De-identification Protocols: Essential for Protecting Privacy

June 25, 2014

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

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

Information and Privacy
Commissioner of Ontario

Commissaire à l'information et à la
protection de la vie privée de l'Ontario

2 Bloor Street East
Suite 1400
Toronto, Ontario
Canada
M4W 1A8

416-326-3333
1-800-387-0073
Fax: 416-325-9195
TTY (Teletypewriter): 416-325-7539
Website: www.ipc.on.ca

Table of Contents

De-identification Protocols: Absolutely Essential for Protecting Privacy

Information is the new currency of our economy. Since the dawn of the digital era, information has become increasingly available, and at a scale previously unimaginable. According to IBM, each day, 2.5 quintillion bytes of information are being created and, over 90 percent of the information currently in existence has been created in the past two years.[1] With technological advances, this information is also becoming easier to collect, retain, use, disclose and leverage for a wide range of secondary uses.

Information is becoming far more valuable as businesses, big and small, seek to learn more about their customers and those of their competitors, and as advertisers seek to gain a competitive advantage by finding new and innovative ways to use information to target advertisements that are most relevant to their consumers. Information is also increasingly being sought for secondary uses that are seen to be in the public interest. For example, the health sector is seeking to use information to support evidence-based decision-making, to improve the quality of care provided, and to identify and achieve cost efficiencies.

However, if organizations do not strongly protect the privacy of individuals in the information being sought out, there may be far-reaching implications for both the individuals and the organizations involved. For example, when individuals lose trust and confidence in the ability of an organization to protect their privacy, the reputation of that organization may be irreparably damaged in the process. This does not include the time and resources needed to contain, investigate and remediate any resulting data breaches or privacy infractions, and the costs associated with any ensuing proceedings and liabilities such as class action lawsuits.

One of the most effective ways to protect the privacy of individuals is through strong de-identification. Despite suggestions to the contrary, de-identification, using proper de-identification techniques and re-identification risk management procedures, remains one of the strongest and most important tools in protecting privacy.

1 Kevin C. Desouza & Kendra L. Smith, "Big Data for Social Innovation" (2014) Stanford Social Innovation Review 39.

Background

In 2011, the Information and Privacy Commissioner of Ontario, Canada, published a joint paper with Dr. Khaled El Emam in response to then current reports questioning the value of de-identification. *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy*[2] argued that proper de-identification techniques, including a robust re-identification risk management framework, are vital to the protection of privacy and that the re-identification of properly de-identified information was far more difficult than suggested by some commentators. Re-identification of properly de-identified information requires significant time and financial resources, in addition to the concerted efforts of those with a specialized skill set.

In 2013, the Information and Privacy Commissioner of Ontario, Canada, published another paper, entitled *Looking Forward: De-identification Developments – New Tools, New Challenges*,[3] which provided further evidence that the re-identification of properly de-identified information was not an insignificant endeavour. It also highlighted key developments in the area of de-identification, including advances in de-identification techniques and strategies for enhancing trust through the implementation of risk management policies, procedures and practices, along with the execution of formal agreements.[4]

Despite these papers, there continues to be considerable misunderstanding surrounding the area of de-identification and the risk of re-identification. Recent reports, including those emanating out of John Podesta's Big Data and Privacy Workshops,[5] have further fuelled this misunderstanding. These reports have caused some commentators to call into question the usefulness of de-identification, to suggest that information can typically only be de-identified at the expense of data quality, and to emphasize the relative ease of re-identification. We again submit that these views are an over-simplification, inconsistent with current evidence, and largely based on the re-identification of poorly de-identified information.

The purpose of this paper is to clarify what it means to properly de-identify personal information, to underscore the value of strong de-identification, to interpret recent research which has been used to call into question the value of de-identification in the protection of privacy, and to emphasize the conclusions that may properly be drawn from this research. We will argue that the vast majority of information may be de-identified in a manner that both provides a high degree of privacy protection, *and* ensures a level of data quality that is suited for the secondary purpose.

2 Ann Cavoukian & Khaled El Emam, "Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy" (2011), online: http://www.ipc.on.ca/images/Resources/anonymization.pdf.

