# Risk-Based De-Identification of Health Data

Khaled El Emam
*Children's Hospital of Eastern Ontario*

Most privacy laws treat the identifiability of information as a binary construct: information is either identifiable (personal) or not. In reality, however, a continuum of identifiability exists, with some data sets being quite easy to re-identify with minimal resources and skill, and others requiring considerable time, effort, cost, and skill to re-identify.

A pragmatic way to reconcile these differences is to define a threshold on the identifiability continuum. If a data set's identifiability is above the threshold, then we consider it personal information; if it's below, then we no longer consider it as such. We might then be able to de-identify personal information by reducing its identifiability to a value below the threshold.

The important question, then, is how to define such an identifiability threshold—where should it be on the continuum? To answer this question, we must solve three problems: define the continuum, objectively measure identifiability, and use a decision rule to decide what threshold to use.

In this article, I describe how to measure identifiability and provide an overview of a threshold decision rule. Colleagues and I at the Children's Hospital of Eastern Ontario have used these approaches quite extensively in, for example, the disclosure of provincial cancer and birth registry data.

The simplified scenario I use here is that of a health data custodian who has received a request to disclose individual-level health data for secondary purposes. The custodian must decide if the data is personal information or not. If the custodian decides that the data is personal information, then it might be necessary to seek patient consent to disclose it, or, alternatively, the data must be de-identified before disclosure.

## The Identifiability Continuum

In a data set, we distinguish between direct identifiers and indirect identifiers (or quasi-identifiers). The former includes names and addresses, whereas the latter includes dates, locations, and socio-economic information. We can re-identify patients from either type of identifier.

Figure 1 defines the continuum of identifiability as comprising five discreet levels. This model has a few important characteristics. First, it aims to coincide with current practices—that is, it's a descriptive scale, rather than a prescriptive model. Second, the levels are cumulative in that each one inherits any de-identification applied at a lower level.

The scale is ordinal, meaning that we can't compare conceptual distances between levels. For example, the difference between levels 1 and 2 isn't the same as that between levels 4 and 5. Also, it doesn't make sense to take ratios, for example, by saying that moving from level 1 to level 3 represents twice the reduction in risk as moving from level 1 to level 2.

Level 1 pertains to data that's clearly identifiable—a database containing names, Social Security Numbers (SSNs), and doctor visit information about individuals would be level 1. At this level, minimal effort is needed to re-identify an individual. If we have people's names and addresses, we know who they are.

Masked data (level 2) manipulates the identifying variables in terms of, for example, randomization and creating reversible or irreversible pseudonyms[1,2] but does nothing to obfuscate the quasi-identifiers. Because these identifiers aren't modified in level 2 data, this is effectively still personal information. As an example, much of the data collected and disclosed in the context of clinical trials is masked data.

The difference between levels 2 and 3 is that, in the latter, the custodian attempts to obfuscate the quasi-identifiers as well as the identifying variables. However, at level 3, the custodian isn't measuring the data's identifiability and thus can't objectively determine whether it's personal information, even if he or she believes and behaves as if it isn't. For example, in a Canadian context, birth date and full postal code uniquely identify many Canadians living in urban areas, making that combi-

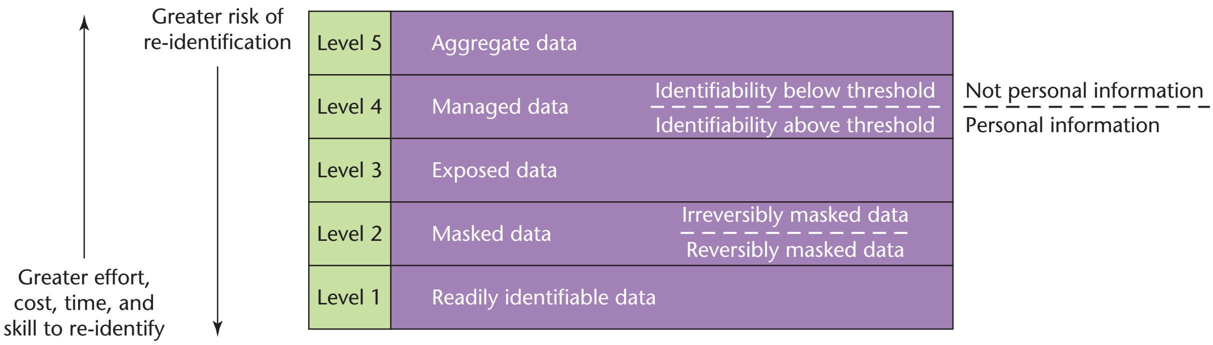

Figure 1. A conceptual five-level model of the identifiability continuum. The discreet levels characterize specific stages that a data set would go through as it's increasingly de-identified.

nation highly identifying. Reducing the postal code's precision to five characters from six doesn't actually reduce identifiability that much, but is quite common among custodians as a mechanism for ostensibly de-identifying data. Because the data's identifiability is uncertain, level 3 data represent a high risk exposure for the custodian.

