Dispelling the Myths Surrounding De-identification:

Anonymization Remains a Strong Tool for Protecting Privacy

Ann Cavoukian, Ph.D.
Information & Privacy Commissioner,
Ontario, Canada

Khaled El Emam, Ph.D.
Canada Research Chair in
Electronic Health Information,
CHEO Research Institute
and University of Ottawa

June 2011

Table of Contents

Introduction

Recently, the value of de-identification of personal information as a tool to protect privacy has come into question. Repeated claims have been made regarding the ease of re-identification. We consider this to be most unfortunate because it leaves the mistaken impression that there is no point in attempting to de-identify personal information, especially in cases where de-identified information would be sufficient for subsequent use, as in the case of health research.

The goal of this paper is to dispel this myth — the fear of re-identification is greatly overblown. As long as proper de-identification techniques, combined with re-identification risk measurement procedures, are used, de-identification remains a crucial tool in the protection of privacy. De-identification of personal data may be employed in a manner that simultaneously minimizes the risk of re-identification, while maintaining a high level of data quality. De-identification continues to be a valuable and effective mechanism for protecting personal information, and we urge its ongoing use.

In this paper we illustrate the importance of de-identifying personal information before it is used or disclosed, and at times, prior to its collection. We will demonstrate that, contrary to what has been suggested in recent articles, re-identification of properly de-identified information is not in fact an "easy" or "trivial" task. It requires concerted effort, on the part of skilled technicians. The paper will also describe a tool that minimizes the risk of the re-identification of de-identified information while also enabling a high level of data quality to be maintained. Our objective is to shatter the myth that de-identification is not a strong tool to protect privacy and to ensure that organizations that collect, use and disclose personal information understand the importance of de-identification for the protection of privacy, and continue to use this tool to the greatest extent possible to minimize potential risks. While our primary focus in this paper is on the value of de-identification in the context of personal health information that is used and disclosed for secondary purposes, the same arguments apply in the broader context of personal information.

Questioning the Value of De-identification

In recent years, studies have demonstrated that, in a surprising number of situations, it is possible to re-identify individuals from information that has previously been claimed to be de-identified. For example, a well publicized study demonstrated that it was possible to execute a successful re-identification attack on claims data on 135,000 patients disseminated by the Group Insurance Commission in Massachusetts. The discharge record for the then Governor was re-identified by matching it with simple demographic

information found in the Cambridge voter registration list which was purchased for $20. This was possible because certain fields in the two databases matched, namely: date of birth, 5-digit residential ZIP code, and gender. In another case, Netflix publicly released movie rating records, with each record containing the movie rated, the assigned rating and the date of the rating. Identifying information was removed, but each user had a unique identifier. After the records were released, researchers demonstrated that an attacker who knows a nontrivial amount about a target individual subscriber's movie viewing habits can potentially identify the subscriber's record if it is present in the Netflix data set.[1] For example, if someone knows approximately when a person in the database had rated six movies, he or she would be able to identify that person 99 per cent of the time. While these studies do not support the conclusion that it is pointless to de-identify data sets, they may create the mistaken impression that de-identification is not a worthwhile exercise to protect privacy.

These, and other studies, have given rise to recent academic articles[2] claiming that privacy cannot be protected through de-identification. These articles argue that the assumption that privacy may be protected through de-identification is incorrect, since it has been demonstrated that easy re-identification is possible.[3] Changes in society and technology have made re-identification of personal information easier and cheaper than ever before. New databases useful for linking are now available, with advances being made in re-identification technology. The Internet has facilitated the collection and distribution of vast amounts of information. Huge amounts of data are generated in the course of everyday life. While much of this information may appear as non-identifying, it can be combined with information from other sources to eventually produce data that may be linked back to specific individuals. People are also sharing personal information about themselves outside of traditional environments, including online chat rooms, personal blogs and social networking sites. As information technology initiatives enable the linkage of data across multiple sources, it becomes more difficult to ensure that de-identified information will remain anonymous.[4]

Some authors have argued that due to easy re-identification, laws such as the EU's *Data Protection Directive* ("*Directive*") are overly broad.[5] The *Directive* covers personal data that can be used to identify a person "directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity." Given the availability of sophisticated re-identification methods, it has been suggested that any database containing facts relating

1 Narayanan A, Shmatikov V (2008) *Robust De-Anonymization of Large Sparse Datasets*. Proc IEEE Security & Privacy Conference. pp.111-125.
2 See for example, Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, available at: http://uclalawreview.org/pdf/57-6-3.pdf.
3 Mark A. Rothstein, *Is Deidentification Sufficient to Protect Health Privacy in Research?*, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1679910.
4 Center for Democracy & Technology, *Encouraging the Use of, and Rethinking Protections for De-identified (and "Anonymized") Health Data,* available at: http://www.cdt.org/files/pdfs/20090625_deidentify.pdf.
5 *Supra*, note 2.

to people, no matter how indirect, would now fall within the *Directive* and be treated as personal information.

