De-identifying Health Data for Secondary Use: A Framework

22nd October 2008

Khaled El Emam
CHEO Research Institute

Document Information

Document Title:	De-identifying Health Data for Secondary Use: A Framework
Original Document Date:	22nd October 2008
Document Version:	Version 6
Copyright:	CHEO Research Institute, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada
Contact:	Khaled El Emam (kelemam@ehealthinformation.ca)
More Information:	http://www.ehealthinformation.ca/

Other Relevant Publications and Reports

K. El Emam: "Heuristics for de-identifying health data." In *IEEE Security and Privacy*, July/August, 6(4):58-61, 2008.

K. El Emam, and F. Dankar: "Protecting privacy using k-anonymity." In the *Journal of the American Medical Informatics Association*, September/October, 15:627-637, 2008.

K. El Emam, E. Neri, and E. Jonker: "An evaluation of personal health information remnants in second hand personal computer disk drives." In *Journal of Medical Internet Research*, 9(3):e24, 2007.

K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, M. Power: "Evaluating common de-identification heuristics for personal health information." In *Journal of Medical Internet Research*, 2006;8(4):e28, November 2006.

K. El Emam: "Overview of Factors Affecting the Risk of Re-Identification in Canada", Access to Information and Privacy, Health Canada, May 2006.

K. El Emam: "Data Anonymization Practices in Clinical Research: A Descriptive Study", Access to Information and Privacy, Health Canada, May 2006.

More information is available from
http://www.ehealthinformation.ca/

Table of Contents

1 INTRODUCTION 2

2 RISK EXPOSURE 4

2.1 DEFINITION 4
2.2 RISK EXPOSURE MODEL 5

3 MANAGING RISK EXPOSURE 7

3.1 USE CASES 7
3.2 MEASUREMENT 7
3.3 RISK MANAGEMENT PROCESS 8
3.4 THE TRADEOFFS 8

4 DISCUSSION 11

5 APPENDIX A 12

6 ACKNOWLEDGEMENTS 14

7 REFERENCES 15

1 Introduction

This document presents a framework for de-identifying health data for secondary use purposes. Although there are variations in the exact definition of secondary use, we will use one from the American Medical Informatics Association where secondary use is defined as any retrospective use of existing data that is not part of providing care to the patient, for example, analysis, research, safety and quality measurement and improvement, public health, payment, provider certification and accreditation, and marketing [1].

The decision to disclose health data for secondary use is often made by research ethics boards. They have the discretion to allow such disclosure and the conditions under which this can happen. For data custodians without ethics boards, such as some of the federal or provincial data repositories, there will be a privacy committee that makes such determinations.

Today, neither the health data custodians nor those seeking the data are satisfied with the status quo. Data custodians are concerned about patient privacy, and therefore they take a long time to decide whether and how to release health data about patients, as well as being conservative in their disclosure decisions. This means that getting access to health data for secondary use can take a long time, and may be refused altogether.

A commonly accepted solution to manage the privacy risks when disclosing health information is for the data custodian to de-identify the data beforehand. De-identification, if done well, balances the need to make sure that the privacy risks to the patients are managed, and the need of the data recipients for high quality data.

If health information is deemed to be de-identified then it is no longer personal information. The implication is that there is no legislative requirement to obtain consent from patients and the other stipulations of privacy laws would not apply. Many research ethics boards would also waive the consent requirement if the information was deemed to be de-identified [2].

However, there are degrees of de-identification. At the extremes are fully identifiable information and no information. But at which point do we pass from de-identified to identifiable information ? We present the current framework as a basis for answering that question in a practical way.

Our assumptions in this framework are:

- In the context of secondary use there is a data custodian and a data recipient. The data custodian has responsibility over the data, while the data recipient wishes to get access to the data.

- De-identification is within the context of a larger process that may include an ethics review and formal contracting arrangements. We will not address these issues directly in this document.
- It is often not practical for the data recipient to seek the consent of patients when data is being made available for secondary use. Therefore, de-identification is an acceptable approach for providing the data; it is acceptable to the patients and it is acceptable to the custodian.
- A data sharing agreement can be enforced and if there is a breach of the agreement there would be sanctions against the recipient. For example, the data custodian can seek damages from the data recipient when there is a breach of contract and/or ban the recipient from getting any data in the future.

