Proposal for the Creation of a European Healthcare Identifier

Catherine Quantina, François-André Allaertb, Béatrice Gouyonb, Olivier Cohenc

a Department of Biostatistics and Medical Informatics. Dijon University Hospital, France
b W2 "data security" European Federation of Medical Informatics Dijon, France.
c Laboratory TIMC-IMAG UMR 5525, CNRS University of Joseph Fourier Medical School of Grenoble. France

Abstract:
In France, the European health card was created in June 2004 to increase the quality of healthcare granted to European citizen anywhere in Europe and to facilitate the reimbursement of the healthcare costs. The patient identifier included in this card is essentially based on the healthcare insurance number of the patient and does not allow any linkage with his (her) previous health care data if he (she) is affiliated to another national healthcare insurance system when working for a long duration outside France. The purpose of this paper is to present the concept of a personal identifier based on familial components which has been validated by the French authority for personal data protection in the framework of a genetic study. Results issued from the Burgundy perinatal network demonstrate the interest and the feasibility of adding a maternal component to the individual component of the newborn to allow Mother/new-born healthcare data linkage after anonymization. The advantage of adding a familial component to the healthcare insurance number is debated. This proposal will permit to link the data of a patient even when residing outside his country in Europe. It will also contribute to establish European public health statistics by matching healthcare data of the patients’ records with other administrative data (mortality, social information ...) after anonymisation of these data in accordance with the European directive on data protection.

Keywords: Data linkage, Human identification, Confidentiality, Hash coding, European health card, Genetics, Perinatology

1. Introduction

Even if the harmonization of healthcare system will still take time, as soon as 1996 the European parliament recommended the development of a European health card in order to allow any European citizens to benefit of adequate care everywhere in Europe and above all when moving from one country to another one. Following the work conducted in the framework of the European project "Netcard" the European health card was created in June 2004. Thus, all French citizens will not use anymore the E111 form when travelling abroad inside Europe but will carry with them their European healthcard which is distributed by the French health insurance system. Today this is still a paper document but it should rapidly move towards a smart card [1]. The setting up of such a European health card intends to simplify the healthcare reimbursement for all European citizens which are travelling for leisure or work but also wants to contribute to a better quality of care by making easier an European access to medical data in the respect of its legal framework. The creation of this card gives also some new possibilities to gather the medical data needed to conduct epidemiological studies at the European level and to provide the global information
required to manage public health at a European level. Such statistics would require to be able to match different categories of data issued from different origins, not only from the personal medical record we try to implement by the GPs in ambulatory care but also from hospital, cancer registers and governmental mortality data, according to some models which have been already used in the United Kingdom [2], in northern European countries and in the USA [3,4], or in Australia [5]. This could also offer the possibility to conduct inter countries comparisons inside Europe. However, this data collection requires solving two main difficulties. First it must respect the European directive on data protection and must provide high guarantees concerning the quality of the linkage of the data of an individual. It implies that the chosen identifier must be sufficiently discriminating to avoid the risk of gathering the data of two set of information belonging to two different patients. The European health card already includes the social insurance number which has been given to the person in his own country. Unfortunately, this number is not an identifier sufficiently powerful to satisfy the objectives previously described: the patient may be also registered in the insurance system of another European country if he works there. Finally, even in the country where this social insurance number is the main identifier in the health domain, epidemiologists agree [6,7], on the fact that this identifier is not by itself sufficient to link healthcare data issued from the medical records with mortality statistics. The purpose of this paper is to describe the concept of family-based identifier. This will allow to show the interest of introducing this identifier in the European health card, apart from the social insurance number, in order to link all the data of a patient even when he moves and permit the production of public health statistics at the European level.


The French authority for data protection has validated this proposal for the linkage of genetic data.

**Principle:** The development of medical genetics has pointed out the importance of the familial dimension of the medical information and the need for introducing a familial component in the individual record of the patient. The interest of such familial component is supported by the identification of genetic factors influencing the outcome of some diseases or the therapeutic response to the drugs. In this framework we have set up an Internet application which gives the possibility to authorised geneticians and researchers to gather all the medical data concerning a patient with a genetic disease. By the use of anonymised family-based identifiers, data concerning the patient will be gathered along with the information issued from the medical records of the other members of his family. With the patient’s agreement, this anonymised identifier is transmitted by a secured system to the data centre called HC-Forum platform [8] with all his medical data. Thus the patient will benefit from a medical record accessible from any medical centre. This record will be regularly updated with his own information and those of his family. In this data centre, for statistical purpose, this identifier will be anonymised a second time in order to avoid any kind of dictionary attacks [9]. The rules governing the information system are strictly in accordance with the European directive for data protection in order to guarantee the confidentiality of his data and of his relatives, and to grant the subject the right to obtain the rectification, of data. The French National Data Protection Authority (CNIL) validated this project in March 2004 (advice n° 04-006). A patent was filed to the French Patent Office in September 2004.