Another thread of reasoning has critiqued some existing de-identification standards as weak. For example, the United States *Health Insurance Portability and Accountability Act* ("*HIPAA*") excludes de-identified health information from regulation. The *HIPAA* Privacy Rule provides two standards for de-identification: (a) the Safe Harbor Standard, and (b) the Statistical Standard. The Safe Harbor Standard requires the removal of 18 specific data elements that could uniquely identify an individual. However, recent analysis shows that when used outside its narrowly defined context, the Safe Harbor Standard may not be sufficiently protective.[6] This is because these enumerated data elements in the Safe Harbor Standard do not include longitudinal data about patient visits, such as hospital name and diagnosis, as well as information that may be collected during the visit, such as profession. This additional information may be used by someone who has auxiliary information to re-identify an individual.

Consequently, one article[7] suggested that legislators should abandon the idea that privacy is protected when personal identifiers are removed and re-examine privacy laws and regulations. In another, the Markle Foundation, Center for Democracy & Technology and Markle Collaborators submitted comments in response to guidance published by the Department of Health and Human Services.[8] One of their recommendations was that the Department of Health and Human Services consider whether de-identified data should always be excluded from breach notification under *HIPAA* given the presumed risk of re-identification, and as part of its study of the *HIPAA* de-identification standard, consider imposing additional requirements on the disclosure of de-identified data. In their comments, they stated that de-identification, particularly through the removal of specific categories of identifiers (the Safe Harbor Standard), does not guarantee anonymity.

By questioning the value of de-identification as a tool to protect the privacy of individuals when personal information is collected, used and disclosed, these articles may lead to the mistaken impression that it is futile to de-identify any data sets. This growing lack of trust in de-identification and focus on re-identification risks may result in data custodians believing they should not waste their time even attempting to de-identify personal information prior to making it available for secondary purposes. Data custodians may also be less inclined to provide third parties with access to information, even if it has been de-identified. This could have a highly negative impact on the availability of de-identified information for potentially beneficial secondary purposes.

6 K. El Emam: "*Methods for the de-identification of electronic health records for genomic research.*" In Genome Medicine, 3:25, 2011.
7 *Supra*, note 2.
8 Markle Foundation, Centre for Democracy & Technology and Markle Collaborators, *Collaborative Comments on Federal Health Data Breach Notification Requirements*, available at: http://www.markle.org/publications/288-collaborative-comments-federal-health-data-breach-notification-requirements#FN4

The Enormous Value of De-identification

We believe that the problem, in part, lies in the pursuit of perfect solutions — "guarantees" of anonymity through the unfailing ability to de-identify personal data 100 per cent of the time. Needless to say, such guarantees do not exist, not in this pursuit, or virtually any other. But the absence of an iron clad "guarantee" does not stop people from attempting to minimize the risks encountered in various avenues of life. Every day, as part of their daily routine, people around the world implement security measures, even though there is no guarantee that such measures will always be 100 per cent effective. For example, before leaving the house, it is customary to lock one's door, in an effort to deter criminals from breaking in. There is no way to ensure that locking the door means that no one will ever break into your house — a skilled thief could still get past the locked door. However, for the most part, locking your door assists in keeping out unwanted intruders. As another example, people regularly put their money into bank accounts. There is always the possibility that a bank could be robbed or a safety deposit box burglarized, however, this is a minimal risk that people are willing to accept. It would be unthinkable to tell people not to lock their doors or put their prized possessions into banks because it is still *possible* that robberies could occur. Even though there are no guarantees that these security measures will work *all* of the time, they dramatically reduce the risk that a negative result will occur. De-identification of personal data is another example of an essential tool that should be routinely used to minimize risks, even though it may not work 100 per cent of the time. De-identification is an important first step that drastically reduces the risk that personal information will be used or disclosed for unauthorized or malicious purposes. While it is clearly not foolproof, it remains a valuable and important mechanism in protecting personal data, and must not be abandoned.

