Looking Forward: De-identification Developments – New Tools, New Challenges

May 2013

Information and Privacy Commissioner,
Ontario, Canada

Table of Contents

Introduction 1

Recent De-identification Updates 3

- The Importance of De-identification 3
- Illustrating the Difficulty of Re-identification 4
- Worldwide Emphasis on De-identification 6
- Enhancing Trust in De-identification 7

Looking Ahead 9

- New Tools 9
- New Challenges 10

Conclusion 14

Introduction

Personal health information is extremely sensitive and can include some of the most intimate details of one's life, such as those related to physical and mental health. Strong privacy and security protections are required as the theft, loss or unauthorized use and disclosure of personal health information can result in serious consequences for the individuals involved, including discrimination, stigmatization, and emotional or economic harm. This is one reason why data minimization is extremely important. Data minimization requires that identifying information not be collected, used or disclosed if other information is sufficient for the purpose and that no more identifying information be collected, used or disclosed than is reasonably necessary to meet the purpose. One method of achieving data minimization is through de-identification. By altering and/or removing identifying information prior to its use or disclosure, individual privacy may be preserved.

In appropriate circumstances, access to health information for secondary purposes that are strongly in the public interest is very important. For example, this information is necessary for public health surveillance, health-related research and improving the quality of care. The availability of information for such purposes can result in direct benefits to individuals as well as to society as a whole. While personal health information is generally necessary for delivering health care to individuals, it is often not needed for secondary purposes. De-identification is a valuable tool in that it enables the protection of individual privacy and drastically reduces the risk that personal health information will be used or disclosed for unauthorized or malicious purposes, while also complying with data minimization practices and enabling the information to be used for authorized secondary purposes.

The value of de-identification of personal information as a tool to protect privacy has been a contentious topic in recent years. Numerous academic articles have claimed that privacy cannot be protected through de-identification. These articles contend that easy re-identification is possible, and argue that it is futile to even attempt to de-identify datasets. In response to these assertions, in 2011 the Information and Privacy Commissioner of Ontario (IPC) released *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy,* co-authored with Dr. Khaled El Emam, Canada Research Chair in Electronic Health Information, CHEO Research Institute and University of Ottawa.[1] The goal of this paper was to shatter the myth that de-identification is not a strong tool to protect privacy. De-identification protects individual privacy while also enabling the information to be used for authorized secondary purposes, such as health research, resulting in benefits to individuals and society. It can be done in a way that both minimizes the risk of re-identification and also maintains a high level of data quality. Re-identification of properly de-identified information is not an easy or trivial task, but rather requires concerted effort on the part of skilled technicians. As long as proper de-identification techniques, combined with re-identification risk measurement procedures, are used, de-identification remains an essential tool in the protection of privacy.

1 Ann Cavoukian and Khaled El Emam, *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy* (June 2011). Available online at: http://www.ipc.on.ca/English/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=1084

Since the publication of the IPC paper, there have been many key developments regarding the topic of de-identification. Numerous studies, papers and guidance documents from around the world have been released emphasizing the importance of de-identification, demonstrating that re-identification is extremely difficult, and providing guidance on de-identification strategies. There has also been a focus on enhancing trust in de-identification, for example through data sharing agreements, policies and procedures. As well, there have been advances in de-identification techniques, some of which still require further development and research. New challenges have also arisen, such as the impact of "Big Data" on de-identification and de-identified data release considerations.

In this paper, we will provide an update on recent de-identification developments. We will also look ahead and focus on new, up-and-coming, issues related to the topic of de-identification. Our objective is to encourage innovative de-identification techniques and to promote knowledge sharing and de-identification best practices. The numerous recently published de-identification papers and guidance documents show that the message that de-identification is a valuable and important mechanism in protecting personal information is being circulated around the globe. It is essential that we continue to spread the word that de-identification is a crucial step in the protection of privacy.

Recent De-identification Updates

The Importance of De-identification

The enormous value of de-identification as a tool to protect the privacy of individuals when personal information is collected, used and disclosed has been discussed in detail in the previous IPC paper. While it is not possible to guarantee that de-identification will work 100 per cent of the time, it is still an essential tool that drastically reduces the risk that personal information will be used or disclosed for unauthorized or malicious purposes. Since the release of our paper, a number of other papers have been published that also emphasize the point that de-identification, while not completely foolproof, must not be abandoned.

