How to protect patient's rights to medical secret in official statistics*

Dr David-Olivier Jaquet-Chiffelle, University of Applied Sciences of Bern
Dr Jean-Paul Jeanneret, SFSO, Swiss Federal Statistical Office

*This article is the resume of a presentation given by D.-O. Jaquet-Chiffelle in September 2001, at the ISSE 2001 Conference in London (ISSE=**I**nformation **S**ecurity **S**olutions **E**urope: www.eema.org/isse).

Dr David-Olivier Jaquet-Chiffelle, University of Applied Sciences of Bern
After having received his Ph.D. degree in Mathematics in 1991, D.-O. Jaquet-Chiffelle went to Harvard University to pursue his research. He was also lecturer in the Department of Mathematics. From 1992 to 1994, he was a scientific associate at the University of Neuchâtel and worked in collaboration with the University of Bordeaux.
In 1994, he joined the Swiss Federal Section of Cryptology (Swiss Government) as a scientific expert and developed in 1996 the concept described in this article in collaboration with the SFSO.
Since 1997, he has been a full time professor of Mathematics and Cryptology at the University of Applied Sciences of Bern (Switzerland).
e-mail : jld@hta-bi.bfh.ch homepage: http://www.hta-bi.bfh.ch/~jld/

Dr Jean-Paul Jeanneret, SFSO, Swiss Federal Statistical Office

J.-P. Jeanneret received his Ph.D degree in Immunology and Epidemiology in 1990.
He has been working for the Swiss Federal Statistical Office since 1996 and was responsible for the development of the new statistics on hospitalizations. He is now responsible for the national health care statistics in Switzerland.

e-mail : jean-paul.jeanneret@bfs.admin.ch

The Swiss Federal Statistical Office (SFSO) is responsible for collecting medical data on all individuals hospitalized in Switzerland. Information on the diagnoses and on the corresponding treatments are given for all patients. The main difficulty lies in the need to recognize multiple hospitalizations without revealing the identity of the patients. A cryptographic solution has been developed. To each person corresponds a uniform anonymous linking code: some identifying data are hashed (no key involved) in the hospitals before the transfer, then these hash values are encrypted (with a secret key) by the SFSO.
Further security measures are taken to protect the hash values as well as the other data during transmission, processing and storage. In order to preserve the anonymity, the level of precision for sensitive data (date of birth, place of residence, nationality) is reduced.
The IT system that we developed is now in use at a very large scale in Switzerland.

In Switzerland, the total cost for hospitals amounts to more than 13 billion SFR/year. How is this money spent? For which kind of treatments? Could we reduce some costs? These questions led the Swiss government to order an exhaustive statistic about all patients hospitalized.

The Swiss Federal Office for Statistics (SFSO) is responsible for collecting medical data on all individuals hospitalized in Switzerland. Information on the diagnoses and on the corresponding treatments are given for all patients.

The first solution proposed by the SFSO was to slightly hide (not to encrypt) the identifying data (for example, the name of the patient was replaced by its SOUNDEX code). Even though the VESKA (Swiss Hospitals Association) had been doing an internal nominative statistic for more than twenty years, the members of the Swiss Medical Computer Science Society (SSIM) reacted very negatively to this first project. They argued that this new statistic would create a large database that would not preserve the confidentiality of the patients' medical records.

From a legal point of view, there are exceptions to medical secret when Federal statistics are involved; the SFSO could have forced health care providers to participate in the statistic. However this would have led to an open conflict. The SFSO therefore contacted the Swiss Federal Section of Cryptology to find a cryptographic expert capable of finding a solution to this problem.

From the statistical point of view, it is not necessary to know to whom a given medical record belongs; however the SFSO needs to recognize that two

different records actually belong to the same person. This is crucial in order to follow the history of the patients. At first both conditions seem incompatible. The solution that we have developed solves this paradox.

Basically we can split the data into two categories:

- medical data (diagnosis, treatment,…),
- non-medical data (last name, first name, date of birth, domicile, …).

Some non-medical data are very identifying.

The epidemiological data form the row data on which all statistical studies will be based. They contain the medical data but also some non-medical data. As long as they do not allow the identification of the patient, they are not sensitive. For this reason, in order to preserve the anonymity, the level of precision of non-medical data has been reduced to the minimum needed for the statistics; for example, we keep only the age instead of the date of birth, or the region instead of the domicile.

The non-medical data that are really identifying are not used directly in the statistics. Essentially, our solution consists in replacing these identifying data by a calculated personal code, called uniform linking code, which characterizes the patient without revealing his/her identity; it ideally satisfies the following properties:

- the identifying data allow to calculate easily the personal code of a patient,
- the personal code of a patient does not allow his/her identification,
- the same person always receives the same personal code,
- two different people always receive two different personal codes (no collision).

