Privacy and Security WG Submission

Khaled El Emam

4th December 2014

Thank you for the opportunity to present to the HIT Policy Committee, Privacy and Security Working Group.

I will focus my comments on what is needed to protect patient or participant (in the case of clinical trials) privacy when their data is used and disclosed for secondary purposes. The key points in this document are as follows:

1. The narrative on re-identification attacks is misleading and not based on an accurate interpretation of the evidence. This is leading data custodians to stop sharing data. There is a need to have a better informed conversation and a proper evidence-based narrative about re-identification risks.
2. There is a need for de-identification standards to help with the adoption of good practices and make it easier for data custodians to share data responsibly.
3. The risks from continued use of the HIPAA Privacy Rule Safe Harbor de-identification method need to be evaluated.
4. There is a need for "ethics councils" within organizations that perform analytics on health data to manage the risks from inferences, and guidance is required on how to set up and run such councils. This is a practical way to manage inference risks from data.
5. The main benefit from safe havens is that they would allow the sharing of data sets that have experienced less de-identification than when raw data is shared. But that data still needs to be de-identified to some extent.
6. Before deploying distributed computation systems in practice, it is important to develop more formal security proofs for them. A number of systems deployed today may be leaking personal information under certain circumstances.

There are generally two types of disclosure that are considered in the context of disclosure control that need to be protected against. One type of disclosure is *identity disclosure*. This is when an adversary is able to assign a correct identity to a record. When discussing identifiability or information that is not personal information, then one is referring to that type of disclosure – that of assigning an identity to a record. A de-identified data set is one where the probability of identity disclosure is very small.

De-identification and Re-identification

There is a large body of work on how to effectively de-identify data sets that spans multiple decades (see the literature covered in some recent books on the topic [1], [2]).

While there have been statements made by some academics and further popularized by the media about the ease of re-identification, such statements miss important nuances and are not consistent with the evidence. There are two critical factors to note:

1. It is necessary to distinguish between different types of health data. Genomic data is difficult to de-identify and good methods for doing so are still in the research phase. These methods utilize secure computation protocols. It will likely be a few years before this research produces results that can be scaled into practice.
2. Other types of data, such as administrative, clinical, trials, and survey data, can be de-identified using existing techniques. There are good de-identification techniques that are based on generally accepted statistical and scientific principles. All known re-identification attacks that have been successful on this type of data were performed on data sets that were not de-identified properly [3]. There is just no evidence that a data set that has been de-identified using generally accepted statistical and scientific methods can be re-identified with a high success rate.

General statements about the ease of re-identification are incorrect at best, and quite misleading in practice because they are not evidence based. As far as we know, de-identification using generally accepted statistical and scientific methods is the one of the best ways to protect patient and participant privacy when data is shared for secondary purposes.

Adoption of De-identification Methods

There is a real challenge with the adoption of de-identification methods. For example, a recent systematic review of evidence on sharing data for public health purposes noted that *"A clear distinction between data containing personal identifiers and fully anonymous data may not always be possible, leading to restrictive policies on all types of data sue to privacy concerns."* [4]. What this is saying is that the misleading statements being made about the ease of re-identification and the publicity surrounding re-identification attacks on data sets that were not de-identified properly, is contributing to creating an environment where data custodians are refusing to share their data. A concerted effort is needed to change that narrative and adopt a more evidence-based and nuanced approach to this issue – otherwise the promise of the benefits of data sharing will take a long time to materialize.

How can we do that ?

There are two things that would be extremely helpful:

- Change the narrative to be more evidence-based. Which means that the correct statements need to be made about re-identification risks. It means that the health data community should be informed that known re-identification attacks were performed on poorly de-identified data (in some cases on data sets that were not even de-identified in that they had individual's names left in them).
- There is an urgent need for de-identification standards. While the HIPAA Privacy Rule provides some methods for de-identification, these need to be operationalized into detailed standards that individuals can follow. The lack of standards means that data custodians have to figure out what they need to do by themselves, and that there isn't a large pool of de-identification experts and an ecosystem of services to support de-identification.