One such paper states that de-identified data is extremely useful for many socially beneficial purposes, the risk of re-identification of properly de-identified data is extremely small and the most significant privacy risk is from improperly de-identified data. Focusing on whether perfect de-identification is possible distracts from other, more serious, privacy threats.[2] Another academic paper discusses the important benefits of using de-identified data for research purposes. The paper emphasizes the difficulty of re-identification and states that the value is not usually worth the effort. For example, it takes significantly less skill to hack into patient records then to re-identify a research dataset. The paper argues that utility and anonymity can, and often do, co-exist: it is possible for a dataset to be useful to researchers while also protecting individual privacy. One does not have to be sacrificed at the expense of the other. The paper also focuses on the fact that the re-identification risk of properly de-identified data is trivially small.[3]

A further example is the *Science as an open enterprise* report, released by the Royal Society, a fellowship of the world's most eminent scientists. The report aims to identify the principles, opportunities and problems of sharing and disclosing scientific information. The report states that the security of personal records in databases cannot be guaranteed through anonymization procedures; however, it advocates a proportionate approach that balances the public benefit against the privacy risks.[4] The Chair of the Science as an Open Enterprise working group, which produced the report, also specifically stated that society should not abandon the use of patient data for research purposes that have already demonstrated public benefit and that have the potential to yield even more.[5]

There have also been numerous papers released in the United States that advocate for legislative reform. In light of the potential re-identification risk, some suggest that there should be stronger privacy protections for de-identified information. Suggestions include revisions to the *Health Insurance Portability*

2 Derek E. Bambauer, *The Myth of Perfection*, 2 Wake Forest Law Review Online 22 (2012).
3 Jane Yakowitz, *The Tragedy of the Data Commons*, (2011). Available online at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1789749
4 The Royal Society, *Science as an open enterprise*, (June 21, 2012). Available online at: http://royalsociety.org/policy/projects/science-public-enterprise/report/
5 Geoffrey Boulton, Risks and benefits of patient data sharing, The Guardian, September 2, 2012. Available online at: http://www.guardian.co.uk/technology/2012/sep/02/risks-benefits-patient-data-sharing

and Accountability Act (*HIPAA*) Privacy Rule,[6] contractual solutions,[7] and re-defining the concept of personally identifiable information.[8] Some of these papers cite re-identification studies and statistics in order to demonstrate that even records that have been de-identified are potentially vulnerable to re-identification. However, these papers do not suggest abandoning de-identification: they still recognize that de-identification plays an important role in the protection of personal information. For example, one article, while discussing concerns about de-identification, highlights the point that "de-identification, if done correctly, provides an important tool for privacy protection while preserving data utility for uses critical to advancing a more effective and efficient healthcare system."[9] Within these legislative critiques the message that de-identification is important still prevails.

Illustrating the Difficulty of Re-identification

Recent studies have continued to demonstrate that the re-identification risk of properly de-identified information is extremely low. Re-identification is not easy. In fact, it is very difficult. In 2011, Dr. Khaled El Emam and his associates conducted a systematic literature review to identify and assess published accounts of re-identification attacks on de-identified datasets.[10] The literature review identified fourteen accounts of re-identification attacks on de-identified data. A review of these attacks found that approximately a quarter of all the records were re-identified and approximately one third of the health records were re-identified. However, several important observations were also noted. Only six of the fourteen re-identification attacks involved health data. As well, out of the fourteen successful re-identification attacks, only two were made on records that had been properly de-identified using existing standards. The remaining eleven attacks only demonstrate that improperly de-identified data can be re-identified. In the two successful attacks on properly de-identified information, the risk of re-identification was found to be very low. As well, only one of those two attacks was made on health data. The one single attack on health data was commissioned by the United States Department of Health and Human Services in order to determine the re-identification risk of data de-identified using the *HIPAA* Safe Harbor Standard. This study demonstrated that only 0.013 per cent of the de-identified records could be correctly identified, which is a very low rate.

