Is De-identification of Electronic Health Records Possible? OR Can We Use Health Record Corpora for Research?

Hercules Dalianis† Gunnar Nilsson†‡ Sumithra Velupillai†

†DSV/KTH-Stockholm University,
Forum 100, 164 40 Kista
‡Department of Neurobiology, Care Sciences and Society,
Center for Family and Community Medicine,
Karolinska Institutet
Sweden
hercules@dsv.su.se, gunnar.nilsson@ki.se, sumithra@dsv.su.se

Introduction and Background

Today an immense volume of electronic health records (EHRs)[1] is being produced. These health records contain abundant information, in the form of both structured and unstructured data. It is estimated that EHRs contain on average around 60 percent structured information, and 40 percent unstructured information that is mostly free text (Dalianis et al., 2009). A modern health record is very complex and contains a large and diverse amount of data, such as the patient's chief complaints, diagnoses and treatment, and very often an epicrisis, or discharge letter, together with ICD-10 codes, (ICD-10, 2009). Moreover, the health record also contains information about the patient's gender, age, times of health care visits, medication, measure values, general condition as well as social situation, drinking and eating habits. Much of this information is written in natural language.

All this information in a health record is currently almost never re-used, in particular the parts that are written in free text. We believe that the information contained in EHR data sets is an invaluable source for the development and evaluation of a number of applications, useful both for research purposes as well as health practitioners. For instance, text mining tools for finding new or hidden relations between diagnoses/treatments and social situation, age and gender could be very useful for epidemiological or medical researchers. Moreover, information concerning the health process over time, per patient, clinic or hospital, can be extracted and used for further research. Another application is the use of this data as input for simulation of the health process and for future health needs. Also, such huge health record databases can be used as corpora for the generation of generalized synonyms from specialized medical terminology constitutes another exciting application. We can also foresee a text summarization system applied to an individual patient's health record, but using knowledge from all text records and conveying the information in the health record at the right level to the specific patient.

The data can also be used for developing methods where clinicians in their daily work get automatic assistance and proposals of ICD-10 codes for assigning symptoms or diagnoses, or for validating the already manually assigned ICD-10 codes.

However, the development of such applications and methods require access to data, and this is not easy since EHRs frequently contain sensitive information given to the physician in trust by the patient and permissions need to be granted through official channels. Health care involves serious ethical issues, and the Hippocratic Oath is a very important principle. The patient communicates with the physician in trust, and the physician knows not to communicate this trust outside the hospital or to any person not involved in the treatment of the patient. In order to address the need for data for research purposes while retaining the respect of patient integrity, it may be possible to de-identify the data prior to releasing it for research. Such methods do however pose many important questions.

Previous Research

In early research regarding EHRs the main concern was medical privacy when moving from paper based health records to electronic ones, but also the possibility to use

[1] In the literature, EHRs may also be called electronic patient records (EPRs) or electronic medical records (EMRs). Here, the term (electronic) health record (EHR) is used.

EHRs for research by performing anonymization[2] of the data. (Barrows and Clayton, 1996). Barrows and Clayton (1996) describe various technical methods to protect the medical privacy of the individual patient during daily clinical work. In Arning et al. (2007) anonymization on genetic data is discussed, and also questions regarding who should have the permission to re-identify genetic data when finding some medical reasons for this. Another issue that is discussed is the ownership of the anonymized data, but none of these articles mention which data to remove or anonymize.

In the U.S., the Health Insurance Portability and Accountability Act (HIPAA 2003) stipulates which types of information risk the possibilities of identifying a patient. They define 18 identifiers, Protected Health Information (PHI), which must be removed in order for the data to be considered de-identified. These have been used for the development of automatic de-identification systems, with modifications in many cases (see for instance Uzuner et al. (2007) and Velupillai et al. (2009)).

Panel Discussion

There is no doubt that the information contained in EHRs is valuable for further research. However, the possibilities of developing methods that could aid both researchers and practicing clinicians are restricted due to confidentiality reasons and the like. De-identifying the data sets prior to distributing them is crucial in order to keep patient integrity intact. However, many questions are left unanswered.

It has been showed that PHIs can be identified to some extent both manually and automatically, but will the records still be useful for research if essential information is removed? Which PHIs should be removed? Why should they be removed? What is actually contained in these PHIs? What about ethnicity or occupation? Can we really remove all PHIs? Will the PHIs that are removed distort the useful information in the patient record? To which extent could the PHIs provide valuable information to the future applications described above?

De-identifying or anonymizing information contained in structured fields of EHRs, such as social security numbers, is technically not complicated and can be done with high confidence. Nevertheless, a lot of identifiable information is also written in free text, such as names and contact details of family members, etc. Due to the nature of natural language, it will never be possible to ensure PHI identification with 100 percent accuracy, manually or automatically. Is such an assurance necessary? If not, what is considered reasonable and why? Also, even if a guarantee is given that a particular EHR does not contain any identifiable instance at all, how do we measure the risks of re-identification given that external data was added? Moreover, how do we know that instances not defined as PHI in combination may constitute a unique identifier?

As researchers, how can we safeguard the patient's trust in the health care system so the patient continues to provide confidential information to the physician and other clinicians? Should we inform the patient that we are using his/her patient record for research? What types of methods of obtaining patient consent could be developed?

We believe that EHRs are going to be (re)used more for various tasks in the future. As more and more health care processes are digitalized, it is crucial to define and discuss ethical issues regarding the use of such private information for various purposes.

[2] Anonymization is often distinguished from de-identification in that identifiable information is removed, whereas it is masked or replaced in the latter case. The entities themselves are, however, the same.

References

Arning, M., Forgó, N. and Krügel, T. 2007. Data protection issues with regard to research in genetic data. In *Proceedings of the 2nd Workshop on Personalisation for E-Health*, June 26th, Corfu, Greece.

Barrows R. C. Jr and Clayton, P. D. 1996, Privacy, confidentiality, and electronic medical records. *Journal of the American Medical Informatics Association*, Vol 3, 139-148.

Dalianis, H., Hassel, M. and Velupillai, S. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. To be published *in Proceedings of the 14th International Symposium for Health Information Management Research, Kalmar, Sweden, 14-16 October, 2009*.

HIPAA. 2003. Health Insurance Portability and Accountability, (HIPAA) Privacy Rule and Public Health Guidance. From CDC and the U.S. Department of Health and Human Services, April 11, 2003. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm. Accessed April 24, 2009.

ICD-10. 2009. International Classification of ICD, http://www.who.int/classifications/icd/en/. Accessed, June 18, 2009

Uzuner, Ö., Luo, Y. and Szolovits, P. 2007. Evaluating the State-of-the-art in Automatic Deidentification. *Journal of the American Medical Informatics Association*, September 2007,14(5): 550-563

Velupillai, S., Dalianis, H. Hassel M. and Nilsson, G. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. In *International Journal. Medical Informatics*. (2009), doi:10.1016/j.ijmedinf.2009.04.005.