The adoption of currently known (and used) de-identification methods will facilitate the sharing of health data. We need to increase these adoption levels.

Do We Need to Change De-identification Regulations ?

The existing de-identification methods stipulated under the Privacy Rule have been in use for approximately a decade, and therefore there is significant knowledge and experience about their strengths and weaknesses.

Much has been written about the weaknesses of the Safe Harbor method for de-identification in that it reduces data utility significantly, but also that it is not that protective (unless its assumptions are met, but that is quite uncommon in practice) [5]. The consequence of that data sets that have a high probability of re-identification are likely being shared under the cover of Safe Harbor.

There are practical advantages to having such a simple de-identification method. But doing so creates additional privacy risks. A useful debate to have is whether this approach to de-identification should be retired.

The second method in the Privacy Rule, the Expert Determination method, has some good common-sense principles from which strong de-identification processes can be developed that can also maximize data utility for complex data sets. These principles can be paraphrased as follows:

- The de-identification of health data should use generally accepted statistical and scientific methods. There is a large body of work on disclosure control which can serve as the basis for what is considered generally accepted.
- De-identification should be performed by experts with the appropriate knowledge and experience with disclosure control.
- Health data sets that are considered de-identified should have a very small risk of re-identification. This means that the concept of risk has to be operationalized for clinical trials and the “very small” threshold needs to be defined.
- The methods and results of the de-identification must be documented and the documentation retained. The period of retention is not specified, but it would be reasonable that it should at least be as long as the data is available for secondary analysis.

To the extent that these can become the guiding principles for the development of de-identification standards, then the health data community would have a strong basis for sharing data in a privacy protective manner.

Learning Something New from the Data: Attribute Disclosure[1]

The second type of disclosure is called *attribute disclosure*. With this type of disclosure an adversary learns something new about individuals from an analysis of the data. Here we are talking about inferences from the data. Inferences can be simple, such as “all 57 year olds in the data set have had

[1] The following section is taken mostly from [6].

heart attacks". This inference has absolute certainty in that all 57 year olds have the attribute of a heart attack. An inference can be more complex where multiple variables are used, for example, to predict the probability of a patient being re-admitted to a hospital or the probability of being diagnosed with a particular type of cancer. In this case the inference is achieved through a statistical or machine learning *model*. For instance, in the cancer risk model, the variables used may be where the patient lives, gender, race, age, and other diagnoses that a patient has.

Model-based inference is often not absolute and there is some uncertainty or inaccuracy. For our purposes, we will consider attribute disclosure to have occurred if a model can be constructed with a high certainty or a high accuracy (predictive or descriptive).[2]

In the following we will use the cancer probability inference to illustrate various points, and assume that the model built from the data is highly accurate.

A model can be built from de-identified data: A data set that has a very small risk of identity disclosure can be used to build that cancer diagnosis model. Also, the same model can be built from identifiable data. In fact, the risk of identity disclosure is orthogonal to model building.

Once a model is built we can start drawing inferences and learning new things. Inferences are used to make decisions. Decisions can be about groups of individuals or specific individuals. Group decisions can be made without knowing the identity of any cancer patients, for example a health authority may develop cancer screening guidelines for all men when they reach a certain age based on the results of the model. Decisions can also be about individuals – for example individuals of a certain race and age who live in high risk areas may be personally visited by a nurse to discuss lifestyle choices.

The individual-level decisions can be targeted at patients who were not even in the data set. Once that model is constructed it can be used at some future point to predict cancer diagnoses for other patients, even those who are not born yet and could not conceivably be in the data set.

All of these distinctions are important because they affect how we deal with privacy risks.

The challenge is that sometimes group or individual decisions are discriminatory, creepy, surprising, or stigmatizing. For example, a property valuation firm may reduce the value of all homes in the high risk areas because there may be a pollution source causing the higher rates of cancer. In this case, all individuals living in those areas suffer an economic harm because the model was used to make a broad group decision. Alternatively, a bank may impose higher interest rates on loans for specific individual customers of a certain gender, race, and age and living in high risk areas.