De-identification tools can mitigate the risk of re-identification of de-identified information. One such tool, The Privacy Analytics Risk Assessment Tool (PARAT), developed by Dr. Khaled El Emam, provides strong privacy protection while also ensuring a high level of data quality. A global data mining competition called the Heritage Health Prize asked participants to predict, by using de-identified claims data, the number of days patients will be hospitalized in a subsequent year. The most accurate prediction model would receive a three million dollar cash prize. Patient privacy is very important when publicly disclosing a large health dataset. The risk of re-identification is a key concern. The Heritage Health Prize decided to

6 Sharona Hoffman and Andy Podgurski, *Balancing Privacy, Autonomy, and Scientific Needs in Electronic Health Records Research,* (2012). Available online at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1923187

7 Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal, (2011). Available online at:* http://ir.lawnet.fordham.edu/iplj/vol21/iss1/2

8 Paul M. Schwartz and Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, (2011). Available online at:http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1909366

9 Deven McGraw, *Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data*, (2012). Available online at: http://jamia.bmj.com/content/early/2012/06/25/amiajnl-2012-000936.full

10 Khaled El Emam et al. *A Systematic Review of Re-identification Attacks on Health Data*, (2011). Available online at:http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028071

use PARAT to create the de-identified claims dataset.[11] Prior to releasing the dataset created using PARAT, the de-identified dataset was subject to strong simulated re-identification attacks in order to ensure that the re-identification probability was acceptably low. It was estimated that the probability of re-identifying an individual in the de-identified dataset was 0.0084, far below the already low threshold specified by the competition. This shows that by using proper de-identification tools that involve re-identification risk measurement techniques, it is extremely unlikely that an individual in a de-identified dataset will ever be re-identified.

A recent paper, written by Daniel Barth-Jones, demonstrates that re-identification is expensive, time consuming, rarely successful, requires advanced skills and it is almost always uncertain as to whether the re-identification is actually correct.[12] The paper critically re-examines Latanya Sweeney's 1997 re-identification of Governor Weld, a famous case which is frequently cited to support beliefs that computer scientists can re-identify individuals with astonishing ease. Latanya Sweeney matched hospital claims data distributed by the Massachusetts Group Insurance Commission with demographic information found in a Cambridge voter registration list. However, Daniel Barth-Jones states that the re-identification of Governor Weld lacked almost all the typical re-identification challenges as it was clear he was in both the hospital claims data (it was widely reported he collapsed and was taken to the hospital) and in the voter registration list (as governor, he would have been the subject of ballot casting photo-ops). Governor Weld was most likely re-identified because he was a public figure who experienced a highly publicized hospitalization.

To be at risk of re-identification, a person would have to be in both the health information dataset (in this case, the hospital claims data released by the Massachusetts Group Insurance Commission) and in the population register (in this case, the voter registration list). However, it is extremely difficult to create a complete and accurate population register. The voter list Latanya Sweeney used was missing almost half of the Cambridge population. In the United States, about 29 per cent of the voting age population is not registered to vote. As well, online data frequently contains errors. For example, people move and do not update their address information. To be certain that an individual has been properly re-identified, the attacker must confirm that all the information in both the population register and health information dataset is correct. The attacker must also confirm that there are absolutely no individuals missing from the population register that could have the same characteristics (e.g. year of birth, gender and three-digit ZIP code) as those in the healthcare dataset. Otherwise a proper re-identification has not taken place. This is an expensive, difficult and time consuming undertaking. If the Governor Weld re-identification were attempted today, under the *HIPAA* Safe Harbor Standard, an attacker would have to confirm that there was not at least one other 50 year old male in the same three digit ZIP code (a population of 1.25 million) that had not registered to vote.

Daniel Barth-Jones cites a widely reported statistic, stated by Latanya Sweeney, that 0.04 per cent (4 in 10,000 or 1 in 2,500) of individuals in the United States within datasets de-identified using the *HIPAA* Safe Harbor Standard can be re-identified on the basis of their year of birth, gender, and three-digit ZIP

11 Khaled El Emam et al. *De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset* (2012). Available online at: http://www.jmir.org/2012/1/e33/

12 Daniel C. 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). Available online at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397

code. However, he states that approximately one third of those individuals would most likely not be re-identified because they are not in voter lists. This means the risk of being re-identified through the use of a voter list is even lower, and the odds would be approximately 1 in 3,500, not 1 in 2,500. This risk falls somewhere between one's lifetime odds of being personally struck by lightning, and the risk of being affected because someone close to you has been struck. This paper highlights the inherent difficultly in attempting to re-identify an individual in a properly de-identified dataset, and emphasizes the extensive time, money and skill involved.