This framework has already served as the basis for more detailed protocol and tool development, and is being applied in secondary use decision making contexts.

2 Risk Exposure

2.1 Definition

An important concept in this framework is that of re-identification *risk exposure*. Risk exposure is commonly defined as the probability of loss multiplied by the actual magnitude of the loss:

$$Risk\ Exposure = Loss \times Probability \quad \text{..... (1)}$$

There are two types of loss that one can consider:

- loss to the patients about whom the data pertains, and
- loss to the data recipient and custodian.

Loss to the patients can be gauged using the invasion-of-privacy test. A checklist that can be used to evaluate this is provided in Appendix A. The checklist covers things like the potential harm to the patients, the number of patients that would be affected by a data breach, and the severity of the injury.

Loss to the data recipient and custodian can most easily be captured in terms of the cost in dollars of dealing with a privacy breach. Costs include those for notification, loss of patients and customers, costs for any credit monitoring services if there is a possibility of financial fraud or identity theft, fines and penalties by regulators, any damages paid to the affected patients, and loss of goodwill. Who ends up actually paying will depend on the particular details of the breach incident, the legal status of the custodian, and any litigation that occurs.

The probability of loss refers to a patient's record being successfully re-identified in the data that is disclosed. It can be expressed as follows [3]:

$$\Pr(identified) = \Pr(identified \mid attempt) \Pr(attempt) \quad \text{..... (2)}$$

This denotes the probability of a patient being re-identified in a data set. Re-identification may be attempted by the data recipient his/her –self, a collaborator, employee, sub-contractor of the data recipient, or the data may be lost or stolen and re-identification is attempted by an unknown party. The probability of re-identification is measured using various re-identification metrics [4].

The probability of a re-identification attempt is a function of "Mitigating Controls" and the "Motives and Capacity" of the data recipient. Mitigating controls discourage the data recipient from attempting re-identification because they require security measures which reduce the chances of success, and because there are legal barriers to doing so that act as a disincentive. For example, if the data recipient is legally required to have a breach notification protocol in place, then a

successful re-identification attempt cannot be kept secret if it is discovered. This is a deterrent because of direct breach costs and the reputational damage which occurs with breach notification. Similarly, if there are many security measures in place, then it would be difficult for, say, other employees at the data recipient site to access the data, making an attempt at re-identification by them unlikely.