De-identification is particularly valuable in the context of personal health information. Health information is highly sensitive and may include some of the most intimate details associated with one's life, such as those related to one's physical and mental health. Personal health information requires the strongest privacy and security protections to prevent unauthorized collection, use and disclosure. However, under appropriate circumstances, it is also important to provide access to this information for vital secondary purposes that are strongly in the public interest. For example, personal health information is essential for public health surveillance and health-related research. It is also used for a variety of legally authorized purposes such as planning, delivering, evaluating and monitoring health programs or services, and improving the quality of care. The availability of information for such purposes results in enormous benefits for individuals and society at large by improving health-care programs and services and by improving the effectiveness of the health-care system. Health research can provide critical information about disease trends,

risk factors, outcomes of treatment, and patterns of care — it has led to significant discoveries including the development of new treatments and therapies.[9]

Frequently, de-identified information is used for secondary purposes, such as research and evaluation, where the recipient of the data has neither the motive nor the intention to re-identify the individuals contained in the data set. In fact, any re-identification would be counter-productive — thwarting the agreed upon terms typical of any data sharing agreement. In such situations, the use of de-identified information, rather than personal information, is of great value. De-identification protects individual privacy, while also enabling the information to be used for authorized secondary purposes, resulting in benefits to both individuals and society as a whole. We call this "positive-sum," not zero-sum. De-identification of personal health information can be done in a way that both minimizes the risk of re-identification and also maintains a high level of data quality. De-identification is an essential mechanism for protecting privacy, especially in circumstances such as sanctioned research where it is highly unlikely that the data recipients would ever attempt to re-identify the data set, since it would go against their best interests to do so.

The routine de-identification of information will also help to prevent privacy breaches in cases where the information is lost, stolen or accessed by unauthorized third parties.[10] If data custodians routinely de-identified information to the greatest extent possible, there would be far fewer data breaches. For example, if a USB key containing de-identified information is lost, it is unlikely that the person who finds the information would have the motive or capacity to attempt to re-identify the individuals in the data set — it is more likely that there would be no invasion of privacy. If, however, the information was not de-identified, there would be a much greater likelihood that the exposed personal information could be used for malicious purposes.

While de-identification may not be a perfect tool in preventing the disclosure of identifiable information, in all circumstances, de-identification will certainly prevent countless third parties, who in the vast majority of cases have neither the motivation, nor the technical expertise, nor the resources necessary to re-identify individuals, from knowing information about identifiable individuals. Further, it is important to not overlook the incidence of data breaches. A nontrivial percentage of these arise from "inside" jobs — by rogue employees who have easy access to identifiable data or accidentally by employees who do not follow or do not understand good data management practices. This could be reduced dramatically by default if far less personal data were retained in identifiable form — instead, being routinely retained with an appropriate amount of de-identification applied.

9 Institute of Medicine, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research*. Washington, DC, The National Academies Press, 2009.

10 For example, see Order HO-008 issued in June 2010, available at: http://www.ipc.on.ca/images/Findings/ho-008_1.pdf. See also Order HO-007 "*Encrypt Your Mobile Devices: Do It Now*" issued in January 2010, available at: http://www.ipc.on.ca/images/Findings/ho-007.pdf.

Challenges in Re-identifying De-identified Information

We believe it is highly misleading to suggest that the re-identification of individuals from de-identified data is an easy task. As long as proper de-identification and re-identification risk measurement techniques are employed, the re-identification of individuals is relatively difficult in actual practice. In fact, a recent review of the evidence indicates that there are few cases in which properly de-identified data have been successfully re-identified.[11] Further, in those cases where properly de-identified data were successfully re-identified, the re-identification risk was very low. The evidence is not consistent with the popular view relating to the fabled failure of anonymization.

For example, a recent study undertaken for the U.S. Department of Health and Human Services' Office of the National Coordinator for Health Information Technology ("ONC") sheds some light on the likelihood of a successful attack on properly de-identified data.[12] The ONC assembled a team of statistical experts to assess whether data properly de-identified under *HIPAA* could be combined with readily available outside data, to re-identify patients. The study was performed under realistic conditions and the re-identifications were verified to be accurate – something that other studies of this nature generally lack. The team began with a set of approximately 15,000 patient records that had been de-identified in accordance with *HIPAA*. Next, they sought to match the de-identified records with identifiable records in a commercial data repository. They conducted extensive searches through commercial data sources (e.g. InfoUSA) to determine whether any of the records in the identified commercial data would align with the records in the de-identified data set. The team was able to accurately re-identify only **two** of the 15,000 individuals, for a match rate of 0.013 per cent. This is an extremely low re-identification risk!