Worldwide Emphasis on De-identification

Governments across the globe are beginning to appreciate the value of de-identification as an important mechanism in the protection of privacy. Recent reports and guidance documents have been released that encourage de-identification and provide new strategies and advice. In Canada, the Pan-Canadian Health Information Privacy Group released a paper consisting of common understandings – principles that the group believes should be adopted consistently to support appropriate and privacy-protective trans-jurisdictional disclosures of electronic health record information.[13] While acknowledging that it is not possible to guarantee that de-identified data will never be re-identified, a number of common understandings in the paper still focus on de-identification. For example, it is a common understanding that "trans-jurisdictional disclosures for secondary uses should, as a general rule, involve aggregated or de-identified information." Another common understanding is that "entities and individuals responsible for handling requests for trans-jurisdictional disclosures of [electronic health record] information for secondary uses should be knowledgeable about de-identification, up-to-date on de-identification techniques, and be able to apply them." The paper emphasizes the importance of de-identification, as well as re-identification risk assessment procedures.

In the United States, a number of reports have been released that acknowledge the necessity of de-identification. The American College of Physicians released a position paper that included policy recommendations related to health information technology and privacy. The paper included numerous policy positions related to de-identified data. For example, using de-identified data wherever possible and appropriate; making appropriate de-identified patient data available for socially important activities, such as health research, with institutional review board approval and adherence to de-identification standards; and educating the public about the benefits to society that result from the availability of appropriately de-identified health information.[14] The Federal Trade Commission issued a report containing a privacy framework that includes best practices for companies that collect and use consumer data.[15] The report states that if a dataset is not reasonably identifiable, the company publicly commits to not re-identifying it, and the company requires users of the dataset to keep it in de-identified form, then that dataset will be outside the scope of the framework. The report also states that companies and researchers are

13 Pan-Canadian Health Information Privacy Group, *Privacy and EHR Information Flows in Canada, Version 2.0*, Canada Health Infoway, (2012). Available online at: https://www.infoway-inforoute.ca/index.php/resources/reports/privacy

14 American College of Physicians, *Health Information Technology and Privacy*, (2011). Available online at: http://www.acponline.org/advocacy/where_we_stand/policy/

15 Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers*, (2012). Available online at: http://www.ftc.gov/opa/2012/03/privacyframework.shtm

encouraged to continue innovating in the development and evaluation of new and better approaches to de-identification.

The United States Department of Health & Human Services Office of Civil Rights also released a guidance document that addresses common questions about the two methods (Expert Determination and Safe Harbor Standard) that can be used to de-identify health data in accordance with the *HIPAA* Privacy Rule.[16] It is intended to help covered entities to understand de-identification, the process by which de-identified information is created and the options available for performing de-identification. The document specifically states that the process of de-identification mitigates privacy risks to individuals and supports the secondary use of data.

The United Kingdom Information Commissioner's Office released a code of practice explaining the issues surrounding the anonymization of personal data and the disclosure of data once it has been anonymized.[17] The code of practice provides practical advice and describes the steps an organization can take to ensure that anonymization is conducted effectively while retaining useful data, showing that effective anonymization of personal data is possible, desirable and can help service society's information needs in a privacy-friendly way. It specifically states that anonymization safeguards individual privacy and is a practical example of the *Privacy by Design* principles that data protection law promotes. The United Kingdom Information Commissioner's Office also announced that it will be funding a new United Kingdom Anonymisation Network (UKAN) that aims to enable sharing of good practices regarding anonymization across the public and private sectors.[18]

