Khaled El Emam (PhD)

Data Anonymization Practices in Clinical Research

A Descriptive Study

University of Ottawa
CHEO RI, 401 Smyth Road
Ottawa, Ontario K1H 8L1

May 8, 2006

This report was written for the
Access to Information and Privacy Division of Health Canada.

Table of Contents

Introduction

There is increasing adoption of IT in health research and practice. For example, the use of stand-alone Electronic Data Collection (EDC) systems is rising in clinical research[1, 2], and the adoption of Electronic Medical Records (EMRs) is growing[3-5]. Both the Canadian and US governments have allocated considerable funds to maximize the adoption of the EMR across the health care system. Researchers are increasingly turning to EMRs as a source of clinically relevant patient data. In cases such as disease registries where research data comes from usual practice, the EMR *is* the electronic data collection system.

Therefore, there is a trend towards collecting, storing, and exchanging health information electronically. The ease of storage and exchange of large volumes of health data electronically has raised privacy concerns[6-9].

A majority of patients, and the public in general, are concerned about unauthorized use and disclosure of their Personal Health Information (PHI) in an era of the electronic medical record[10-14]. Some of these concerns are justified in that privacy violations have ocurred[15, 16]. There is some evidence that existing medical training does not provide adequate coverage of practices for protecting EMR privacy[17]. Basic practices for protecting patient privacy are not followed[18]. Serious vulnerabilities in Internet-based health data collection systems have been reported[19].

One of the mechanisms to safeguard PHI is to anonymize it. This means remove or obfuscate any identifying information about the individual patients in the data set, hence making the re-identification of those individuals very difficult.

The focus of this report is to understand how PHI is anonymized, if at all, in the Canadian clinical research context. Our objectives are to answer the following three questions:

- Under what conditions is data being anonymized in clinical research studies in Canada ?
- If data is being anonymized, how is it being anonymized ?
- Are these anonymization practices adequate ?

To answer these questions we conducted an interview study on anonymization practices in clinical research. In the next section we provide a set of brief definitions and describe the methodology used to collect the data. This is followed by the detailed findings, and we conclude with recommendations on how to improve anonymization practices.

Methods

We followed an exploratory, descriptive research method, namely *grounded theory* [20]. This involved an iterative process of data collection and analysis to identify anonymization practices, and where applicable integrating these findings with the literature. The result is a summary of anonymization practices that is grounded in the realities of clinical research projects.

Different stakeholder groups were interviewed. Twenty individuals (2 clinical researchers, 2 members of academic Research Ethics Boards (REBs), 9 experienced study coordinators, 5 privacy specialists, and 2 IT/security specialists) were interviewed. Interviewees were identified through a literature search and expert recommendations. The role of interviewees was weighted heavily towards study coordinators because it is they who had the most detailed information about what actually happens during the research projects and the reasons for that. All interviewees were in Ontario or Quebec sites.

The interviews were conducted in the Summer and Fall of 2005. The interview plan asked for a recent list of clinical research studies that an interviewee was involved with, and then for each study we probed to identify the anonymization practices that were followed.

Results

In presenting the results we will denote a person conducting the study (e.g., collecting data) as I , and a subject in the study as S . We assume that all data collected is kept in a r*esearch database*.

1 Examples of Inappropriate Disclosure

We will start by presenting some examples of inappropriate disclosure practices for protecting this kind of data that were identified during the interviews. This clearly illustrates the need for anonymization practices in clinical research settings:

1. In one case engineering and mathematics graduate students were participating in a study that involved the analysis of medical images. These students did not receive sufficient education on privacy issues and how to handle PHI. Consequently, they were exchanging the personal data of subjects among themselves by email without any encryption.
2. There were reported cases of study participants taking data home to finish some work off by saving it on to a memory stick or emailing the information to public accounts that they can access from home (e.g., Sympatico or Rogers accounts). The data that was taken home was not encrypted.
3. In one study progress notes had to be completed in an electronic data collection system during a patient visit. There were cases where the physician or nurse completing the clinical notes mentioned the patient's name, family physician name, sibling or parent names, or other identifying information in what they wrote. Therefore, even if the structured questionnaires used to collect data in a clinical research study exclude any identifying or potentially identifying information, patients can potentially be identified from the clinical notes that were submitted as part of the study.
4. Another scenario involved the audit trails. If say a nurse saved identifying information in the notes or comments section in an electronic data collection form and then subsequently deletes that information, the information remains in the audit trail. In this scenario patients were re-identifiable through data that was available in the audit trails.
5. When an EDC was used, examples of password sharing (to avoid having to re-login every time an individual was to work on a shared computer) and passwords written on notes posted on monitors were common.