**Composition of the family-based identifier:** The choice of the criteria included in the individual part of the patients’ identifier is based on a previous work we have conducted, with the French Department for the Modernization of Health Information Systems, which have demonstrated that the key information for patient identification are his first and last
names and his date of birth [10,11,12]. The first name is the first one recorded in the register of birth and the last name is the family name which means for the women their maiden name and not their marital name for example. The familial components (last name, first name and date of birth of the mother and the father) were added to the individual component following the work conducted by our department on the linkage of the mother and her babies medical records in the Burgundy perinatal network [13] which has pointed out the necessity of introducing the maternal component in the individual identifier of the baby. However, we also need a paternal component to rebuild the genealogic tree.

**Anonymisation of the family-based identifier:** Hash algorithms are not reversible and cannot be deciphered and one of the most reliable [18,14] and freely available one is the Standard Hash Algorithm (SHA). The three key variables included in the individual part of the patient identifier (Last name, first name and date of birth) are separately hashed in order to maintain a higher security level. For the individual component, we have three variables: Hn: anonymous number corresponding to the last name of the subject; Hp: anonymous number corresponding to the first name of the subject; Hdn: anonymous number corresponding to the subject’s date of birth. The family based identifier includes 9 variables. This anonymisation is made locally before any kind of data flows in such a way that only anonymised data are sent and made available on the HC forum data centre.

**Familial linkage:** By using the link existing between the identifiers of the subjects of a same family, the genealogic tree of the patient can be described. Thanks to this linkage, one can thus build the family tree from a "vertical" point of view, i.e. ascending-descending of an individual. This linkage also makes it possible to build the family tree from an "horizontal" point of view, i.e. within the same generation, for the following cases: - the phratry: by sorting the dates of birth of all the individuals having the same parents; - half-brothers and sisters: by sorting the dates of birth of all the individuals having the same father or the same mother;

### 3. Results obtained in the Burgundy perinatal network

The Burgundy perinatal network was initiated in 1992 to improve the quality of perinatal care in Burgundy, a French region with 1.800.000 inhabitants and 18.500 annual births and including 18 private and public hospitals. A multidisciplinary working group previously chose and precisely defined specific indicators. These items could be mainly obtained from the discharge abstracts mandatory for each hospitalised patient in both public and private hospitals in France. These data are previously rendered anonymous before being sent from each hospital to the committee in charge of the assessment of the perinatal network’s performance. The use of the ANONYMAT software has been authorised by the French Department for Information System Security (SCSSI) and the management of perinatal data in Burgundy was specifically authorised in 1998 by the CNIL. An optimal assessment of perinatal care obviously needs an effective linkage between maternal and neonatal data. Indeed, taking into account the mechanisms of neonatal diseases made essential the linkage between: 1) abstracts of a mother and her corresponding neonate even if they were not cared for in the same hospital, 2) all discharge abstracts obtained from the same patient, who may have several discharge abstracts from different hospitals. We present here the results of the performance assessment of the file-linkage process of maternal and neonatal data used for the evaluation of the Burgundy perinatal network.

**Population:** All deliveries and newborn births, whenever alive or not, are considered in the perinatal network if the gestational age is at least of 20 weeks in pregnancy and/or if the birth weight is greater than 500g. For the purpose of this study, we took into account the population included since 1998, year of the beginning of the data collection, up to 2003. In
1998, only nine hospitals participated into the data collection, whereas since 2001 all the 18 hospitals involved in the regional perinatal care have provided items for 100 % of the 18,500 annual births.

**Data Collection:** Apart from discharge abstracts collected for all mothers and all neonates, five identification items (maiden names, first names and of dates of birth of mothers, first names and of dates of birth of neonates) have been added for both mothers and neonates to allow the linkage between their discharge abstracts. This identifier corresponds thus to the family-based identifier previously described, except for the identification of the father. The last name of the baby was not recorded as this name often changes during the first days after birth.

**File linkage** Once rendered anonymous, data were transmitted to the regional audit committee located in the University Hospital of Dijon (France) and were included in a regional database. A statistical model, taking into account the five identification variables, accomplished the linkage[15]. Three sets of possible decisions can be determined as follows: 1) the pair is matched, 2) no determination is made, 3) the pair is not matched. In case of no determination, a paediatrician validated the linkage by checking the data accuracy of the five identification variables as well as all medical information contained in the discharge abstracts of the neonate(s) and the mother.

**Data validation:** The exhaustiveness for the number of mothers and neonates registered in the regional database was assessed from hand-written notebooks which are used in each hospital for the registration of birth and/or admission of sick neonates in units caring for these infants. The quality of both medical items and linkage items were monitored as follows: 1) Exhaustiveness for the five linkage items was controlled for each discharge abstract; 2) Exhaustiveness for gestational age and birth weight was controlled for each neonatal discharge abstract; 3) For each patient, a paediatrician looked for inconsistency between medical data or between dates of exit and admission for successive hospitalisations. Computerized procedures have been developed to disclose those discrepancies. Correction of erroneous data was then performed on the nominative files in each hospital.

