

In our opinion, social files included in official statistics should no longer be devoid of any identifier, but should carry an SSN that has been hashed in an identical fashion, and then specifically encrypted. Preserve the future while providing protection against any unauthorized matching would be an economical and productive policy. Thinking about their future, epidemiologists have opened up a fruitful pathway for social and more generally official statistics. The example of the new Constances cohort shows the direction for possible matching between data for careers and healthcare. It is hoped that the same dynamic approach will be implemented for educational careers and professional careers, and between professional career and periods of unemployment. The wealth of information in administrative files is a source of great hope.

For official French statistics, the emergence of a new system, the new 'centre d'accès sécurisé distant aux données: CASD' (centre for secure remote access to data) of the CREST-INSEE also represents a significant progress for the analysis of sensitive data. The use of our technique in this framework could provide some help in linking official statistics with other data. Let us hope that current reflection on this strategy will lead to new ambitious projects.

References:

Borst, F., F. A. Allaert, et al. (2001). "The Swiss solution for anonymously chaining patient files." Stud Health Technol Inform **84**(Pt 2): 1239-1241.

Cornet, B., J. B. Gouyon, et al. (2001). "[Using discharge abstracts as a tool to assess a regional perinatal network]." Revue d'epidemiologie et de sante publique **49**(6): 583-593.

Couris, C. M., C. Foret-Dodelin, et al. (2004). "[Sensitivity and specificity of two methods used to identify incident breast cancer in specialized units using claims databases]." Revue d'epidemiologie et de sante publique **52**(2): 151-160.

Fellegi, I. P. and A. B. Sunter (1969). "A theory for record linkage." Journal of the American Statistical Association **64**(328): 1183-1210.

Fournel, I., M. Schwarzinger, et al. (2009). "Contribution of record linkage to vital status determination in cancer patients." Studies in health technology and informatics **150**: 91-95.

Goldberg, M., E. Jougla, et al. (2012). "The French health information system." Statistical Journal of the IAOS **28**(31-41).

Quantin, C., E. Benzenine, et al. (2012(a)). "Advantages and limitations of using national administrative data on obstetric blood transfusions to estimate the frequency of obstetric hemorrhages." Journal of public health.

Quantin, C., E. Benzenine, et al. (2012(b)). "Estimation of national colorectal-cancer incidence using claims databases." Journal of cancer epidemiology **2012**: 298369.

Quantin, C., H. Bouzelat, et al. (1998). "How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure." International journal of medical informatics **49**(1): 117-122.

Quantin, C., M. Fassa, et al. (2008). "Combining hashing and enciphering algorithms for epidemiological analysis of gathered data." Methods of information in medicine **47**(5): 454-458.

Quantin, C., B. Gouyon, et al. (2006). "Methodology for chaining sensitive data while preserving anonymity: application to the monitoring of medical information." Courrier des Statistiques, English Series **12**.

Sagot, P., P. Mourtialon, et al. (2012). "Accuracy of blood transfusion in postpartum hemorrhage to assess maternal morbidity." European journal of obstetrics, gynecology, and reproductive biology **162**(2): 160-164.