3 Information and Privacy Commissioner of Ontario, Canada "Looking Forward: De-identification Developments – New Tools, New Challenges (2013), online: http://www.ipc.on.ca/images/Resources/pbd-de-identifcation_developments.pdf.

4 *Ibid*.

5 Executive Office of the President, Big Data: Seizing Opportunities, Preserving Values (Washington: The White House, 2014) and Executive Office of the President, President's Council of Advisors on Science and Technology, *Report to the President Big Data and Privacy: A Technological Perspective* (Washington: The White House, 2014).

The What, Why and How of De-identification

De-identified information is information that cannot be used to identify an individual, either directly or indirectly. Information is de-identified if it does not identify an individual, and it is not reasonably foreseeable in the circumstances that the information could be used, either alone or with other information, to identify an individual.[6]

Personal information is de-identified through a process involving the removal or modification of both direct identifiers **and** indirect or quasi-identifiers, unlike "masking" which only involves the removal or modification of direct identifiers.

Direct identifiers are fields of information that may be used to directly identify an individual; they include name, home address, telephone number, health number and social insurance/ security number. Indirect or quasi-identifiers are fields of information that may be used on their own or in combination with other indirect or quasi-identifiers, or other information, to indirectly identify an individual. They include information such as gender, marital status, race, ethnic origin, postal code or other location information, significant dates, or one's profession. Some indirect or quasi-identifiers may be more likely to lead to the re-identification of individuals in a dataset due to their rare occurrence. Characteristics which are highly uncommon in the population or in the dataset, such as an unusual occupation or medical diagnosis, can increase the likelihood of the identity of an individual being revealed.

In addition to the removal or modification of direct and indirect or quasi-identifiers, de-identification often involves the implementation of a robust re-identification risk management framework. An assessment must be conducted to identify the risks of re-identification in the particular circumstances involved, having regard to such factors as the motives and capacity of the organization or individual to re-identify the information. Once these risks have been determined, a re-identification risk management framework must be implemented to mitigate the risks identified, including the implementation of privacy and security policies, procedures and practices and the execution of agreements.[7]

At a minimum, policies, procedures and practices should be implemented, including those in respect of privacy breach management and de-identification. The de-identification policies, procedures and practices should be based on the assessment of the risks identified, and be reviewed on a regular basis to ensure that they continue to be consistent with industry standards and best practices, technological advancements, legislative requirements, and emerging risks. Further, agreements should be entered into with the organization or individual using or receiving the de-identified information. These agreements should: 1) prohibit the use of de-identified information, either alone or with other information, to identify an individual; 2) place restrictions on any other use or subsequent disclosure of the de-identified information; 3) ensure that those who have access to the de-identified information are properly trained and understand

6 *Personal Health Information Protection Act, 2004*, SO 2004, c 3, s 4(2); *Personal Health Information Act*, SNS 2010, c 41, s 3; and *Personal Health Information Act*, SNL 2008, c P-7.01, s 5(5).

7 Khaled El Emam, *Guide to the De-Identification of Personal Health Information* (Boca Raton: CRC Press, 2013), Khaled El Emam, *Risky Business: Sharing Health Data While Protecting Privacy* (Bloomington: Trafford Publishing, 2013) and Khaled El Emam & Luk Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started* (Sebastopol: O'Reilly Media, 2013).

their obligations in respect of such information; 4) require the recipient to notify the organization of any breach of the agreement, and 5) set out the consequences of such a breach.[8]

Provided that proper de-identification techniques, in conjunction with re-identification risk management procedures are used, the collection, use and disclosure of de-identified information has a number of advantages over the use of personally identifiable information:

- Proper de-identification greatly reduces the risk of a privacy breach in the event that the information is lost, stolen or accessed by unauthorized persons, since it is far less likely that individuals can be identified from information that has been properly de-identified;
- De-identification allows organizations to comply with data minimization principles[9] – principles that are the cornerstone of privacy legislation and fair information practices, all across the globe;[10] and
- Greater use may be made of de-identified information since properly de-identified information falls outside the scope of privacy legislation and is not subject to the same restrictions and limitations that are imposed on the collection, use and disclosure of personally identifiable information.