The above of course is not a comprehensive list, but it does illustrate the types of disclosures that are occurring today in Canadian clinical research studies. These examples strengthen the case for ensuring that clinical research data be anonymized.

2 When Anonymization Is Required

Anonymization practices in clinical research studies are strongly influenced by the REB requirements. In many cases the REB requirements were not clear, therefore the *perception* of the REB requirements played an important role in determining anonymization practices.

The following were the conditions we identified where an REB was more likely to require that the data be anonymized:

1. If a study did not require consent (if the REB waived the consent requirement for secondary use of data, for example), then anonymization was often a requirement. This would occur in retrospective studies involving chart reviews or in prospective disease registries.
2. If there is an intention of linking collected patient records to another database then the REBs typically insisted on additional measures to anonymize the linked data. For example, if the clinical research data that was going to be collected was to be linked with an external administrative database.
3. When electronic data collection systems were used and any PHI was going to leave the institution's custody (e.g., to another hospital) then anonymization was often required. The concern was that if data about the institution's patients was going into someone else's hands then it required additional protection.
4. When mobile devices were used to collect data, anonymization was often required. This was driven by fear that if the mobile device was lost or stolen then S's would be easily identifiable by examining the data on the lost/stolen device.
5. When sensitive data was being collected as part of the study (e.g., HIV status, mental health information) an REB was more inclined to require that the data be anonymized.

For other studies that did not meet the above criteria there was no clear guidance about whether the data should be anonymized or not: a decision was made on a case by case basis.

It should be noted that, as a general statement, the above concern about mobile devices is unfounded because these can be easily encrypted and remotely managed (for example, disable the device if it was reported stolen).

3 Anonymization Practices

Data anonymization can be applied during different activities in a typical clinical research study: collection, retention, and disclosure. These three activities are sequential in a study: data is collected, retained for the duration of the study, and then disclosed. The diagram in Figure 1 shows the workflow among these activities.

If data is collected anonymously, then by definition it is anonymized during retention and disclosure. If the data is anonymized during retention then that data will be anonymized during disclosure. The converse is also true. If data is identifiable when it is collected then it will still be identifiable when it is retained in the research database unless something specific is performed to anonymize it.