The same model can be used to make socially acceptable and socially beneficial decisions as well as to make stigmatizing or discriminatory decisions. In our example of the cancer diagnosis model, it can be used to develop improved health care services to high risk communities, or it can be used to discriminate against these communities. The model is not the problem, it is the decisions that are made from the model. The determination of whether a particular decision is appropriate or not will be subjective and contingent on prevailing social norms.

Algorithmic techniques are sometimes proposed as a solution to this inference problem, t-closeness and l-diversity. These techniques are not used in practice (I do not know of a single real world application). The reason is that techniques which modify the data to limit inferences significantly diminish the analytic utility of the data. There are two reasons for this:

- These techniques assume that all models from the data will be used to make inappropriate decisions, and therefore the data needs to be modified to ensure that no models can be built. For

[2] We will not define "high" more precisely here because that definition will not affect the logic of our argument. We will assume that it can be defined in a context-specific manner.

example, the data would be modified so no cancer diagnosis models can be built since it is possible to make inappropriate inferences and decisions from such models.

- Because of the above, many useful models cannot be built from the data and therefore the data becomes quite useless for analytics purposes.

The determination of whether an inference from a data set or a decision from a model is appropriate or not is a subjective decision. This is why it is not amenable to automation.

Governance Mechanisms to Manage Risks from Inferences

To protect individuals from inappropriate decisions, it is important to manage the risks from the *use* of the models. An appropriate solution to attribute disclosure then is to put in place governance mechanisms that oversee the development and use of the models. Let's call this *privacy ethics*. A group of individuals within a data controller or data processor would advise the business about whether the model and its uses are discriminatory, stigmatizing, creepy, or surprising. Let's call this a *privacy ethics council*. The ethics review process has been in use for a long time in the research community and has worked quite well to ensure ethical data collection, analysis, and decision making. We are proposing to replicate a lighter version of that type of review more broadly. A privacy ethics council would have a lay person representing the data subjects, a privacy expert, an ethicist, a person representing the business, and a person representing the brand (public relations). This council needs to be independent in order to give un-coerced advice.

An earlier opinion by the Article 29 Working Party provides some good criteria that such a council can consider to determine whether the model and its use would be appropriate [7]:

- The relationship between the purposes for which the data have been collected and the purposes for model-based decision making.
- The context in which the data have been collected and the reasonable expectations of the data subjects as to their further use.
- The nature of the data and the impact of the model-based decisions on the data subjects.
- The safeguards applied by the controller to ensure fairness in decision making and to prevent any undue impact on the data subjects.

It should be noted that the application of such criteria is going to be subjective, and it may not always be possible for a data custodian to know in advance all the possible models and decisions that can be made with a de-identified data set that is shared. For example, how would an ethics council know a priori if a cancer diagnosis model would be used for discriminatory purposes? It would be a problematic outcome if they erred on the conservative side because then they would likely not share any cancer data due to a possibility of some data processor using the data to make discriminatory decisions. In such cases conditions of use may accompany the de-identified data to manage those risks.

The combination of de-identification techniques that address identity disclosure only and governance mechanisms in the form of an ethics council would address the risks from identity and attribute disclosure.

Should "Uses" be Regulated ?

Continuing with the example above, it will be difficult to formulate regulations that limit or constrain uses because it will be difficult to anticipate all possible uses in all contexts. That is why a council of individuals within an organization are best positioned to make these admittedly subjective decisions. Regulations can require the creation of such a council whenever there are analytics initiatives within an organization.

What would be useful is guidance for setting up these councils and for making these decisions. The guidance needs to be principles-based so that it would not need to be updated on an on-going basis, and it would apply to commercial as well as academic research data custodians.