8 *Supra*, note 3 at 7 - 8.

9 Data minimization requires organizations not to collect, use or disclose personal information if other information would serve the same purpose and not to collect, use or disclose more personal information than is necessary to meet the purpose.

10 See for example Organisation for Economic Co-operation and Development, *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data* (Paris: OECD, 2002); *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5, Schedule 1, 4.4 and 4.5; and *Personal Health Information Protection Act, 2004*, SO 2004, c 3, s 30.

The Research on Re-identification – What's New?

While there is always a residual risk of re-identification, the ease of re-identification should not be overstated. Recent studies used by some commentators to highlight the ease of re-identification, have certain features that limit the conclusions that may be drawn, as well as the generalizability of their findings. First, many of these studies were based on personal information that was not properly de-identified.[11] Second, while one of the research studies demonstrated that an individual may be unique within a dataset, no individuals were actually re-identified. In fact, no attempts were made by the researchers to re-identify any individuals.[12] Third, any conclusions that may be drawn from these studies must take into consideration the significant time, resources and effort of the highly skilled experts required to re-identify individuals, based on properly de-identified information. Fourth, the generalizability of the findings is limited by jurisdictional considerations such as the privacy legislation in force and the information publicly available within the jurisdiction. Finally, a number of these studies involved the re-identification of genetic information, for which we readily concede, there is currently no effective method of de-identification. Each of these will be discussed in detail below.

It is Easier to Re-identify Improperly De-identified Information – No Kidding!

Over the years, a number of studies have successfully re-identified individuals from information that was purportedly de-identified. For example:

- In 1997, the medical records of the then governor of Massachusetts were re-identified by matching information made publicly available by the Group Insurance Commission with demographic information found in a voter registration list purchased for 20 dollars;[13]
- In 2006, reporters for the New York Times were able to identify a single individual in a list of web search queries released by AOL, using the searches that the individual had made over a three-month period;[14] and
- In 2008, researchers were able to re-identify the movie rating records of subscribers in a dataset publicly released by Netflix by comparing those records to the movie ratings posted on the Internet Movie Database, knowing very little information about the subscribers.[15]

11 Latanya Sweeney, "Matching Known Patients to Health Records in Washington State Data" (2013), online: http://arxiv.org/pdf/1307.1370v2 and Latanya Sweeney, Akua Abu & Julia Winn "Identifying Participants in the Personal Genome Project by Name" (2013), online: http://dataprivacylab.org/projects/pgp/1021-1.pdf

12 Yves-Alexandre de Montjoye et al., "Unique in the Crowd: The Privacy Bounds of Human Mobility" (2013) 3:1376 Sci. Rep. 1.

13 Latanya Sweeney, "k-Anonymity: A Model for Protecting Privacy" (2002) 10:5 International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 557.

14 Michael Barbaro & Tom Zeller Jr., "A Face Is Exposed for AOL Searcher No. 4417749," *New York Times* (9 August 2006), online: http://select.nytimes.com/gst/abstract.html?res=F10612FC345B0C7A8CDDA10894DE404482.

15 Arvind Narayanan & Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets" (2008), online: http://arxiv.org/pdf/cs/0610105.pdf.

These, and other studies, gave rise to a number of academic articles arguing that privacy could not be protected through de-identification, given the "astonishing ease" of re-identification.[16]

However, in 2011, a systematic literature review was conducted to identify and assess published accounts of re-identification attacks on de-identified datasets, including many of the re-identification attacks listed above.[17] The literature review identified 14 accounts of re-identification attacks on ostensibly de-identified information, **but found that only two of the 14 attacks were made on records that had been properly de-identified. The remaining 12 attacks were made on records that had not been properly de-identified.**[18]

For example, the claims database of the Group Insurance Commission that had been used to re-identify the governor of Massachusetts contained gender, date of birth and full five-digit ZIP codes, therefore failing to meet the U.S. *Health Insurance Portability and Accountability Act* (*HIPAA*) Safe Harbor standard for de-identification,[19] which requires the removal or generalization of 18 fields of data.[20] It also failed to meet the second de-identification standard, the Expert Determination Method (also known as the Statistical Method).[21] This is not surprising, however, since the re-identification attack occurred before these *HIPAA* standards came into effect. Therefore, the claims database did not meet any de-identification standards, then or now.