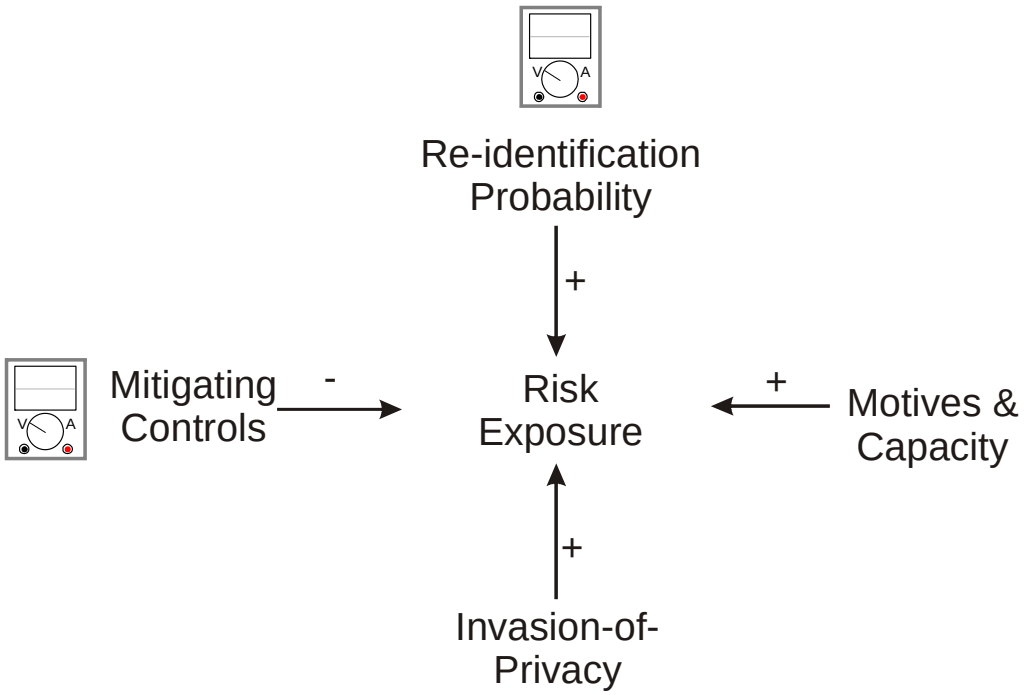

Figure 1: Conceptual view of the factors that influence overall re-identification risk exposure. The sign over the arrows indicates the direction of the relationship: positive or negative association.

2.2 Risk Exposure Model

The overall re-identification risk exposure of the data disclosure set is a function of four factors: (a) the re-identification probability, (b) the mitigating controls that are in place, (c) the motives and capacity of the data recipient to re-identify the data, and (d) the extent to which an inappropriate disclosure would be an invasion of privacy. The data custodian can manipulate (a) and (b) to manage the overall re-identification risk exposure. Factors (c) and (d) are intrinsic to the data recipient and the data set respectively and therefore would be very difficult to change. The four factors are illustrated in Figure 1.

The more motivated the data recipient is to re-identify the data, and to the extent that s/he has the capacity to do so, the greater the overall re-identification risk exposure. The greater the invasion-of-privacy (if there is an inappropriate disclosure), the greater the overall re-identification risk

exposure. Because these two factors are difficult to manipulate, they provide a baseline risk exposure for a particular data disclosure.

The higher the re-identification probability, the greater the overall re-identification risk exposure. The more mitigating controls that are put in place, the smaller the overall re-identification risk exposure. Because these two factors work in opposite directions, they can be manipulated by the data custodian to balance each other.

What is an acceptable risk exposure ? Thus far this remains somewhat subjective. However, we can define a *default* risk exposure which captures current practices today. To the extent that data custodians and patients are comfortable with the amount of risk being taken today, then we can use that as an anchor.

3 Managing Risk Exposure

3.1 Use Cases

An important implication of the concept of risk exposure is that risk exposure will depend on the specific data set and data recipient. For example, the disclosure of a data set X and data set Y to the same data recipient may have different risk exposures. There are two general use cases for applying these concepts:

1. In a case-by-case basis where every recipient requesting data goes through a risk assessment process and an appropriate risk exposure is determined. This means that potentially every data recipient will get a different data set (with a different level of de-identification).

2. Different classes of data recipients are identified in advance and the risk assessment is performed on the classes. When a new data recipient approaches the data custodian, the data recipient is classified, and the data set that is appropriate for that class is provided to the data recipient.

The first use case applies when it is difficult to predict which data sets will be in demand. For example, a hospital has a large number of possible data sets and may get a request for secondary use of any one or combination of them. Whereas a provincial data custodian, such as a prescribed registry in Ontario, will have one (or a small number of) data set(s) and may have common data requests that fall under the second use case above.

3.2 Measurement

The re-identification probability is managed by de-identifying the data. There are many techniques for de-identifying data sets [4-6]. For example, a commonly used rule is to ensure that there are no less than five records with the same values on the indirectly identifying variables in the data set. Therefore, to de-identify the data this rule can be implemented on the data set and any records which do not meet this criterion are removed from the data set.

The other three factors shown in Figure 1 are measured using a checklist. This checklist is provided in Appendix A. The items in the checklist are based directly on existing guidelines and recommendations [7-10]. The data recipient completes the “Mitigating Controls” section, and the data custodian completes the “Motives and Capacity” and “Invasion-of-Privacy” sections.

3.3 Risk Management Process

Within this framework, the data custodian and the data recipient must work together. This is because the process is iterative whereby the data custodian and data recipient must negotiate a suitable level of risk.