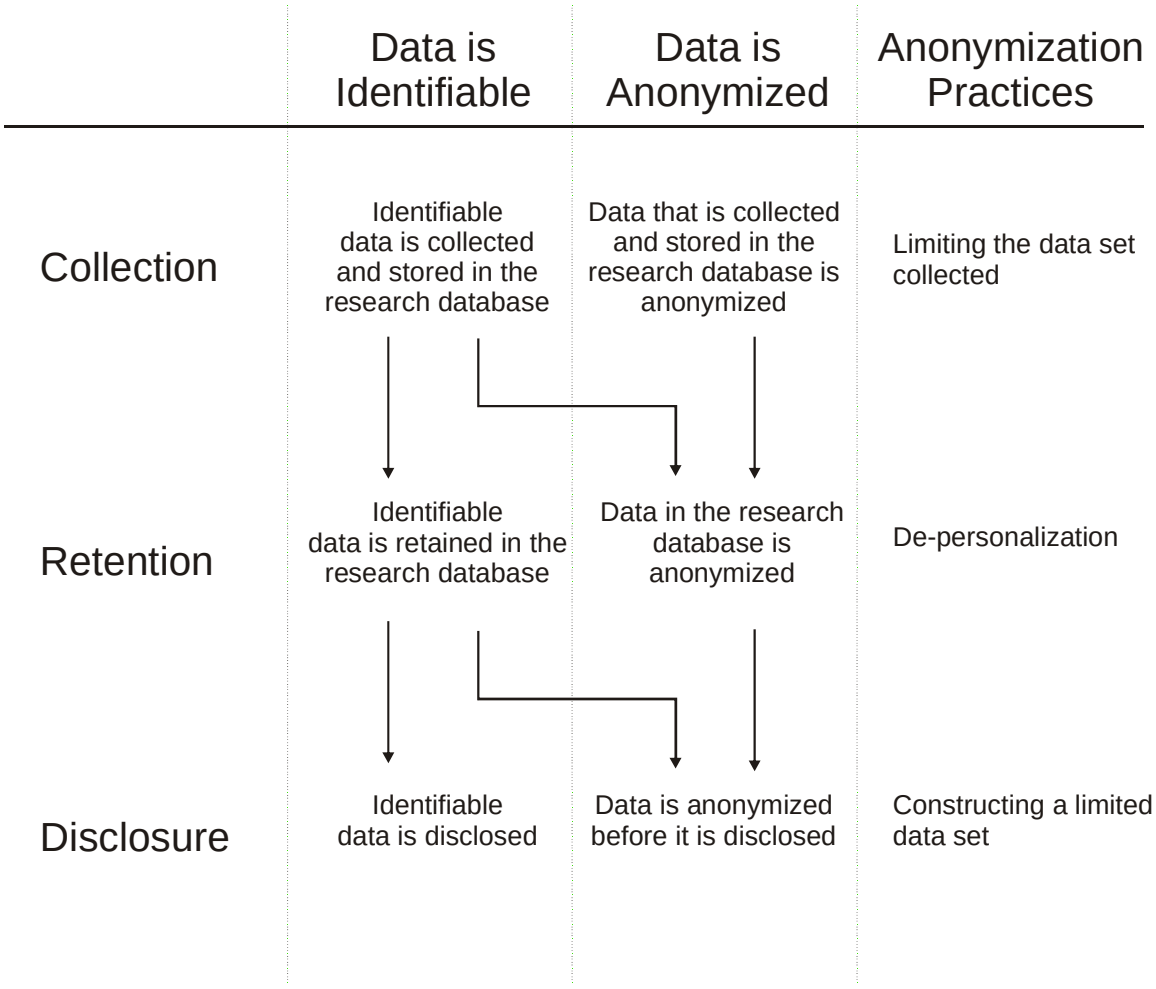

Figure 1: *Workflow summarizing anonymization practices. The assumption is made here that no specific action is taken to re-identify an anonymized data set.*

In the following sub-sections we describe the practices that are followed to anonymize data sets during each of the three activities.

3.1 Collection

Data can be collected in an anonymous way. There are two scenarios for anonymous data collection: (i) at the outset data is collected from S anonymously, and (ii) identifiable data is collected but then destroyed soon after collection.

When I does not know who the S is then data can be collected anonymously. For example, if I is performing a survey on the web and does not collect identifying information in the on-line questionnaire, does not retain the IP addresses of the client machines in the log files, nor use cookies, then one can make the case that data is being collected anonymously.

When I does know who the S is, I may still not collect any identifying information about S to be put in the research database. For example, S may be a patient of I and is being recruited into a registry. The data entered into the registry may be anonymous.

The second scenario is when identifying information is collected but then removed soon after collection. For example, during an on-line survey the email address of the respondent is collected to send the respondent a confirmation email that their response has been received. Then the email address and the copy of the confirmation email sent are destroyed. That way the identifying information is collected temporarily but not stored in the research database.

In both scenarios it is important to determine which variables are identifying or potentially identifying so as not to collect them or to destroy them soon after collection. This practice is called limiting the data set. The REB most commonly provides guidance on which variables not to collect.