Also, while AOL had replaced names with pseudonyms, it had not de-identified the search queries themselves. Some of the queries included the actual names of individuals, which was the case for the individual who was re-identified ("vanity search queries").

Further, in the two successful attacks on properly de-identified information, the risk of re-identification was found to be extremely low. For example, in one of the two successful attacks on properly de-identified information, there was only the slightest chance of re-identification – only 0.013 percent of the properly de-identified records could be correctly re-identified.[22]

16 Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" (2010) 57 UCLA L. Rev. 1701; Gregory D. Curfman, Stephen Morrissey & Jeffrey M. Drazen, "Prescriptions, Privacy, and the First Amendment" (2011) 364 N Engl J Med 2053; and Mark A. Rothstein, "Is De-Identification Sufficient to Protect Health Privacy in Research?" (2010) 10 Am J Bioeth 3.

17 Khaled El Emam et al., "A Systematic Review of Re-identification Attacks on Health Data" (2011), online: http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0028071&representation=PDF.

18 *Ibid* at 6.

19 U.S. Department of Health and Human Services, Standards for Privacy of Individually Identifiable Health Information, Final Rule, Federal Registrar 2002; 45 CFR, Parts 160-164.

20 The eighteen fields of data include the following identifiers of the individual or relatives, employers or household members of the individual: names, all geographic subdivisions smaller than a state, all elements of dates (except year) related to an individual, telephone numbers, fax numbers, electronic mail addresses, social security numbers, medical record numbers, health plan beneficiary numbers, account numbers, certificate/license numbers, vehicle identifiers and serial numbers, device identifiers and serial numbers, Web Universal Resource Locators, Internet Protocol address numbers, biometric identifiers, full face photographic images and any other unique identifying number, characteristic or code.

21 This method requires a person with appropriate knowledge of, and experience with, generally accepted statistical and scientific principles and methods for rendering information not individually identifiable to apply such principles and methods and to determine that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information, and to document the methods and results of the analysis that justify such a determination.

22 *Supra*, note 17 at 7.

More recently, an article entitled, *Matching Known Patients to Health Records in Washington State Data*,[23] demonstrated that it was possible to re-identify individuals from publicly available state-level hospital discharge data by matching that data to other publicly available sources of information, including news stories containing the word "hospitalization." The hospital discharge data used to re-identify individuals did not include names or addresses, but it did include an individual's full five-digit ZIP code, age in years and months, race, ethnicity, gender; as well as the hospital, month of discharge, number of days in the hospital, admission type, source and weekend indicator, discharge status, how the bill was paid, diagnosis codes, procedure codes, and list of attending physicians.[24]

Re-identification was accomplished by collecting information on individuals who were hospitalized, obtained from news stories, and linking it with information in public registers available online to obtain the individual's date of birth and any five-digit ZIP code associated with the individual.[25] Once the name, date of birth and ZIP code were found, this information was then linked to the publicly available hospital discharge data. Using this method, the researcher was able to accurately re-identify individuals in 35 of the 81 cases reported in news stories – approximately 43 percent of the individuals in the sample.[26]

But even here, flaws in the initial de-identification process contributed significantly to re-identification. The researcher acknowledged that the hospital discharge data used to re-identify individuals was not de-identified in accordance with either of the *HIPAA* de-identification standards. The data included full five-digit ZIP codes, and dates were reported in years and months as opposed to years alone.[27] In a further survey of publicly available state health databases, researchers found that only a woeful three of the 33 states that release hospital discharge data, released the data in a form that complied with the *HIPAA* de-identification standards, across all fields.[28]

In addition, it is important to note that what appears to be a high re-identification rate of approximately 43 percent (35 cases), only applies to the hospital discharge records of the 81 individuals identified in the news stories containing the word "hospitalization." In fact, the hospital discharge dataset contained approximately 600,000 records. Therefore, the re-identification rate for the entire hospital discharge dataset was only 35 out of the 600,000 records, or 0.0058 percent.