At the outset the data custodian decides whether the risk exposure is acceptable or not. If the risk exposure is not acceptable, then the data custodian can either: (a) de-identify the data further, and/or (b) put in place more mitigating controls.

The data recipient may perceive that the extent of de-identification is too extensive and has resulted in a large amount of information loss (for example, many records have been suppressed). In such a case the data recipient must agree to more mitigating controls if s/he wants less de-identification to be done to the data.

The data recipient may perceive that the amount of mitigating controls is too extensive and that s/he cannot afford to put them all in place. The data recipient therefore wants to reduce the amount of mitigating controls for a particular data release. Then the data custodian will have to increase the extent of de-identification to compensate for the reduction in mitigating controls.

Once the mitigating controls are agreed to, they are included in the data sharing agreement. In effect, the mitigating controls become binding obligations on the data recipient. This is the reason why the data sharing agreement must be enforceable; otherwise, there is little incentive to implement the mitigating controls.

Admittedly the process as described is qualitative and will depend on the expertise of the privacy analyst. But using the checklist in Appendix A, the negotiation between the data recipient and the custodian can be transparent, and more importantly, the tradeoffs become explicit.

3.4 The Tradeoffs

We will discuss the tradeoffs that need to be made with reference to Figure 2. We will assume that the potential harm to the patients is fixed. The re-identification probability is dichotomized into high and low, and so are the mitigating controls.

The de-identification of health data will result in some information loss. This means that the granularity of the variables will be reduced and some values will have to be suppressed. The figure shows that when the re-identification probability is reduced, the quality of the data goes down. Therefore, the data recipient does not want too much de-identification done to the data.

The more mitigating controls that are put in place the higher the overhead costs on the data recipient. When extensive mitigating controls are stipulated, individual researchers, journalists, and not-for-profit advocacy groups, for example, would not have the financial and technical

resources to implement them. Therefore, high mitigating controls ensure that only the very well funded data recipients can meet the requirements.

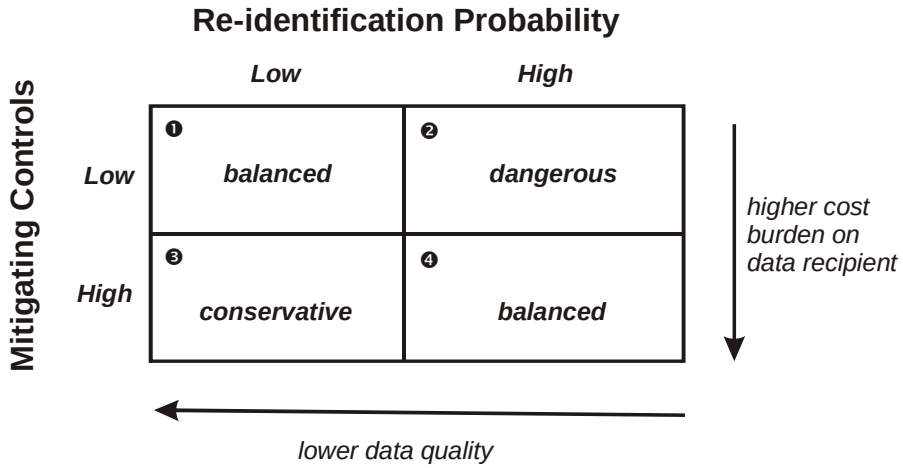

Figure 2: Conceptual diagram illustrating the tradeoffs that need to be made among the re-identification probablity and managing the mitigating controls.

Conceptually, a balance is attained when the mitigating controls are low and the re-identification probability is low (quadrant one). This quadrant is suitable for data recipients who do not have the resources to put into place strong mitigating controls, or if the data will be made publicly available where it is difficult to put in place strong mitigating controls. Note that extensive de-identification in quadrant one can go as far as not releasing any data (i.e., all records are suppressed). Another balance is attained if the data is only lightly de-identified but significant mitigating controls can be put in place (quadrant four).

Quadrant number two is dangerous in that the data has a high probability of re-identification and insufficient mitigating controls are in place. Some organizations may be in that quadrant because they do not realize that the probability of re-identification in their data is high, and/or because they mistakenly believe that they have sufficient controls in place (i.e., the risk management process is informal and not enforced).