The obvious identifying variables were always removed during anonymization. These include variables such as: the health card number, hospital ID, name, telephone number, and address. However, there was wide variation among interviewees on which additional variables have a high risk of re-identification, individually or in combination. There was uncertainty about whether postal codes should be collected or not. Some studies used patient initials as an identifier (for example, to identify the patient when they come back for a second visit). But because initials are not unique in large studies, date of birth was used as well as initials. There was uncertainty about whether date of birth can be used to re-identify individuals.

Some sites used the 18 variables specified in HIPAA[*] as a basis for anonymization. If a researcher wanted to collect any data elements out of those 18 then they would need special permission from the REB after making a case for why that variable was necessary and what mechanisms will be put in place to protect the data once it is collected.

3.2 Retention

Data retention here means storing the data during the course of the study. Whether it makes sense to anonymize data during retention will depend on the specifics of the study design.

There are a number of situations where permanent anonymization of data in the research database would create practical difficulties, for example:

- If each S needs to come back to the study site for multiple visits, then it is important to match the S with their record in the research database at each visit. In such a design some identifying information must be kept in the research database.
- If an adverse event or a serious adverse event were to occur during or soon after the study completion, it would be necessary to identify the individual's record in order to complete the appropriate adverse event reporting forms,

[*] HIPAA is the Health Insurance Portability and Accountability Act in the US. As part of this legislation there is a Privacy Rule that specifies a list of 18 items that would make a data set identifiable.

link with previous adverse events, and possibly file a report with Health Canada.

- If an S is allowed to withdraw from the study and withdrawal implies removal of all of their data then it is important to be able identify which records in the research database belong to that S.
- During the regular monitoring of say a clinical trial the monitors would need to check the records in the research database against the source documents (e.g., patient charts and lab results). This would be very difficult to do if it was not possible to identify the records in the research database.

Because of the need to know who the Ss are under certain conditions, data was often *de-personalized* rather than permanently (and irreversibly) anonymized. De-personalization means that data with identifying information is collected, but the identifying information is then severed from the PHI in the research database and is stored separately in an identification database. A code is attached to the research database and the same code is used to link with the identifying database. This scenario is illustrated in Figure 2. The identification database may be electronic (e.g., a spreadsheet) or on paper. Each site in a study would need to have someone with access to the identification database. The identification database may be held by (senior) researchers or may be held by a third party.

Figure 3 shows another scenario where a linking database is used. This is most suitable if a third party will hold the linking database to ensure that the records in the research database cannot be re-identified by any of the participants in the study.

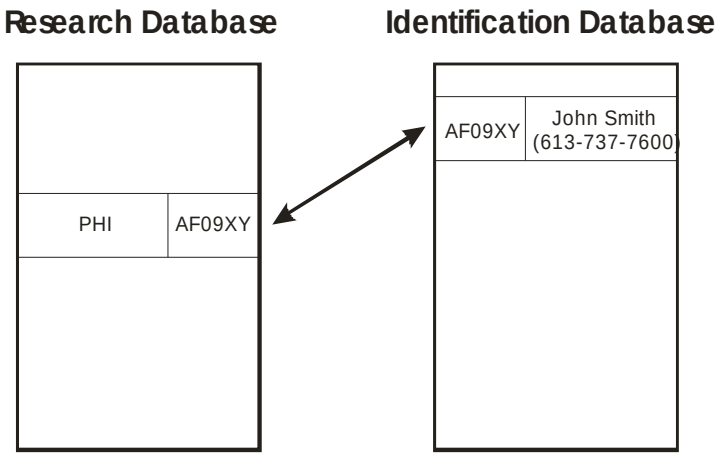

Figure 2: *An example of severing the PHI from the identifying information and using a code to link the two data sets.*