Privacy Architectures for Big Data: Safe Havens[3]

The discussion above emphasized data de-identification and governance focused on uses as the primary mechanisms for dealing with the privacy issues when sharing data for secondary purposes. Alternatives approaches have been proposed.

One approach is data enclaves or safe havens.

Safe havens are becoming a popular institutional model for making health data available for secondary purposes. The most common secondary purpose is data analysis. The basic idea is that instead of giving data to the data user (e.g., on a DVD or by allowing them to download it) so that the user can do the analysis on their own machines, the data user is given read-only access to the data. This access can be achieved through a number of different mechanisms, such as a portal, for example. The thinking is that this approach is safer than giving the users actual data.

It should also be noted that the term "safe havens" is sometimes used to refer to many different things, and some of the things that we characterize as safe havens are called something else. We have a terminology problem. For example, terms such as "virtual data laboratories", "remote access", and "research data centers" are other terms used to describe what we mean by safe havens. So let's start off by being a bit more precise.

What is a Safe Haven

The objective of a safe haven is facilitate data access without giving the users actual data files. There are variants to the safe haven design. The first is that the data user accesses the data remotely using a secure application. The data user can then execute whatever analytics they want remotely, and is not able to copy the data back to their own machines. The data custodian provides a secure computing environment. The data user is limited in what kind of analysis tools are provided to them in this environment.

A variant of that is to limit the data user to being able to access the safe haven remotely from specific locations. For example, remote access may only be allowed from the library of the university. This is similar to the first type except that the locations where remote access is allowed has to be defined upfront.

The final variant is not to allow remote access at all, but to require the data user to come on-site and work on the data at the safe haven itself. On-site access requires constant monitoring and review of analysis output and individuals to ensure that they do not take data or printouts of data off the premises (i.e., manually enforce the "do not download data" requirement).

[3] This section is based on an article that appeared in Risky Business Magazine, 2014.

Gone Spear Phishing

Remote access to data has to be constructed to avoid phishing attacks. These are targeted attacks on the data users to steal their credentials (such as their username and password). If the data user is an academic, then the easiest way to do that is to construct a fake email that looks like it is coming from the CIO of the data custodian (e.g., by using their email signature and spoofing their email address). The email warns the data user that the custodian's security has been compromised and that it is necessary for everyone to change their passwords right away. Or the email can say that their password has expired and they need to renew it. The email would then include a link to change the password. The URL in the email looks like a link to an institutional web page but it really is going to another site that has been setup to look exactly like the data custodian. The user then types their old password and their new password. They get a fake thank you email and now their credentials to access the data have been compromised.

This is a very simple attack on data users to get their credentials and essentially give access to the data to someone else. To avoid this risk it is necessary to add another authentication mechanism beyond a username and password. For example, the user receives a text message with a randomly generated code each time they try to login.

Safe havens that limit the user to remotely access the data to specific locations and that require data users to come on-site are less prone to this type of attack because the user still has to go to a specific location where their identity can visually or manually be verified. This can be considered as another authentication mechanism.

And of course if the data user has to come on-site then phishing attacks are not a concern any more.

Screen Scraping

Another risk is that the data user will take screen shots of the data set or manually copy the data they view on the screen down. This is not practical for large number of individuals, but is easily doable for small numbers of individuals. This is a form of a deliberate attack on the data.

From a risk management perspective this is not desirable because this data can then be re-identified by the data user. The risk analysis for safe havens assume that the user is not going to download data, and such screen scraping defeats that.

One way to manage the risks from such a deliberate attack is to require the data user to sign a contract that prohibits them from copying raw data. The contract should also prohibit them from attempting to re-identify any individuals in the data and prohibits them from contacting any of the individuals in the data.