Another recent study attempted to re-identify 579 of the 1130 public profiles copied from the website of the Personal Genome Project using public records.[29] The 579 profiles copied from the Personal Genome Project[30] website contained the date of birth, gender and full five-digit ZIP code of each individual. Of the

23 Latanya Sweeney, "Matching Known Patients to Health Records in Washington State Data" (2013), online: http://arxiv.org/pdf/1307.1370v2.

24 *Ibid* at 4.

25 *Ibid* at 8.

26 *Ibid* at 9.

27 *Ibid* at 11.

28 Sean Hooley & Latanya Sweeney, "Survey of Publicly Available State Health Databases" (2013), online: http://thedatamap.org/1075-1.pdf.

29 Latanya Sweeney, Akua Abu & Julia Winn "Identifying Participants in the Personal Genome Project by Name" (2013), online: http://dataprivacylab.org/projects/pgp/1021-1.pdf.

30 The Personal Genome Project is a long-term cohort study that aims to sequence and publicize the complete genomes and medical records of 100,000 informed volunteers to enable research into personal genomics and personalized medicine by creating a publicly available resource that brings together genomic, behavioral, environmental and human trait data.

579 profiles copied, 241 unique names were matched to individuals.[31] It is important to note, however, that the profiles copied from the website of the Personal Genome Project were again not de-identified in accordance with the *HIPAA* de-identification standards. For example, they included full five-digit ZIP codes. Further, 103 profiles were directly identifiable through names that appeared in the documents copied. The researchers themselves acknowledged that the risk of re-identification could be greatly reduced by making the values, such as date of birth and full five-digit ZIP code, less specific and by removing the names of individuals from the documents uploaded to the Personal Genome Project website.[32]

Further, in none of these studies did the researchers suggest that, based on their research, efforts to de-identify information should be abandoned. The research was largely aimed at improving de-identification techniques and re-identification risk management procedures, rather than encouraging the abandonment of de-identification as a tool for protecting privacy.

Context is Everything – While Individuals May be Unique, They May Not be Identifiable

Another important consideration in interpreting the recent research is that while an individual may be unique within a dataset, this does not necessarily mean that the individual was actually re-identified or that the individual may be re-identified in the future.

Unique in the Crowd: The Privacy Bounds of Human Mobility[33] describes a study conducted using non-aggregated mobility data for one and a half million individuals over a period of 15 months from April 2006 to June 2007. Each time that individuals interacted with their mobile phone by means of incoming or outgoing calls or text messages, the closest antenna recorded their transactions. The study found that in a dataset where the location of an individual was specified hourly, knowing as few as four random spatio-temporal points was enough to uniquely identify 95 percent of the mobility traces in the sample. It was also determined that 11 points were sufficient to uniquely characterize all the traces in the sample.[34] Citing other well-known studies that have shown individuals can be re-identified by linking seemingly de-identified information with publicly available sources of information, the researchers suggested that because an individual's mobility data is highly unique, there is a greater likelihood of re-identifying individuals.[35]

However, while the researchers in the study demonstrated that an individual's mobility data is highly unique, they did not actually re-identify any individuals from the mobility traces. This was not the goal of the study. The goal was to uniquely identify mobility traces in a dataset using randomly selected spatio-temporal points from each mobility trace. There was no attempt to ascertain the identity of any particular individual in the dataset. The identity of the individuals remained unknown.

31 *Supra*, note 29 at 3.
32 *Ibid* at 4.
33 *Supra*, note 12.
34 *Ibid* at 3.
35 *Ibid* at 2.

Further, although the researchers in the study suggested that re-identification could be accomplished by linking mobility data to other sources of information, they did not demonstrate how 95 percent of the one and half million individuals in the sample could be re-identified. In addition to having access to the comprehensive dataset of mobility traces, one would have to know at least four spatio-temporal pieces of information about each individual in the sample in order to re-identify the individuals. We believe that the feasibility of amassing such information from publicly available sources would be prohibitive.