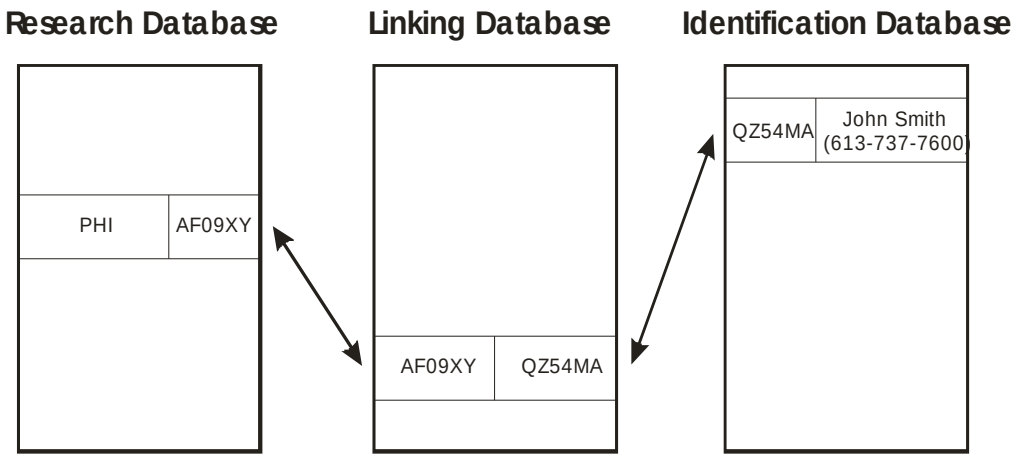

Figure 3: *An example showing the use of a separate linking database to match the subjects in the research database with the subjects in the identification database.*

It should be noted that in none of the cases that were covered during the interviews was it reported that the data was encrypted in the research database.

3.3 Disclosure

Disclosure occurs when the research database is made available either outside the research institution or to individuals not taking part in the study. For example, in a multi-center clinical trial the data is eventually sent to a study coordinating center that collects all of the data and produces a consolidated and validated data set for the whole study. The coordinating center may be in another part of the country or in another country; it may be an academic institution or a Contract Research Organization. Disclosure also occurs if the data set is sent to an external researcher, for example, to perform additional analysis on the research database.

Anonymization for disclosure always means the removal of direct identifiers (e.g., name and address). Most sites will take additional actions to anonymize the data. The most common practice is to limit the data set (as we discussed for *collection*) before it is disclosed.

Discussion and Recommendations

None of the interviewees knew about the use of statistical techniques to anonymize data during any stage of the research process. The two most common anonymization techniques were: limiting the data set and de-personalization. The use of these practices was driven largely by actual or perceived REB expectations and requirements.

There was wide variation in what were believed to be the key variables that must be removed during data limitation. The variation we found makes the conduct of multi-center studies very difficult. For example, if one center considers date of birth as a potentially identifying variable and all of the others did not, it would be necessary to change the protocol for just that one center. Taking into account each center's sensitivities may result in multiple critical variables being excluded.

Even though the privacy landscape is very different between the US and Canada, the use of HIPAA as a guide for what variables to exclude is a reasonable approach. Removal of the HIPAA variables would constitute a conservative approach because not all risky variables in the US will be risky in Canada.

De-personalization practices are based on the assumption that the information to link individuals with the research database is stored in a safe way and that there are procedures in place to protect it. We were not able to ascertain that in detail during our interviews.

Within each site, the anonymization practices were the same across studies. This is consistent with the finding that the REBs drove anonymization practices, reducing variability within sites.

There is a need for improving existing clinical research practices as they pertain to privacy. We can therefore make a number of recommendations:

1. Provide privacy training to participants in clinical research studies. The use of IT in research projects and the extra practices needed to protect data in that environment should be covered.
2. There is a need for stronger privacy expertise on REBs to allow them to provide clear guidance on how to anonymize data.
3. To reduce the inconsistency in anonymization practices across REBs in Canada, research is needed to determine which potentially identifying variables are really risky and under what conditions. Thus far no work on re-identification risk has been performed in Canada.
4. A stronger emphasis should be placed on the use of statistical disclosure control techniques to anonymize data sets when they are disclosed. Removing variables may limit the types of analysis that can be performed,

hence reducing the value of the data set. However, other statistical techniques may provide protection against re-identification but still retain the variables in the data set.

This descriptive study has a number of limitations related to its scope. First, the scope was Ontario and Quebec only. Individuals in Western and Maritime provinces were not interviewed. Second, we did not address the collection of DNA samples and their use in health research. The focus was on clinical variables.