Enhancing Trust in De-identification

De-identification is of enormous value and dramatically decreases the likelihood that personal information will be used or disclosed for unauthorized or malicious purposes; however, there cannot be 100 per cent certainty that an individual in a de-identified dataset will never be re-identified. While proper de-identification techniques and re-identification risk measurement procedures are crucial, it is also extremely important to have mitigating controls for the de-identified data. Mitigating controls, such as a data sharing agreement, can discourage data recipients from re-identifying individuals in the dataset. At minimum, the data sharing agreement should include provisions that prohibit the recipient from re-identifying, or attempting to re-identify, the data; limit the uses and disclosures of the data to those specifically mentioned in the agreement; ensure that the recipient has strong administrative, physical and technical safeguards in place; permit audits at the recipient's site and of the recipient's practices and procedures; and require staff privacy training and signed confidentiality agreements for those having access to the de-identified database. The data sharing agreement should contain provisions related to breach notification and should also include penalties for breaching the terms of the agreement. This brief list of the types of provisions

16 Department of Health & Human Services Office of Civil Rights, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability (HIPAA) Privacy Rule*, (2012). Available online at: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

17 Information Commisioner's Office, *Anonymization: managing data protection risk code of practice*, (2012). Available online at: http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/anonymisation.aspx

18 Information Commissioner's Office, New anonymization code sets out how to manage privacy risks and promote transparency, (2012). Available online at: http://www.ico.gov.uk/news/latest_news/2012/new-anonymisation-code-sets-out-how-to-manage-privacy-risks-and-maintain-transparency-20112012.aspx

that should be included in a data sharing agreement is not comprehensive. Specific details of what to include in data sharing agreements should always be reviewed with legal counsel.

An organization involved in de-identification should make sure to have a comprehensive de-identification policy and procedures in place. The policy and procedures should be developed and implemented prior to any de-identification activities and are necessary regardless as to whether the de-identification will be done in-house or by an external vendor. The policy and procedures should address the situations when de-identification needs to occur and the associated controls that need to be in place. The policy should specify the person responsible for maintaining the policy and procedures, the person responsible for enforcement and the people who must follow the policy and procedures. The policy should also address who is responsible for overseeing the de-identification process and should specify that this person will ensure that only authorized de-identification methods are used and that the de-identified data is secure and not disclosed inappropriately.[19] The de-identification procedures should specify the de-identification methods that will be used. These methods should be verified and include risk measurement assessment procedures, having regard to information that can be used to identify an individual both directly and indirectly. If the de-identification will be conducted outside of the organization, the policy should specify the approved and qualified vendors. The policy and procedures should also prohibit re-identification, or attempted re-identification, of de-identified datasets. The policy and procedures should also address the consequences of a breach.

It is not enough to simply develop a de-identification policy and procedures. The policy and procedures must also be reflected through the actual practices of the organization. All employees, especially those who will be accessing the de-identified data, should understand the de-identification policy requirements. Employees who will be de-identifying data must be properly trained on the appropriate de-identification procedures. Effective communication is a crucial step towards ensuring appropriate implementation of the de-identification policy and procedures. Employees should also be made aware of any notable amendments or updates to the policy and procedures.

As well, de-identification procedures should be reviewed and assessed periodically to ensure that the re-identification risk remains appropriately low. De-identification techniques are continually evolving and improving. At the same time, advances are being made in re-identification technology. It is important to continually reassess and strengthen de-identification risk management techniques.

Many uses of de-identified data occur with limited public knowledge and this lack of transparency can contribute to public concerns about de-identification. Greater transparency can help build public trust in the use of de-identified datasets.[20] To enhance trust, organizations can explain why personal information is de-identified, describe in general terms the techniques that will be used, be open about any potential risks and describe the safeguards that are in place to minimize the risk of re-identification.[21] By providing information about de-identification, organizations can assist in raising awareness of de-identification as a necessary tool to protect individual privacy while also enabling valuable secondary uses of the data that are in the public interest.

19 Rebecca Harold, *Implementing a Data De-identification Framework*, (2012). Available online at:http://www.infosecisland.com/blogview/22733-Implementing-a-Data-De-Identification-Framework.html?utm_source=twitterfeed&utm_medium=twitter
20 *Supra*, note 9.
21 *Supra*, note 17.