Even if four specific spatio-temporal pieces of information were known for a handful of individuals and these individuals were re-identified, it is important to note that the mobility data in this study was not properly de-identified. The researchers used what they referred to as a "simply anonymized dataset that does not contain name, home address, phone number or other obvious identifiers."[36] Perturbation of the times and locations in an appropriate manner would have reduced the risk of matching four known spatio-temporal pieces of information about an individual with the mobility data.

Therefore, while the study highlighted the unique and sensitive nature of mobility data and demonstrated that this uniqueness makes it possible to isolate an individual with limited information, the researchers did not actually ascertain the identity of any particular individual by linking the four random spatio-temporal points to other sources of information. No one was actually re-identified.

Context is Everything – The Need for Time, Money and Expertise

Despite repeated claims of its ease, the re-identification of properly de-identified information is a technologically rigorous and expensive endeavour, with very limited success rates.[37] It requires the investment of significant time and financial resources and the concerted effort of individuals possessing unique skill sets. In contrast, where information has not been properly de-identified in accordance with current standards, no unique skill sets are required – simple matching exercises would likely be sufficient to re-identify individuals.

One of the common methods for conducting re-identification attacks is to use a population register, such as a voter registration list in the United States or a private property security register in Canada. One of the main reasons that re-identification is time-consuming, labour-intensive and a costly exercise in these types of attacks is because, in order to achieve a high success rate in a re-identification attack, one of the critical steps is to minimize the number of individuals missing from the population register[38] and to ensure that the quasi-identifiers in both the dataset and the population register are correct.[39] As stated by one commentator:

36 *Ibid* at 1.

37 Ubaka Ogbogu et al., "Policy Recommendations for Addressing Privacy Challenges Associated With Cell-Based Research and Interventions," (2014) 15:7 BMC Medical Ethics 1 at 3.

38 A population register is a publicly available database of personal information which is used to link or match to de-identified information in order to re-identify individuals, such as voter registration databases.

39 Daniel Barth-Jones, "The 'Myth of the Perfect Population Register' and Re-Identification Risk Assessment" in Khaled El Emam, *Risky Business: Sharing Health Data While Protecting Privacy* (Bloomington: Trafford Publishing, 2013) at 151.

> The problem is that creating a "perfect population register" – one that is complete and accurate is a tremendous challenge for even the U.S. Census Bureau and would typically be far beyond the likely abilities of a hypothetical data intruder. Not surprisingly, disclosure risk scientists themselves cannot afford to complete this final exhaustive step when making their re-identification risk estimates. So they wisely skip this last essential task and instead make easily obtained, but highly conservative, estimates of the true re-identification risks.[40]

With an incomplete or inaccurate population register, re-identification attempts will have a reduced success rate, and although individuals may seemingly be re-identified, the re-identification may be incorrect.

In 2010, researchers conducted a study to determine the re-identification risk of information de-identified in accordance with the *HIPAA* Safe Harbor standard, and to evaluate the risks and costs associated with a specific re-identification attack using voter registration lists.[41] The researchers found that the cost of obtaining the voter registration list varied widely from one jurisdiction to another, ranging from a low of $0 to a high of $17,000.[42]

Context is Everything – Location, Location, Location!

The success of re-identification attacks will necessarily be jurisdiction-dependent due to differences in the publicly available sources of information and the privacy legislation in force in that jurisdiction. Therefore, research demonstrating a successful re-identification attack in one jurisdiction will not necessarily be successful in another jurisdiction. [43]

The article entitled, *Matching Known Patients to Health Records in Washington State Data*,[44] demonstrated that it was possible to re-identify individuals from publicly available hospital discharge data by matching that data to other publicly available sources of information.[45] A similar re-identification attack, however, would not likely succeed in other jurisdictions where the amount of hospital discharge data and demographic information released is more limited, or where the types of individuals or organizations to whom the hospital discharge data are released is restricted.

For example, in the province of Ontario, Canada, hospital discharge data that includes indirect identifiers, such as complete date of birth and full postal code, is not made available to the general public because of the limitations and restrictions placed on the collection, use and disclosure of personal health information

40 *Ibid* at 150.