References

1. Paul J, Seib R, Prescott T. The Internet and Clinical Trials: Background, Online Resources, Examples and Issues. *Journal of Medical Internet Research.* 2005;7(1):e5.
2. Borfitz D. Conspiring Forces Behind EDC Adoption. *CenterWatch.* 2003;10(2).
3. Irving R. *2002 Report on Information Technology in Canadian Hospitals*: Canadian Healthcare Technology; 2003.
4. HIMSS. *Healthcare CIO Results*: Healthcare Information and Management Systems Society Foundation; February 2004.
5. Andrews J, Pearce K, Sydney C, Ireson C, Love M. Current State of Information Technology Use in a US Primary Care Practice-based Research Network. *Informatics in Primary Care.* 2004;12:11-18.
6. Norsigian J, Whelan S. Privacy and medical records research. *New England Journal of Medicine.* 1998;338(15):1076-1078.
7. American College of Epidemiology. Statement on health data control, access, and confidentiality. http://www.acepidemiology.org/data.html. Accessed 1st May, 2005.
8. Health Privacy Working Group. *Best Principles For Health Privacy*: Institute for Health Care Research and Policy, Georgetown University; 1999.
9. Upshur R, Morin B, Goel V. The privacy paradox: Laying Orwell's ghost to rest. *Canadian Medical Association Journal.* 2001;165(3):307-309.
10. HarrisInterractive. Health information privacy (HIPAA) notices have improved public's confidence that their medical information is being handled properly. http://www.harrisinteractive.com/news/allnewsbydate.asp?NewsID=894. Accessed 4th April, 2005.
11. California Health Care Foundation. *Medical privacy and confidentiality survey* 1999.
12. Grimes-Gruczka T, Gratzer C, The Institute for the Future. *Ethics survey of consumer attitudes about health web sites*: California Health Care Foundation; 2000.
13. Willison D, Kashavjee K, Nair K, Goldsmith C, Holbrook A. Patients' consent preferences for research uses of information in electronic medical records: Interview and survey data. *British Medical Journal.* 2003;326:373.
14. Mitchell E, Sullivan F. A descriptive feast but an evaluative famine: Systematic review of published articles on primary care computing during 1980-97. *British Medical Journal.* 2001;322:279-282.

15. Shalala D. Testimony before the US Senate Committee on Labor and Human Resources. http://aspe.hhs.gov/admnsimp/pvctest.htm. Accessed 4th April, 2005.

16. Goldman J, Hudson Z. Virtually exposed - Privacy and ehealth. *Health Affairs.* 2000;19(6):140-148.

17. Davis L, Domm J, Konikoff M, Miller R. Attitudes of first-year medical students toward the confidentiality of computerized patient records. *Journal of the American Medical Informatics Association.* 1999;6(1):53-60.

18. Mole D, Fox C. Electronic data protection: Procedures need drastic improvement. *British Medical Journal.* 5 March 2005;330(537).

19. Goldman J, Hudson Z, Smith R. *Report on the privacy policies and practices of health web sites*: California Health Care Foundation; 2000.

20. Glaser B, Strauss A. *The Discovery of Grounded Theory: Strategies for Qualitative Research*: Aldine; 1967.

About The Author

Dr. El Emam is an Associate Professor at the University of Ottawa, Faculty of Medicine and a Canada Research Chair in Electronic Health Information. Previously he was a senior research officer at the National Research Council of Canada, where he was the technical lead of the Software Quality Laboratory, and prior to that he was head of the Quantitative Methods Group at the Fraunhofer Institute for Experimental Software Engineering in Kaiserslautern, Germany. In both of these latter roles he was working on the development of predictive models of software quality, and on developing and evaluating audits of software processes to ensure good project outcomes. In 2003 and 2004, Khaled was ranked as the top systems and software engineering scholar worldwide by the *Journal of Systems and Software* based on his research on measurement and quality evaluation and improvement, and ranked second in 2002 and 2005. He holds a Ph.D. from the Department of Electrical and Electronics Engineering, King's College, at the University of London (UK).