By the estimates reported in the August 23, 2007 testimony of one expert in de-identification, only **0.04 per cent** (4 in 10,000) of the individuals within data sets de-identified, using the Safe Harbor Standard under *HIPAA,* may be potentially re-identifiable.[13] This means that the use of de-identified data under *HIPAA* has been shown to reduce re-identification risks dramatically by at least 2,500 fold (less than 0.05%) over the identification risks that would have resulted from access to identifiable health information.

The ONC study and the risk estimate results further demonstrate that when applied within the appropriate context, the *HIPAA* Safe Harbor Standard can provide strong protections against re-identification. If the data does not meet the narrow assumptions

11 K. El Emam, E. Janker, B. Malin, "*A Systematic Review of Re-Identification Attacks on Health Data*." Submitted for publication. 2011.

12 Deborah Lafkey, *The Safe Harbor Method of De-Identification: An Empirical Test*, ONC Presentation, October 8, 2009, available at www.ehcca.com/presentations/*HIPAA*West4/lafky_2.pdf

13 *See* National Committee on Vital and Health Statistics Report to the Secretary of the U.S. Department of Health and Human Services, *Enhanced Protections for Uses of Health Data: A Stewardship Framework for 'Secondary Uses' of Electronically Collected and Transmitted Health Data*, December 19, 2007, *available at* http://www.ncvhs.hhs.gov/071221lt.pdf.

of the Safe Harbor Standard, then other standards and methods can and should be used, rather than abandoning de-identification altogether as some have suggested.

Professor Khaled El Emam has conducted research to evaluate the ease of re-identifying individuals in a large database of medical records that were de-identified using sophisticated de-identification tools. His study found that while it was possible to re-identify some individuals in the data set using public information, the number was minimal. Individuals who were re-identified were mostly adolescents. Information about these individuals in online social networking sites and school/college websites proved to be useful in the re-identification. However, it was a very small percentage of individuals that could actually be re-identified with any certainty (less than 0.5 per cent). Further, the re-identification of each individual was not an easy exercise, in some cases requiring many hours to complete! This supports the conclusion that the re-identification of individuals is a difficult and time-consuming task, on the part of skilled technicians.

The Implications of Including De-identified Information under Privacy Legislation

It has been suggested that in light of the risk of re-identification, it may be prudent to re-examine privacy laws with a view to including de-identified information among the types of information that must be protected. Applying legislative privacy protections to de-identified information may seem like the logical next step to the re-identification problem. However, we do not believe it would be an ideal solution since it may result in unintended consequences — there are other, more straightforward ways of managing the risk of re-identification. One unintended consequence of applying privacy laws to de-identified information may be that it would reduce the incentive to routinely de-identify personal data. Further, in those cases where personal information has been properly de-identified, and the re-identification risk is low, applying the additional conditions and restrictions imposed by privacy legislation would be unnecessarily burdensome, without the added benefit of enhancing privacy.

For example, if de-identified health information were included under Ontario's *Personal Health Information Protection Act* ("*PHIPA*"), this would mean that pursuant to section 29, health information custodians would only be able to collect, use or disclose de-identified information if the individual consents or if the collection, use or disclosure of personal health information is specifically permitted or required by *PHIPA*.

In some situations, obtaining consent would not be possible or practical. For example, in the context of non-interventional or database health research, the size of the population represented in the data may be too large to obtain consent from everyone, or many patients may have relocated or died, making it difficult or impossible to obtain consent.

Further, even if obtaining consent were possible and practical, strong concerns have been expressed about the negative impact of consent requirements on the ability to conduct certain types of health research.[14] In some cases, consent may have severe consequences for data quality since individuals who consent tend to have different characteristics than those who do not consent (e.g. age, gender, socioeconomic status, whether they live in rural or urban areas, disease severity, and level of education). These differences can result in significantly skewed results and biased findings. Further, it has been demonstrated that recruitment rates decline significantly when individuals are asked to consent. Seeking consent will likely result in a reduction of the number of individuals who agree to provide data or allow their data to be used. Consequently, where it is not practical to obtain consent, a case may be made for waiving the consent requirement.

If de-identified information were subject to *PHIPA*, in cases where it is not practical to obtain consent, health information custodians would only be permitted to collect, use or disclose personal health information without consent for the limited and specific purposes set out in *PHIPA* and all of its requirements would apply to the collection, use or disclosure. For example, in the case of the collection, use and disclose of de-identified information for research purposes, the requirements set out in section 44 of *PHIPA* would have to be satisfied, including the preparation of a detailed research plan, approval of the plan by a Research Ethics Board and, in the case of disclosure, the health information custodian would have to enter into an agreement with the researcher. Given the low risk that properly de-identified information will be re-identified by authorized researchers, the requirements and restrictions in *PHIPA* would appear to be unnecessarily burdensome and would do little to enhance privacy, over and above that which would be achieved through the proper de-identification of information.