The third quadrant is quite common in practice. This is where the data custodian takes a conservative approach: the re-identification probability is maintained very low *and* strong mitigating controls are put in place. This has the twin effect of disclosing data of lesser quality and imposing a high cost burden on the data recipient ensuring that only the well resourced can get access to this data. This approach attracts complaints from the full spectrum of potential data recipients: those that are well resourced complain that they do not get good quality data despite

the efforts in establishing strong mitigating controls, and those that are not well resourced because they cannot put all of the mitigating controls in place.

Following a conservative approach can be partially justified if all potential data custodians are funded to implement strong security and privacy systems and procedures. For example, if the government funds the implementation of secure facilities in all Canadian universities and provides the technical, legal, and analytical resources to operate them, then that ensure that practically all research data recipients can meet half the requirements (i.e., mitigating controls are high). However, the disclosed data will still be extensively de-identified if a conservative approach is followed. This means that even with such an expense, the research data recipients would not be satisfied with the outcome.

A conservative approach can be justified if data sharing agreements cannot be realistically enforced. For example, if the data custodian can require strong security and privacy practices, but has no way of enforcing these or has no legal or financial ability to seek damages in case of a breach of contract, then conservatism is a sensible choice.

Where data sharing agreements can be enforced, then we contend that the most appropriate quadrants are one and four. These provide reasonable tradeoffs among the competing demands.

4 Discussion

The risk management approach described in this framework provides a spectrum of solutions for secondary uses. The extent of de-identification to the data set is balanced against the potential harm done to patients and custodians. But all data recipients would get some data – except that the quality of the data is different.

In practice, this risk management approach has four consequences:

- It results in the data recipient and the data custodian working together to find a reasonable solution that is satisfactory to both. The data recipient understands the tradeoffs that need to be made and why. The data custodian can be transparent about how the decisions on de-identification are made and the need to manage the risk exposure.
- This approach provides a strong incentive for the data recipient to improve their privacy and security practices. To the extent that more of the mitigating controls can be put in place, the less de-identification will be performed on the data set.
- Data recipients with few funds and resources to put in place good records management practices, and procedures and systems for managing security and privacy, will likely still get some data. But this will be lower quality data because they would be considered high risk data recipients.
- Each data recipient needs to have a customized data sharing agreement that accounts for the specific mitigating controls required to manage the risk exposure for them. For example, one data custodian may have an annual audit provision in their agreement, while another would not. The former data recipient would get higher quality data with less de-identification compared to the latter.

The items in the checklist have some content and face validity, in that they are based on the checklists and questionnaires that are currently used in Canada to evaluate risk, and we have been using them and getting good outcomes. However, they are also currently undergoing further detailed empirical validation. The validation is attempting to address three questions: (a) do we have a complete set of factors, (b) within each factor do we have a complete list of items, and (c) what are the weights associated with each item.

5 Appendix A

This appendix contains the checklist that can be used to evaluate the subjective factors when computing risk exposure.

Mitigating Controls *(to be completed by the data recipient)*			
Item	**Yes**	**No**	**N/A**
The recipient has directly or indirectly worked/collaborated with the data custodian in the past			
The data sharing agreement forbids the recipient from disclosing the database to third parties			
The data sharing agreement is enforceable in all jurisdictions where the recipient will use the data			
The data sharing agreement will allow surprise audits of the recipient's record management system and practices			
The data sharing agreement stipulates that regular third party privacy and security audits need to be performed at the recipient site and of the recipient's practices			
The data sharing agreement imposes strong limits linking the database with other administrative or clinical data sources			
The recipient has a written privacy policy			
There is a person responsible for privacy at the recipient's site			
All members of the team on the recipient site have signed a confidentiality agreement as a condition of them having access to the disclosed database			
A threat and risk assessment has been completed on the recipient			
Strong security procedures for the collection, transmission, storage and disposal of personal information, and access to it, have been documented			
IT & database staff are sufficiently trained in the requirements for protecting personal information			
Systems are designed so that access and changes to personal information can be audited by date and user authentication			
User accounts, access rights and security authorizations are fully controlled by a system or record management process			
The recipient has an adequate breach notification protocol in place and their staff are trained in its implementation			
Computer systems are housed in a physically secure environment			
There is no public access to areas where computers holding the data will be			
The data will be destroyed once its purpose has been accomplished (e.g., the study has been published or other funding agency data retention period expires)			
Access rights are only provided to users on a 'need to know' basis consistent with the stated purpose for which the data was collected			