41 Kathleen Benitez & Bradley Malin, "Evaluating Re-Identification Risks With Respect to the *HIPAA* Privacy Rule" (2010) 17 J Am Med Inform Assoc 169.

42 *Ibid* at 175.

43 *Supra*, note 17 at 6.

44 *Supra*, note 23.

45 *Ibid* at 4.

for secondary purposes contained in Ontario's *Personal Health Information Protection Act*.[46] However, efforts are underway to make individual-level Canadian hospital discharge data that has been properly de-identified available to researchers through a pilot project that includes a detailed evaluation of the risks of re-identification, to ensure that the risk of re-identification remains very small.[47]

The Special Case of Genetic Information

The debate about the value of de-identification in protecting privacy has also been refuelled by recently published studies demonstrating that individuals may be re-identified from their genetic information.

For example, the study which matched names to individual profiles from the Personal Genome Project website[48] has been used by some commentators to suggest that individuals can be re-identified from de-identified genetic information. However, as acknowledged by the researchers themselves, the ability to re-identify individuals was not based on their genetic information but rather on their demographic information.[49]

While genetic information was not used to re-identify individuals in the particular study above, it was used to re-identify individuals in another study. In *Identifying Personal Genomes by Surname Inference*,[50] researchers at the Whitehead Institute used surname inferences from commercial genealogy databases and Internet searches to re-identify nearly 50 participants in genomics studies. The researchers began by analyzing unique genetic markers of men whose genetic material was collected by the Center for the Study of Human Polymorphisms, and whose genomes were sequenced and made publicly available as part of the 1000 Genomes Project. They then linked the unique genetic markers to publicly accessible databases that housed genetic information by surname. The surnames were then linked to other information sources, including Internet record search engines, obituaries, genealogical websites and public demographic data to identify nearly 50 research participants who had participated in the Center for the Study of Human Polymorphisms genomics study. The researchers projected that they could re-identify the surnames of approximately 12 percent of Caucasian males using this method.[51]

In drawing conclusions from this particular study, it is important to keep in mind the differences between genetic information and other types of information that are more commonly used for secondary purposes. Unlike other types of information, there are currently no standards for de-identifying genetic information. Because genetic information characterizes individuals in highly unique ways, it is difficult to de-identify this information in a manner that maintains a sufficient level of data quality necessary for most secondary

46 *Personal Health Information Protection Act, 2004*, SO 2004, c 3.

47 Khaled El Emam et al., "De-Identifying a Public Use Microdata File From the Canadian National Discharge Abstract Database" (2011) 11 BMC Medical Informatics and Decision Making 53.

48 *Supra*, note 29.

49 *Ibid* at 1.

50 Melissa Gymrek et al., "Identifying Personal Genomes by Surname Inference" (2013) 339 Science 321.

51 *Ibid* at 322.

purposes. Genetic information, unlike most other types of information, is highly sensitive to any kind of distortion, and is not amenable to standard de-identification approaches.[52] Therefore, it is not possible to generalize the findings of studies involving genetic information to other information that is more commonly used for research and other secondary purposes, such as administrative, survey, and clinical data.

There is a need for improved methods for the de-identification of genetic information[53] and additional safeguards when genetic information is used or disclosed for secondary purposes.

52 Distortion of the original data tends to diminish the utility of genetic information because the conclusions that can be drawn from the analysis of the distorted data are not the same as the conclusions that can be drawn from the original data.

53 Khaled El Emam, "Methods for the De-Identification of Electronic Health Records for Genomic Research" (2011) 3:25 Genome Medicine 1.

De-identification and Data Quality

The suggestion that information can typically only be de-identified at the expense of data quality is based on an outdated, zero-sum paradigm which suggests that two interests, in this case, privacy and data quality, are mutually exclusive and that each may only be attained at the expense of the other. While this argument may currently hold true for genetic information, it certainly does not hold true for the vast majority of information used for secondary purposes.