It could be argued that Research Ethics Board approval of a research plan would be necessary even where the data that the researcher intends to use has been de-identified since, in practice, health information custodians tend to rely on Research Ethics Boards to determine whether data are sufficiently de-identified. However, in cases where a Research Ethics Board determines that the data are sufficiently de-identified, there would be no obligation to satisfy all of the requirements of section 44 of *PHIPA* and the Research Ethics Board would be free to adjust its review process to be compatible with the lower degree of risk associated with de-identified data.

As another example, section 52 of *PHIPA* gives an individual the right of access to a record of personal health information about the individual that is in the custody or control of a health information custodian, subject to certain exceptions. If de-identified information were treated in the same manner as personal health information, then health information custodians would have to re-identify information every single time that

14 Khaled El Emam, Elizabeth Jonker and Anita Fineberg, *The Case for De-identifying Personal Health Information*, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1744038.

an individual wished to access his or her records, creating an undue burden on health information custodians. Once a data set has been de-identified, it would be difficult to determine if the data set contained information about a specific individual. Whole data sets may have to be re-identified in order to provide individuals with access to de-identified health information, even though information about that individual may not even be contained within the data set!

As well, in the event that personal health information is stolen, lost or accessed by unauthorized persons, section 12(2) of *PHIPA* requires health information custodians to notify individuals at the first reasonable opportunity. If de-identified information is not excluded from the application of *PHIPA*, then these notification requirements would apply to de-identified information as well as personal health information, regardless of the fact that there may be a very low probability that an individual would ever be identified. Further, ironically, the de-identified information may have to first be re-identified in order to determine which individuals to notify about the breach. Finally, notification of a breach involving de-identified information may cause undue alarm or be perceived as a nuisance where the risk of re-identification is relatively low. These days, individuals may already be bombarded with too many breach notices under U.S. privacy laws, causing them to ignore such notices.[15] Expanding the scope of breach notification requirements to include de-identified information would only compound this problem.

If de-identified information were treated in the same manner as personal health information under *PHIPA*, there would be far less incentive for health information custodians to de-identify any information, as the same restrictions and requirements would apply regardless of the identifiability of the data. This may result in a decrease in the use of one of the most effective tools for the protection of privacy in the context of the collection, use and disclosure of personal information for secondary purposes. Further, it may drastically reduce the availability of de-identified information for potentially beneficial purposes.

De-identification in the Context of Privacy Legislation

Privacy legislation generally permits the collection, use and disclosure of personal health information for secondary purposes such as research, under appropriate circumstances, and treats information that does not relate to an identifiable individual as falling outside the scope of the legislation. For example, in Ontario, *PHIPA* permits the collection, use and disclosure of personal health information for secondary purposes, including health research, in specific circumstances. Where the collection, use or disclosure is not specifically permitted by *PHIPA*, health information custodians must either obtain

15 Fred H. Cate, *Another notice isn't answer*, USA Today, February 27, 2005, available at http://www.usatoday.com/news/opinion/2005-02-27-consumer-protection-oppose_x.htm

consent from individuals or de-identify the health information in a manner such that it falls outside the scope of *PHIPA*.

The general limiting principles of *PHIPA* place an obligation on health information custodians to de-identify personal health information to the greatest extent possible whenever they collect, use or disclose personal health information. Section 30(1) of *PHIPA* states that health information custodians must not collect, use or disclose personal health information if other information, e.g. de-identified information or aggregate information, would serve the purpose of the collection, use or disclosure. Section 30(2) of *PHIPA* states that health information custodians must not collect, use or disclose more personal health information than is reasonably necessary to meet the purpose of the collection, use or disclosure. This means that health information custodians must collect, use and disclose de-identified health information rather than personal health data if the de-identified information would be sufficient to serve the purpose. These principles apply whether or not individuals have consented to the collection, use and disclosure of their health information. As well, section 37(1)(f) of *PHIPA* specifically states that health information custodians may use personal health information about an individual for the purpose of disposing of or modifying the information in order to conceal the identity of the individual. Therefore, health information custodians not only have an obligation to de-identify personal health information, to the greatest extent possible, but they also have the legal authority to use personal health information for the purpose of de-identification. Once de-identified, in a manner such that it falls outside the scope of *PHIPA*, the information may then be used and disclosed for secondary purposes, without the consent of the individual.