Motives and Capacity *(to be completed by the data custodian)*			
Item	**Yes**	**No**	**N/A**
The disclosed database has potential commercial or criminal value			
There is a likely non-commercial motive for the recipient to try to re-identify the disclosed database			
The recipient has the technical expertise to attempt to re-identify the disclosed database			
The recipient has the financial resources to attempt to re-identify the disclosed database			
The recipient may want to harm or embarrass the data custodian			
If the recipient does have a possible motive to attempt re-identification, they can achieve their objectives through other means apart from re-identification			
Invasion-of-Privacy *(to be completed by the data custodian)*			
Item	**Yes**	**No**	**N/A**
The personal information is highly detailed			
The database is large / many people would be affected if there was a breach			
The information is of a highly sensitive personal nature			
The information comes from a sensitive context (for example, data about individuals participating in a youth employment program are less sensitive than a similar list containing names and addresses of Hepatitis C and HIV compensation victims)			
The conditions that were established at the time the information was first collected from the individuals are consistent with the intended purpose of the recipient			
There is a commitment or promise not to disclose to any third party or institution			
There is a caveat stating that information can be disclosed in a manner consistent with the original purpose for its collection			
The information was compiled or obtained under guarantees that preclude some or all types of disclosure			
The information was unsolicited or given freely or voluntarily with little expectation of it being maintained in total confidence			
Disclosure of the information carries a probability of causing measurable injury (e.g., identity theft, fraud, etc)			
There is a risk in terms of the possible application of foreign laws			
The potential injury to the patients in case of an inappropriate disclosure is grave or serious			

6 Acknowledgements

The concepts presented here have benefited from discussions with Ed Brown (Memorial University). Funding for this work was provided by the Ontario Centers of Excellence.

7 References

1. Safran, C., M. Bloomrosen, E. Hammond, S. Labkoff, K.-F. S, P. Tang, and D. Detmer, *Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper.* Journal of the American Medical Informatics Association, 2007. 14: p. 1-9.
2. Willison, D., C. Emerson, K. Szala-Meneok, E. Gibson, L. Schwartz, and K. Weisbaum, *Access to medical records for research purposes: Varying perceptions across Research Ethics Boards.* Journal of Medical Ethics, 2008. 34: p. 308-314.
3. Marsh, C., C. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft, D. Lievesley, and N. Walford, *The case for samples of anonymized records from the 1991 census.* Journal of the Royal Statistical Society, Series A (Statistics in Society), 1991. 154(2): p. 305-340.
4. El Emam, K. and F. Dankar, *Protecting privacy using k-anonymity.* Journal of the American Medical Informatics Association, 2008. 15: p. 627-637.
5. El Emam, K., *Heuristics for de-identifying health data.* IEEE Security and Privacy, 2008: p. 72-75.
6. El Emam, K., S. Jabbouri, S. Sams, Y. Drouet, and M. Power, *Evaluating common de-identification heuristics for personal health information.* Journal of Medical Internet Research, 2006. 8(4): p. e28.
7. Elliot, M. and A. Dale, *Scenarios of attack: the data intruders perspective on statistical disclosure risk.* Netherlands Official Statistics, 1999. 14(Spring): p. 6-10.
8. *CIHR best practices for protecting privacy in health research*. 2005, CIHR.
9. Treasury Board of Canada Secretariat, *Guidance document: Taking privacy into account before making contracting decisions*. 2006, Government of Canada.
10. Treasury Board of Canada Secretariat, *Privacy impact assessment guidelines: A framework to manage privacy risks*. 2002, Government of Canada.