These properties are very similar to those of a cryptographic one-way hash function.

First of all we must choose on which identifying data the calculations will be based. These data should always be available, they should stay constant over time. If we are too restrictive in our choice, we will have many collisions. If we take too much data, they will not always be available and they could change over time with a higher probability.

Eventually, we decided to restrict the identifying data to the following set, called the minimal set of identifying data: date of birth, sex, last name and first name. Of course the last name can change during someone's life but hopefully it will not change too often and the consequences on the statistics remain acceptable. Given the practical conditions and constraints of the problem, this choice appears experimentally to be optimal.

In order for the personal code to be always the same for a given patient, it is crucial to minimize the consequences of spelling mistakes. Therefore we decided to pass the identifying data through a robust compression transformation. Again however, if the rate of compression is too high, we risk to introduce collisions (two different patients who are not distinguished).

Our robust compression transformation was tested in 1997 at a real scale on the database of the University Hospitals of Geneva (more than 222'000 records). Results were extremely encouraging: the rate of collision was only 0.3%. Moreover our transformation detected several doubles (two seemingly different patients who are actually the same person) in the hospital database; our results could help correct the database of the University Hospitals of Geneva.

If we want to recognize multiple hospitalizations, the cryptographic transformation applied by the hospitals on the minimal set of identifying data has to be the same in all hospitals and also the same over time. It would not be reasonable to make the security of this transformation depend on a secret key; a long-term secret key dealt to about 400 hospitals cannot be trusted... However if this transformation is public, the resulting linking code is not completely resistant against a dictionary attack. As a consequence, the linking code has to be encrypted, first during the transmission from the hospitals to the SFSO, then in the SFSO database.

The session key used to encrypt the linking code during transmission is generated in the background by hospital computers; entropy is given by measuring for example the time in milliseconds between two keystrokes and/or the acceleration of the mouse.

A public-key cryptosystem (RSA) is then used to transmit the value of the session key to the SFSO. Some redundancy is introduced in order to control the origin of the session key and to test the specific implementation in the hospital of both the robust compression transformation and the encryption algorithm.

After reception, the encrypted linking codes are first decrypted, then directly and uniformly re-encrypted by the SFSO: they become the uniform anonymous linking codes used as personal codes.

The RSA-private key of the SFSO and the master key that is used to uniformly encrypt the linking codes are both very sensitive secret keys. In our protocol, those two keys form a secret that is shared between three independent individuals. The protocol is based on a secret sharing scheme.

The IT-system that we developed is now in use at a very large scale in Switzerland. During the first year, it was used for exactly 482'089 cases of hospitalizations (about 38% of all hospitalizations in Switzerland during the year 1998). The uniform linking code showed that these **482'089** cases concern only **393'974** patients. It allows describing very precisely the distribution of multiple hospitalizations without revealing the identity of the patients:

Multiple hospitalizations in Switzerland[1]
Year 1998

Nb of stays	Nb of patients	Multiple hospital. rate	Nb de cases
1	333'274	85%	333'274
2	44'730		89'460
3	10'132		30'396
4	3'352		13'408
5	1'220		6'100
6	581		3'486
7	298	15%	2'086
8	149		1'192
9	90		810
10	47		470
11	32		352
12+	69		1'055
Total	393'974	100%	482'089

For 1999, more than 700'000 records have been transmitted with their encrypted linking codes. Data for the year 2000 will be sent to the SFSO at the end of the summer 2001. More than 80% of the hospitals have already made their cryptographic module validated. The SFSO expects to receive this year more than 1 million hospitalizations with a linking code. By the end of this year, all hospitals should use the system.

Since we have developed the concept, several other institutions have been interested in adapting the system and/or the protocols to their own need in Switzerland, in other European countries and in Canada. <

[1] *Source SFSO:*
http://www.statistik.admin.ch/stat_ch/ber14/statsant/statsante1_2001.pdf
This statistic for 1998 takes into consideration only 38% of all cases of hospitalizations, more precisely those that have been received with a valid linking code.

Further readings:

• Recent links
http://www.statistik.admin.ch/stat_ch/ber14/fffr14.htm
http://www.statistik.admin.ch/stat_ch/ber14/statsant/statsantel_2001.pdf
http://www.statistik.admin.ch/stat_ch/ber14/gewe/ftfr14k.htm (French)
http://www.statistik.admin.ch/stat_ch/ber14/gewe/dtfr14k.htm (German)

• Historical links
http://www.statistik.admin.ch/stat_ch/ber14/fffr14.htm (French)
http://www.statistik.admin.ch/stat_ch/ber14/statsant/fffr1402.htm (French)
http://www.statistik.admin.ch/stat_ch/ber14/statsant/fffr1403.htm (French)
http://www.statistik.admin.ch/stat_ch/ber14/statsant/fffr1406.htm (French)