As a starting point, de-identification must be considered prior to collecting, using and disclosing personal health information. It has been argued that personal health information should always be de-identified before it is collected, used or disclosed for secondary purposes, since de-identified data is sufficient for such purposes. However, while this may generally be the case, there are some situations where it will be necessary to collect, use or disclose personal health information for secondary purposes. For example, in the context of research, especially epidemiological research, personal health information as opposed to de-identified information may be required in order to link and match data from various sources, over time. However, once the data are linked, the information can then be immediately de-identified as the personal health information is no longer necessary for the purpose.

Section 4(1) of *PHIPA* defines personal health information as identifying information about an individual in oral or recorded form if the information relates to the physical or mental health of the individual; the provision of health care to the individual; payments or eligibility for health care or health-care coverage; the donation of any body part or bodily substance of the individual; is a plan of service within the meaning of the *Long-Term Care Act*; is the individual's health number; or identifies an individual's substitute

decision-maker. Section 4(2) of *PHIPA* defines identifying information as information that identifies an individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify an individual. Health information that is de-identified such that an individual cannot be re-identified, or that it is not reasonably foreseeable that an individual can be re-identified, would fall outside the scope of *PHIPA*. This de-identified health data would no longer be considered personal health information and would not be subject to any of the limitations and restrictions imposed by *PHIPA*.

To reduce the re-identification risk to the level where re-identification is not reasonably foreseeable in the circumstances, health information custodians may alter and/or remove identifiers prior to using or disclosing personal health information. Direct identifiers (or identifying variables) are variables that provide an explicit link to a data subject and can directly identify an individual. Examples of identifying variables include name, email address, home address, telephone number, health insurance number and social insurance number. [16]

It is frequently assumed that by removing the direct identifiers, the privacy of the individuals whose information is being used or disclosed would be protected.[17] However, dealing only with direct identifiers is insufficient to ensure that the information is truly de-identified. The problem of de-identification involves quasi-identifiers, those variables that may not directly identify individuals, but can still be used for indirect re-identification. These quasi-identifiers can be used, either by themselves or in combination with other available information, to uniquely identify individuals. Examples of quasi-identifiers include gender, marital status, postal code or other location information, a significant date (e.g. birth, death, hospital admission, discharge, autopsy, specimen collection, or visit), diagnosis information, profession, ethnic origin, visible minority status, and income.

Due to their rarity, some quasi-identifiers may be more likely to result in identification of individuals in a data set. For example, highly uncommon characteristics of an individual (e.g. an unusual occupation or an unusual medical diagnosis) can increase the likelihood of the individual's identity being revealed. As well, a particular quasi-identifier does not necessarily always have the same likelihood of re-identification. For example, the postal codes of rural areas contain many more individuals than urban areas. It is more likely that individuals in a data set with an urban postal code could be re-identified than those in a rural area. Quasi-identifiers may also differ across data sets. For example, gender will not be a meaningful quasi-identifier if all of the individuals in the data set are female.[18] However, if all the individuals in a data set are female except one, there

16 Khaled El Emam and Anita Fineberg, *An Overview of Techniques for De-identifying Personal Health Information*, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1456490.
17 Privacy Analytics Inc., *De-identification Reduce Privacy Risks when Sharing Personally Identifiable Information*, available at: http://privacyanalytics.ca/documents/de-identification_explained.pdf.
18 Khaled El Emam et al., *Pan-Canadian De-identification Guidelines for Personal Health Information*, available at: http://www.ehealthinformation.ca/documents/OPCReportv11.pdf.

is a higher likelihood that the one male in the data set could be re-identified. Data sets should be analyzed to asses the risk of re-identification and determine the proper de-identification techniques that should be employed.

Solving the Zero-Sum Paradigm

De-identification is one of the most important methods of protecting privacy and should not be casually abandoned. It should be designed directly into all processes in which personal health information is collected, used and disclosed for secondary purposes, a concept advanced by *Privacy by Design*.[19] Taking a *Privacy by Design* approach is characterized by proactive rather than reactive measures in an effort to anticipate and prevent the privacy harm from ever arising. Privacy is embedded into the design of systems and business practices as the default setting, ensuring that personal health information is automatically protected. *Privacy by Design* is a comprehensive approach that extends securely through the entire life cycle of the information involved. Visibility and transparency, as well as respect for user privacy, are paramount.

An issue in respect of de-identification is the traditional zero-sum paradigm. To reduce the risk of re-identification, direct identifiers and quasi-identifiers may be altered and/or removed prior to the collection, use or disclosure of personal health information, for secondary purposes. However, as more variables are altered or removed, the quality of the information is reduced. Individual privacy is achieved at the expense of data quality. Conversely, limiting the variables that are de-identified can result in higher data quality, but at the expense of individual privacy. For example, one can aggressively de-identify all the information in a data set, but then the information is no longer useful to researchers. Alternatively, one can choose not to de-identify any of the personal health information in the data set; however, this results in significant privacy concerns. Utility and privacy of data are intrinsically connected. As the usefulness of the data increases, privacy has tended to decrease.