The second author of this paper, Dr. Khaled El Emam, the Canadian Research Chair in Electronic Health Information and senior investigator at the Children's Hospital of Eastern Ontario Research Institute, has developed a widely used tool, the Privacy Analytics Risk Assessment Tool (PARAT), which de-identifies information in a manner that simultaneously minimizes both the risk of re-identification and the degree of distortion to the original database. This tool provides a high degree of privacy protection, while also ensuring a level of data quality that is appropriate for the secondary purpose. This tool has been used to assist in the establishment of Canada's first multi-disease electronic record surveillance system, in public health monitoring and to enhance access to research data in areas such as cancer, child and maternal health, chronic diseases and disorders of the brain, including cerebral palsy, epilepsy, neurodegenerative disorders, neurodevelopmental disorders, and depression.[54]

54 http://www.privacyanalytics.ca.

Conclusion

Although skepticism about the value of de-identification persists, there is no new evidence to support the conclusion that re-identification is a trivial task and consequently, that de-identification should be abandoned as a means of protecting privacy. Such conclusions continue to be based on research involving information that was not properly de-identified. Nor do they take into account the prohibitive amount of time, energy and resources required to successfully re-identify individuals using properly de-identified information, and that re-identification in one jurisdiction may not necessarily be replicable in another jurisdiction.

De-identification, using proper de-identification methods, combined with a rigorous re-identification risk management framework, remains a strong tool for protecting privacy while ensuring a level of data quality commensurate with the secondary purpose for which the de-identified information will be used. There are, in fact, considerable risks in abandoning de-identification efforts, including the fact that individuals and organizations may simply cease disclosing de-identified information for secondary purposes, even those seen to be in the public interest. As eloquently stated by one epidemiologist, Dr. Daniel Barth-Jones:

> The reality is that, while one can point to very few, if any, cases of persons who have been harmed by attacks with verified re-identifications, virtually every member of our society has routinely benefited from the use of de-identified health information. De-identified health data is the workhorse that routinely supports numerous healthcare improvements and a wide variety of medical research activities. But in the same way that it would be difficult to point to exactly who has had their lives saved by speed limit laws, it should be quite clear that some among us owe our lives to the ongoing research and health system improvements that have been realized because of the analysis of de-identified data. Hopefully, these advancements will continue in generations to come, but unfounded fears of re-identification could have the power to derail this progress.[55]

Let us not derail this progress. Let us abandon zero-sum paradigms and faulty generalizations about the alleged "ease" of re-identification, which will take us a long way. We need to encourage the consistent use of known and effective de-identification techniques, combined with re-identification risk management procedures that are currently available.[56] We need to encourage respected public and

55 Daniel Barth-Jones, "The 'Re-identification' of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risks and Privacy Protections, Then and Now" (2012), online: http://www.futureofprivacy.org/wp-content/uploads/The-Re-identification-of-Governor-Welds-Medical-Information-Daniel-Barth-Jones.pdf.

56 One the seven principles of *Privacy by Design*, Principle 4 entitled Full Functionality — *Positive-Sum*, not Zero-Sum, seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. *Privacy by Design* avoids the pretense of false dichotomies, such as privacy ***vs.*** security, demonstrating that it ***is*** possible to have both.

private organizations to develop standards for de-identification suitable for their use cases and types of data. In addition, as noted by the Federal Trade Commission Chairwoman Edith Ramirez, we need to foster the continued improvement of proper de-identification techniques and re-identification risk management frameworks,[57] and spur research into the development of de-identification methods for genetic data. This will ensure that de-identification remains an essential tool in protecting privacy, both now, and well into the future.

57 Federal Trade Commission Chairwoman Edith Ramirez, "Protecting Consumer Privacy in a Big Data Age" delivered at The Media Institute (8 May 2014), online: http://www.ftc.gov/system/files/documents/public_statements/308421/140508mediainstitute.pdf.

**Information and Privacy Commissioner,
Ontario, Canada**

2 Bloor Street East
Suite 1400
Toronto, Ontario
Canada M4W 1A8

Web site: www.ipc.on.ca
Privacy by Design: www.privacybydesign.ca

June 2014

Information and Privacy
Commissioner of Ontario

Commissaire à l'information et à la
protection de la vie privée de l'Ontario