**Reliability of the Linkage Procedure** It was assumed that an erroneous link between a mother and a non-corresponding neonate was excluded in the case of a perfect agreement between two records on the five identification items. Indeed, it was highly unlikely that in the same maternity hospital, two women having the same maiden name, the same first name, the same date of birth would give birth the same day to babies with the same first name. Additionally, the risk for an homonym error was very low ($10^{48}$) with the standard hash algorithm [11]. In a perinatal network, it is obvious that every neonate has a mother and vice versa. So, the fact that no link was found between a neonate and a mother will indicate either a linkage error or the lack of the mother record. The identification of linkage errors was performed using again the linkage method on the basis of only five items or even four. To verify these potential links, each hospital was asked to control the identification items of the records corresponding to the given anonymous numbers. The corrected data were rendered anonymous before being sent again from the hospitals to the regional database. Finally, the linkage is only performed if the five identification items are perfectly identical in the mother’s and neonate’s records.

**Results:** In 1998, 9 hospitals were involved in the collection of discharge abstracts. The percentage of mothers retrieved in the regional database after the validation procedure was 99.1 % of all eligible mothers; those percentages were of 98.7 % for neonates. In 2003, 18 hospitals were involved in the collection of discharge abstracts and the overall exhaustiveness for both mothers and neonates was 100 % after validation procedure. In 1998, the five items used for the linkage procedure were recorded in 80 % of discharge abstracts before validation and in 99 % after. In 2003, the exhaustiveness of these items
were 93.7% before and 100% after validation. Before validation, the percentage of neonates linked to their corresponding mothers on the basis of the identification items were 71% in 1998, and 86.3% in 2003. After validation, 99.9% of neonates were linked to their mothers whatever the year concerned.

4. Discussion

4.1 Discussion of the result obtained in the burgundy perinatal network.
Optimal assessment of perinatal care needs a linkage procedure between successive files from the same patient and between files of the mothers and their corresponding neonate(s). This latter linkage was found to be mandatory in assessing the postnatal consequences of antenatal risk factors and maternal diseases. For instance, one maternity hospital showed in 1999 a significant increase both in the rates of Caesarean section and of neonatal hospitalisation for respiratory distress as compared with regional and national data rates. The linkage procedure disclosed that the excess in neonatal hospitalisation rate was related to an excess in caesarean section rate. This finding was a strong argument in reducing the Caesarean section rate in this hospital. The fact that at each mother corresponded a neonate and vice versa was particularly helpful in testing the linkage procedure. Indeed, coupling the direct linkage of anonymous data files to the validation procedure gave very satisfactory results, on a regional scale, with 99.9% of neonates linked with their mothers. Different types of errors were found during the validation procedure. The most frequent one corresponded to errors in spelling names leading to phonetic changes not subsequently corrected by the spelling process included in the ANONYMAT software. The inversion of the married name and the maiden name was also responsible for linkage failures. These results thus demonstrate the importance of a specific validation of the patient’s identification, within each hospital databases, before the extraction of any information to the perinatal network committee, using appropriate methods for the reduction of the doubleton’s rate [17, 18].

4.2 Proposal for using the family-based identifier to create a unique European identifier.
We propose to add in the European health card, apart from the national social insurance number of the patient in each country, a family-based identifier which could contribute to harmonise patients' identification at the European level. This solution would lead to the creation of a European health identifier which would allow to gather data patients anywhere in Europe, whatever their location, even if their social insurance number change according to their country of residence. This will be useful, at the individual level, to provide higher quality in health care due to a better follow-up of the patient and to facilitate the reimbursement of health care costs. At the community level, this will increase the reliability of public health statistics. The great advantage of this European health identifier based on the family component is to be founded on very basic information available for everybody, easily checkable and permanent during all the patient's life which is not the case for the social insurance number. To fulfil the rules of the European directive on data protection and of the medical deontology, the use of hash algorithms with different keys would allow to create, from the information of the family-based identifier, different anonymised identifiers for a same patient which could be used in the different countries for different applications ie Medical personal record, healthcare network, epidemiological or clinical studies. The fact that all these anonymised identification numbers are issued from the same initial information will facilitate the linkage of the data of a same patient, coming from different health information systems, in a secure environment. If an error occurs on one of the identification criteria (Last name, first name or date of birth) of the patient or of the parents, the use of a probabilistic linkage algorithm as described for the Burgundy perinatal network, could be
used to recreate the link with the other parts of the medical information and to rebuild for example the medical record of a patient. This is not always possible today when there is an error in the social insurance number, particularly when this number is not significant, meaning that it is not made using characteristics of the patient (date of birth, gender …).

5. Conclusion

Our proposal for using the family-based identifier to create a unique European identifier will permit to link the data of a patient even when residing outside his country in Europe. It will also contribute to establish European public health statistics by matching healthcare data of the patients' records with other administrative data (mortality, social inform, ) after anonymisation of these data in accordance with the European directive on data protection.

6. References


Address for correspondence

Professeur Catherine QUANTIN, Service de Biostatistique et Informatique Médicale, Centre Hospitalier Universitaire, BP 77908 21079 DIJON CEDEX, Tel: 33 3 80 29 36 29, Fax: 33 3 80 29 39 73
Email: catherine.quantin@chu-dijon.fr

Section 13: Public Health Informatics, Clinical Trials