However, through the use of proper de-identification techniques and re-identification risk measurement techniques, it is possible to achieve a high degree of privacy, while at the same time preserving the required level of data quality necessary for the secondary purpose. Maximizing both privacy and data quality enables a shift from a zero-sum paradigm to a positive-sum paradigm, a key principle of *Privacy by Design*. This doubly-enabling win-win approach avoids unnecessary trade-offs and illustrates that it is possible to de-identify information in a manner that maintains both privacy and data quality.

19 Ann Cavoukian, *Privacy by Design: The 7 Foundational Principles*, available at: http://www.ipc.on.ca/images/resources/7foundationalprinciples.pdf.

Re-identification Risk Assessment

To reduce the risk of re-identification, it is essential to properly de-identify the information contained in a data set. There are many tools available that mitigate the risk of re-identification of de-identified information. Some automated de-identification programs remove or suppress direct identifiers in data sets, while others use sophisticated algorithms in order to address the risks of re-identification in data sets from which direct identifiers have already been removed. Professor Khaled El Emam has developed a tool that provides a high degree of privacy protection while also ensuring a high level of data quality, providing an excellent example of a positive-sum approach to de-identification.

Privacy analytics involves analyzing the data set itself in order to measure the re-identification risk and then deciding how to best de-identify the data.[20] The de-identification process is risk-based and takes into consideration the usefulness of the data. The analyst must be able to evaluate the risk of re-identification. Once this risk is known, decisions may then be made about how, and the extent to which the data should be de-identified. The first step is to determine the quasi-identifiers. The next step is to determine the acceptable level of risk. The risk threshold should reflect the amount of re-identification risk that the health information custodian is willing to take. For example, one approach is to ensure that for each record contained in the data set that describes characteristics of a data subject, there are at least four other individuals also represented by records in the data set who share these same characteristics.[21]

The risk of re-identification should be evaluated. Criteria that can be used to determine the re-identification risk exposure of a data disclosure is a function of four factors: the re-identification probability; the mitigating controls that are in place; the motives and capacity of the data recipient to re-identify the data; and the extent to which an inappropriate disclosure would be an invasion of privacy.[22] The re-identification probability is the probability of an individual being re-identified in a data set and it is managed by de-identifying the data set. Mitigating controls discourage the data recipient from re-identifying the data. A data sharing agreement between the health information custodian and the data recipient can document the expectations and obligations of each party. For example, mitigating controls such as prohibiting re-identification of the data, limiting use and disclosure of the data, requiring strong security protections and a breach notification protocol, and permitting the health information custodian to audit compliance, can be included in the data sharing agreement.

20 Health System Use Technical Advisory Committee Data De-identification Working Group, *'Best Practice' Guidelines for Managing the Disclosure of De-identified Health Information*, available at: http://www.ehealthinformation.ca/documents/Data%20De-identification%20Best%20Practice%20Guidelines.pdf.

21 Khaled El Emam and Fida Kamal Dankar, *Protecting Privacy Using k-Anonymity*, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/.

22 Khaled El Emam, *De-identifying Health Data for Secondary Use: A Framework*, available at: http://www.ehealthinformation.ca/documents/SecondaryUseFW.pdf.

To assess motives and capacity of the data recipient to re-identify the data involves evaluating criteria such as whether the data recipient may want to harm the health information custodian, whether the data recipient has the financial resources or technical expertise to attempt to re-identify the database and whether the database has commercial or criminal value. To assess the extent to which an inappropriate disclosure would be an invasion of privacy involves evaluating criteria such as whether the information in the database is highly sensitive, whether the information is highly detailed and whether the database is very large and many people would be affected if there was a breach.