Looking Ahead

New Tools

De-identification Techniques

There continues to be exciting advancements in de-identification technology, resulting in enhanced privacy protection and further assurance that de-identified information will remain anonymous. Dr. Khaled El Emam has developed a protocol for securely linking databases without sharing any identifying information.[22] This protocol has been used to identify and locate records relating to an individual that exist in more than one dataset. For example, if a person with cancer lives in New York for part of the year and in Florida for the other part, and has been hospitalized in both locations, that person could be in more than one state cancer registry. Population based cancer registries monitor the frequency of new cancer cases every year in order to recognize and reduce risks. The state in which the patient resides at the time of the diagnosis is very important. For data accuracy, it is important that patients are only counted once and that information is not duplicated in more than one state registry. One way to ensure the information is not duplicated is to collaborate and have each state exchange data to identify residents of other states that are in more than one registry. However, this raises privacy concerns because the personal health information of patients whose information is only in one registry will now be disclosed to another registry. Dr. Khaled El Emam's secure matching protocol uses an encryption system to identify an individual that may exist in multiple datasets. It involves encrypting personal identifiers in each dataset and then comparing the encrypted identifiers using mathematical operations. A list of matched records can be obtained without revealing any personal health information of anyone in either dataset. This is a win-win, positive sum solution that enables both privacy and data quality to be maintained.

Alternatives to De-identification

Differential privacy is a relatively new privacy model that has been gaining attention in recent years. Under the differential privacy model, personal information in a large database is not modified or released. Instead, a third party, such as a researcher, can submit questions about the information in the database by going through an intermediary piece of software that serves as a privacy guard. The privacy guard establishes the privacy risk of the question based on that question and any others that have preceded it. It then gets the answer from the database, and mathematically distorts it before sending it back to the third party. The distortion is small enough that it does not significantly affect the quality of the answers, while large enough to protect the identity of any individual whose data is in in the database. A privacy budget is assigned to the database and the guard keeps track of the cumulative privacy cost of all the questions. If the question has the potential to create a privacy breach, the guard will increase the amount of distortion. At this point the answer may not be useful so the third party may abandon the query or ask

22 Khaled El Emam et al., *A Protocol for the Secure Linking of Registries for HPV Surveillance*, (2012). Available online at: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0039915

a more general question. The third party does not ever access or see the contents of the database and the underlying data does not need to be modified or distorted.[23]

Differential privacy is still in its early stages. There are potential mathematical, computational and policy challenges that need to be addressed. There are also numerous healthcare specific considerations. Differential privacy requires data distortion, which means that the released data will be at least slightly imprecise. However, certain healthcare studies require perfect reporting. Distortions to the data can produce results that do not make sense (e.g. showing two drugs that are known to interact in a way that can be damaging to a patient's health, or showing a drug that would normally never be prescribed with a particular treatment). This can erode the trust of data analysts and potentially act as barriers to the acceptability of techniques used to protect data privacy.[24] There have been limited disclosures of differentially private health data and consequently few examples of useful and valid analytical results. While further research into its theoretical and practical limitations is required, differential privacy has the potential to be a valuable tool in the protection of privacy.

New Challenges

De-identification of Genetic Information

Genetic information is becoming more prevalent in research and healthcare. Due to recent technological advances in biomedical research, genetic analysis enables an increase in knowledge about disease processes and individual variations in treatment effectiveness or susceptibility to disease. Genomic research can help decipher the roles that genetics and the environment play in the origin of common but complex diseases, such as cancer and diabetes.[25] Genetic information has the potential to greatly advance clinical care and general health and can provide a vast amount of information for addressing long standing questions about health and disease. However, patient privacy is also a serious concern that must be addressed. Privacy concerns extend beyond those of the participant. Genetic testing may also reveal personal information about the family members of the individual, who most likely did not consent to the use of their genetic information.[26] Genetic information may also affect an individual's employment and insurance status, potentially resulting in discrimination. For example, employers may avoid hiring an individual who they believe may have health problems and would be likely to be absent frequently, take sick leave and/or resign or retire early.