Another important control is to de-identify the data itself to make sure that any attempt at re-identification will have a high probability of failing. Sometimes we get asked "why is it necessary to de-identify the data if the user signs a contract not to attempt to re-identify ?". There are multiple reasons for this. First, it is important to have multiple layers of protection to manage risks. This is a basic security principle. Each layer reduces the risks to the organization by a certain amount, but never to zero. The cumulative risk reduction from multiple layers is significant and starts approaching a very small number. For example, a contract acts as a strong deterrent for a data user but the probability of a data user deliberately attempting to re-identify an individual is still not zero. If we *only* relied on the contract as our risk management technique then that means it is acceptable to share personal health information for secondary purposes without consent using only contracts as the control mechanism.

Inadvertent Re-identification

When a data user is accessing data they may still recognize someone they know in the data. For example, the data user may be viewing a raw data file in their favorite statistical analysis tool and recognize a date of birth or ZIP code of their ex-spouse. Curious, their eyes move down the data row to look at the other fields and more values match the person they know. Suddenly it is evident that the record belongs to that ex-spouse.

This is called “inadvertent re-identification”. Such a re-identification can occur when a data analyst accidentally recognizes someone, say a neighbor or a relative. Recognition can occur because the variable values match those for someone they know in real life or a famous person. It is not deliberate like in the situation considered above, but rather inadvertent.

The only way to manage this type of risk is through the de-identification of the data.

Concluding Points on Safe Havens

The use of safe havens to share data does not preclude the need to de-identify the data. The extent of de-identification may be different than what you would perform if data is actually released (it could potentially be less). Nevertheless, some amount of de-identification is necessary to manage two kinds of risks: the deliberate and the inadvertent re-identification risks.

Privacy Architectures for Big Data: Distributed Computation

A number of different schemes for distributed computation have been developed and used [8], [9]. The basic idea is that instead of collecting data and pooling it into one location, the analysis can be performed at the data source and then then the intermediate results would be combined to get a final analysis result that is the same or almost the same as if the data was pooled. By not pooling the data the expectation is that there would be a lesser privacy problem.

The challenge that we see is that a number of the systems that are implemented in practice are not actually secure and therefore do not address the privacy problem. This means that they can leak data which can potentially identify individual patients.

For example, systems that query remote databases and then pool the results are vulnerable to the “inference from statistical databases” problem (see [2]). A malicious data user can run multiple overlapping queries at each site to reconstruct the original data set. Another example are distributed regression modelling tools that can, under certain circumstances leak information about the raw data (see the analysis described in [8]).

Just because a protocol looks like it does not pool data does not mean that it is safe to use from a privacy and security perspective. More formal poofs are needed to demonstrate that these protocols and systems are not leaking information.

References

[1] K. El Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O’Reilly, 2013.

[2] Khaled El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

[3] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, “A Systematic Review of Re-identification Attacks on Health Data,” *PLoS ONE*, vol. 6, no. 12, 2011.

[4] W. G. van Panhuis, P. Paul, C. Emerson, J. Grefenstette, R. Wilder, A. J. Herbst, D. Heymann, and D. S. Burke, “A systematic review of barriers to data sharing in public health,” *BMC Public Health*, vol. 14, no. 1, p. 1144, Nov. 2014.

[5] K. El Emam, *Risky Business: Sharing Health Data while Protecting Privacy*. Trafford, 2013.

[6] K. El Emam and C. Alvarez, “A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques,” *International Data Privacy Law*, 2014.

[7] Article 29 Working Party, “Opinion 03/2014 on Purpose Limitation,” WP203, Apr. 2013.

[8] K. El Emam, S. Samet, L. Arbuckle, R. Tamblyn, C. Earle, and M. Kantarcioglu, “A secure distributed logistic regression protocol for the detection of rare adverse drug events,” *Journal of the American Medical Informatics Association*, Aug. 2012.

[9] K. El Emam, L. Arbuckle, A. Essex, S. Samet, B. Eze, G. Middleton, D. Buckeridge, E. Jonker, E. Moher, and C. Earle, “Secure Surveillance of Antimicrobial Resistant Organism Colonization or Infection in Ontario Long Term Care Homes,” *PLoS ONE*, vol. 9, no. 4, p. e93285, Apr. 2014.