The motives and capacity of the data recipient and the extent to which an inappropriate disclosure would be an invasion of privacy are inherent to the data recipient and would be difficult to change. The greater the potential invasion of privacy and the more motivated and capable the data recipient is to re-identify the data, the greater the overall risk of exposure. However, the re-identification probability and the mitigating controls can be changed by the health information custodian in order to manage overall risk exposure.[23]

Once a data request has been received, the health information custodian can determine if the overall risk exposure is acceptable. If the risk exposure is not acceptable, the health information custodian can add more mitigating controls or further de-identify the data set. The health information custodian and the data recipient must work together to achieve the level of data quality that is necessary for the data recipient's purposes and the level or risk exposure that is acceptable to the health information custodian.[24] The data recipient may prefer less de-identification (to preserve greater data quality) and so must agree to more mitigating controls in the data sharing agreement. Conversely, the data recipient may not be able to afford to put in place a large number of mitigating controls and must then agree to increase the amount of de-identification to the data. Balance occurs when the mitigating controls are low and the re-identification probability is low (e.g. there is more de-identification of the data but less mitigating controls) or where the re-identification probability is high and the mitigating controls are high (e.g. there is less de-identification but more mitigating controls). Each data recipient must have a customized data sharing agreement that accounts for the specific mitigating controls required to manage risk exposure.

This approach takes into consideration the fact that de-identification of information is not a perfect tool, as there may always be some risk of re-identification. However, the re-identification risk can be assessed and managed through a variety of means.

23 *Ibid*.
24 Ann Cavoukian and Khaled El Emam, *A Positive-Sum Paradigm in Action in the Health Sector*, available at: http://www.privacyanalytics.ca/documents/positive-sum.pdf.

Conclusion

The claim that the de-identification of personal data has no value and does not protect privacy due to the ease of re-identification is a myth. If proper de-identification techniques and re-identification risk measurement procedures are used, re-identification remains a relatively difficult task. However, we recognize that this is not a static exercise – it is ever-changing. As re-identification techniques become more sophisticated and more personal information becomes available to facilitate re-identification, it is important to reassess and strengthen de-identification and re-identification risk management techniques.

While there may always be a residual risk of re-identification, in the vast majority of cases, de-identification will protect the privacy of individuals, as long as additional safeguards are in place. While de-identification may not be a perfect solution to reduce all privacy risks when personal information is being considered for secondary purposes, it is an important first step that should be used as part of an overall risk assessment framework. We urge you not to abandon your efforts to de-identify personal data in a comprehensive and responsible manner.

About the Authors

Dr. Ann Cavoukian

Information and Privacy Commissioner, Ontario, Canada

Dr. Ann Cavoukian is recognized as one of the leading privacy experts in the world. Her concept of *Privacy by Design* seeks to proactively embed privacy into the design specifications of information technology and accountable business practices, thereby achieving the strongest protection possible. In October, 2010, data regulators from around the world unanimously passed a landmark Resolution recognizing *Privacy by Design* as an essential component of fundamental privacy protection. This was followed by the U.S. Federal Trade Commission's inclusion of *Privacy by Design* as one of its three recommended practices for protecting online privacy – a major validation of its significance.

Dr. Cavoukian's leadership has seen her office develop a number of tools and procedures to ensure that privacy is strongly protected, not only in Canada, but around the world. She has been involved in numerous international committees focused on privacy, security, technology and business, and endeavours to focus on strengthening consumer confidence and trust in emerging technology applications.

Dr. Cavoukian serves as the Chair of the Identity, Privacy and Security Institute at the University of Toronto, Canada. She is also a member of several Boards including, the European Biometrics Forum, Future of Privacy Forum, RIM Council, and has been conferred as a Distinguished Fellow of the Ponemon Institute. She was named by *Intelligent Utility Magazine* as one of the "Top 11 Movers and Shakers for the Global Smart Grid industry for 2011," and has been honoured with the prestigious *Kristian Beckman Award* for her pioneering work on *Privacy by Design* and privacy protection in modern international environments.

Dr. Khaled El Emam

Canada Research Chair in Electronic Health Information, CHEO Research Institute and University of Ottawa

Dr. Khaled El Emam is an Associate Professor at the University of Ottawa, Faculty of Medicine, a senior investigator at the Children's Hospital of Eastern Ontario Research Institute, and a Canada Research Chair in Electronic Health Information at the University of Ottawa. His main area of research is developing techniques for health data anonymization and secure disease surveillance for public health purposes. Previously Khaled was a Senior Research Officer at the National Research Council of Canada, and prior to that he was head of the Quantitative Methods Group at the Fraunhofer Institute in Kaiserslautern, Germany. He has co-founded two companies to commercialize the results of his research work. In 2003 and 2004, he 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, King's College, at the University of London (UK).

Information & Privacy Commissioner of Ontario
2 Bloor Street East, Suite 1400
Toronto, Ontario M4W 1A8
CANADA

Telephone: (416) 326-3333
Toll-free: 1-800-387-0073
Fax: (416) 325-9195
TTY (Teletypewriter): 416-7539
Website: www.ipc.on.ca
E-mail: info@ipc.on.ca