Genomic privacy is particularly challenging given the vast amount of data that is required. There is a trend toward setting up large scale population biobanks. These biobanks can be very helpful for research purposes; however, they can pose privacy risks as well. For example, recently researchers were able to determine the identity of nearly fifty individuals who had anonymously submitted genetic samples that

23 Microsoft, *Differential Privacy for Everyone*, (2012). Available online at: http://www.microsoft.com/en-ca/download/details.aspx?id=35409

24 Fidal Kamal Dankar and Khaled El Emam, *The Application of Differential Privacy to Health Data*, (2012). Available online at: http://dl.acm.org/citation.cfm?id=2320816&dl=ACM&coll=DL&CFID=173632698&CFTOKEN=11462744

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

26 Presidential Commission for the Study of Bioethical Issues, *Privacy and Progress in Whole Genome Sequencing*, (2012). Available online at: http://www.bioethics.gov/cms/node/764

were listed in a publicly-accessible research database.[27] The researchers determined the identity of the individuals through internet searches and genealogy websites. It took a single researcher with an internet connection about three to seven hours per person.

De-identification is, of course, an essential way of protecting the privacy of participants. However, de-identification of genetic data is still in its early stages. Improved methods for the de-identification of genome sequences or genomic data are needed.[28] Even if parts of a DNA sequence are suppressed, a skilled geneticist can most likely reconstruct the missing pieces. It is important to continue to accelerate the development and use of emerging privacy enhancing technologies that can allow for the analysis of genetic information while also protecting the privacy of participants. In the meantime, until the de-identification of genetic data is no longer in its infancy, individuals should be cautious about posting their raw genetic data publicly.

Big Data

Today is the age of "Big Data." Increasing computer power combined with a greater availability of information has resulted in vast amounts of data being collected, stored and analyzed. This data is generated from an endless array of online sources, such as social networking sites, search queries, and online transactions. Big Data has benefits to both individuals and society as a whole. For example, it can boost the economy, help businesses increase productivity, create new opportunities through the use of business intelligence, analysis and analytics and advance scientific research.[29] In the healthcare sector, a frequently cited example of the benefits of Big Data is Google Flu Trends, which can predict and locate flu outbreaks through aggregated Google flu-related search queries. Early detection of a disease outbreak can reduce the number of people affected. Up-to-date estimates can enable better response to seasonal epidemics and pandemics.

Although Big Data has many potential benefits, it also poses many risks to privacy. As masses of information are linked across multiple sources it becomes more difficult to ensure the anonymity of the information. There is a risk that information, while appearing non-identifying, can be combined with information from other sources to eventually produce data that may potentially be linked back to a specific individual. Privacy risks may largely be addressed through the use of proper de-identification techniques combined with risk measurement techniques. However, as different sources of data are continually combined, there is a constant need to evaluate whether the information remains unidentifiable. The de-identification applies to data that has been linked with other information, as well as the original data. It is essential to ensure that the risk of re-identification continues to be acceptably low.[30]

27 The Boston Globe, *Using simple tools, scientists show privacy of research participants is at risk*, (2013). Available online at: http://www.boston.com/news/science/blogs/science-in-mind/2013/01/17/using-simple-tools-scientists-show-privacy-research-participants-risk/dq3G4lXohz93YqvE14FvuO/blog.html

28 Khaled El Emam, Methods for the de-identification of electronic health records for genomic research, (2011). Availlable online at:http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129641/

29 Omar Tene and Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, (2012). Available online at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2149364

30 Khaled El Emam, *Perspectives on Health Data De-identification*, (2012). Available online at: http://www.privacyanalytics.ca/knowledgebase/index.php?/article/AA-00280/0/Perspectives-on-Health-Data-De-identification-compilation.html

While Big Data can have privacy risks, it can also make re-identification more unlikely. There is an extensive amount of information currently available online. Much of this information is identified data rather than anonymous data. For example, people may use their real names while writing blogs and using social networking sites. Given the data mining opportunities available with identifiable information, it is highly unlikely that an adversary will find it worth his money and time to learn the complex skills necessary to properly re-identify a database.[31] An advantage of having Big Data to analyze is that it makes de-identification easier to achieve. There is a greater likelihood that there are more similar people in a large dataset than in a small one. By contrast, smaller datasets are more challenging to de-identify as it is easier to be unique in a small dataset.[32] As uncertainty is introduced is it harder to know if the information truly corresponds to a specific individual. Big Data has its benefits, as well as its risks. It is necessary to ensure privacy best practices are used. One of the most important best practices is to continue to de-identify information at the earliest opportunity.

Limited Access or Open Access

The risk of re-identification of a de-identified dataset can vary according to the way the de-identified information is released. There is a clear difference between releasing de-identified information to the world at large, where anyone can access the information and potentially attempt to re-identify it, and limited access to the de-identified data, for example to a specific research community where there are data sharing agreements that restrict further disclosure.[33] Limiting access to a closed community minimizes the risk that re-identification will occur as robust safeguards can be put into place. It is essential to reduce the risk of re-identification and ensure the privacy of individuals whose information is contained in the de-identified databases.

Despite the advantages of limiting access, there are many benefits that occur due to open access to large de-identified databases. Nearly every recent public policy debate has benefited from mass dissemination of anonymized data. For example, public-use birth data has led to advances in understanding the effects of smoking on fetuses and research performed using Medicare and Medicaid data is central to debates about U.S. health care reform.[34] There is increasing pressure for researchers to make data publicly available. For example, as of January 1, 2013, the Canadian Institutes of Health Research (CIHR), the Government of Canada's health research investment agency, requires CIHR-funded researchers to make their peer-reviewed publications accessible at no cost within twelve months of publication.[35] They also require certain data, such as gene sequences, to be deposited into public databases. De-identified datasets can be of critical importance for research. New technology is increasing the benefits from analysis of large datasets in ways that may not be predictable in advance. Some of the most useful data may be originally collected for a completely unrelated and unanticipated purpose. Restricting the flow of information to a limited number of researchers may prevent new discoveries and overly constrain

31 *Supra*, note 3.
32 *Supra,* note 25.
33 *Supra*, note 17.
34 *Supra*, note 3.
35 Canadian Institutes of Health Research, *CIHR Open Access Policy,* (2013). Available online at: http://www.cihr-irsc.gc.ca/e/32005.html

information flow.[36] Open access also encourages transparency and replicability and creates an important research dialogue. It is important that data is assessable so that judgements can be made about its reliability. Open access facilitates early detection of errors or even fraud. It enables researchers to test, refute, reinforce or built on the results of other researchers.[37]

A balance must be made between protecting privacy and supporting secondary uses of the data. It is important to take a proportionate approach and weigh the potential benefits of open access against the potential harms. Data release considerations include disclosure risk (there is less risk of re-identification if the data is properly de-identified using verified risk assessment techniques and tools); utility of the data (is the data still useful for research after de-identification); sensitivity of the data (de-identified personal health information is especially sensitive and may be more suitable for limited access); evolving technical risks; and developments in techniques designed to safeguard privacy. It is important to recognize that there is no way to completely guarantee that the information will never be re-identified (although, as mentioned previously, the re-identification risk of properly de-identified information is extremely low); however, it is also important to take into account the potential vast public benefits of open access to the data. There is no one size fits all answer. It is necessary to weigh both the risks and the benefits in each case in order to determine whether limited access or open access is preferred.

36 *Supra*, note 8.
37 *Supra*, note 4.

Conclusion

Proper de-identification remains an extremely important and effective method of protecting privacy. While it is encouraging to note that recent papers have been published that promote the value of de-identification and demonstrate the difficulty of re-identification, we all must continue to encourage innovative de-identification techniques. There have been many advances made; however, certain areas require further development and research.

Awareness is just the first step. The importance of knowledge sharing and promoting best practices cannot be overstated. It is essential to stay ahead of the curve and to be informed of the most up-to-date de-identification techniques and re-identification risk measurement procedures. In order to foster proper de-identification techniques and best practices and demonstrate the importance of de-identification, my office has developed a De-identification Centre – a de-identification webpage found on the *Privacy by Design* website. We all must continue to spread the word of the importance of de-identification. However, we must also constantly and consistently encourage innovation and promote knowledge sharing in order to ensure that proper de-identification remains a key step in the protection of privacy.

Information and Privacy Commissioner of Ontario

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

Web site: www.ipc.on.ca

May 2013

Information and
Privacy Commissioner,
Ontario, Canada