Most of the data sets that have been re-identified to date have been level 2 data, and some level 3 data—for example, the Chicago homicide database[3] and the governor of Massachusetts cases.[4] These data can't be considered de-identified and therefore their re-identification isn't a surprise.

With level 4 data, the custodian objectively measures identifiability and can substantiate claims that it's above or below a specified threshold. Level 4 data can be microdata or appear in tabular form, and only at this level can data move from being personal information to not being personal information. Level 4 data is called *managed* because the data custodian can manage the risk of re-identification.

Level 5 pertains to information that clearly isn't identifiable. Aggregate data consists of unstratified counts, frequencies, or rates. For example, we would consider a table showing that 25 people have died from the H1N1 pandemic influenza in Canada in 2009 to be aggregate data. If the table is stratified by quasi-identifiers (for example, date, age, and gender), then it wouldn't be aggregate data anymore because such cross-tabulations can still have quite high identifiability.

So, to be able to make credible claims about information being personal or not, we must obfuscate the identifying variables (that is, perform masking) and be able to measure identifiability.

## Measuring Health Data's Identifiability

We generally assume that an adversary can re-identify one or more individuals from a data set if he or she has some background information about those individuals. Such an adversary presents two general risks. First, he or she could determine the identity associated with a record in the disclosed data set. This would occur, for example, if the adversary correctly determines that record number 7 pertaining to a 25-year-old male is actually about John Smith living on 401 Smyth Road. Second, the adversary could discover something new about individuals in the data set without knowing which records belong to them—for instance, if all 20-year-old males in the data set are HIV positive, and the adversary knows that John Smith is in the data set. In this case, the adversary discovered something new about an individual without knowing which record was about him. I focus on the first type of re-identification, called *identity disclosure*.

Identity disclosure presents three kinds of re-identification risk that we can measure objectively with specific probabilistic metrics.[5,6]

*Prosecutor risk* is relevant when the adversary is attempting to re-identify a specific (target) individual (for example, a famous person or a neighbor), has background knowledge on the individual, and knows that he or she is in the disclosed data. For example, if a data custodian is disclosing a provincial renal cancer registry, then everyone with kidney cancer will be in that data. If an adversary knows that John Smith had kidney cancer, then he or she will also know that John is in the registry, and this risk applies.

*Journalist risk* is relevant when the adversary is attempting to re-identify an individual in the disclosed data set but doesn't know with certainty whether this individual is actually included. For example, when the data set is based on a chart review of a random sample of patients in a hospital, or if a data custodian is creating a random sample from a population registry that will be made publicly available, then the journalist risk would apply because an adversary wouldn't know whether John Smith was in that sample.

This uncertainty is captured in the probabilistic measure of re-identification risk.

*Marketer risk* is relevant when the adversary is attempting to re-identify as many people as possible in the disclosed data (unlike the previous two metrics, in which the adversary is trying to re-identify a single individual). For example, if a community hospital is disclosing discharge abstract data, then a risk exists that an adversary would match adults in that data set with the county's voter list and re-identify multiple patients based on their demographics (for instance, date of birth and gender).

For prosecutor and journalist risk, we can assign a probability of re-identification to each individual in the disclosed data. The whole data set would have a risk equivalent to the maximum probability, so these metrics tend to be conservative. Marketer risk computes the expected number of individuals that would be re-identified, in which case, a data set where four out of 1,000 people would be expected to be re-identified has a lower marketer risk than one in which 500 people would be expected to be re-identified.

Data custodians must decide which metric represents a plausible attack scenario for their data sets. In some cases, more than one risk would apply. Numerically, prosecutor risk will be higher than journalist risk, which will be higher than marketer risk (as a proportion). So, a custodian would start from the top and must only manage the first plausible risk.

## Identifiability Threshold Decision Rule

Consider two quite different scenarios for a cancer registry: a researcher requesting the data and the creation of a public-use microdata file (PUMF).

An identifiability threshold for a PUMF would have to be extremely low because once a data set becomes public, we can't control who gets access to it or what the user will try to do with it in terms of re-identifying patients and maybe even contacting them. To protect such a cancer registry, it would have to go through extensive de-identification, which will reduce the data's quality.

The researcher would represent less risk to the data custodian because the researcher can sign a data-sharing agreement stipulating that he or she won't try to re-identify the data, can undergo regular third-party security audits and privacy impact assessments to ensure that he or she is implementing good practices for managing the data, and can be compelled to destroy the data once the research project is over. With all these provisions in place, it would seem sensible to accept a higher identifiability threshold for the researcher because he or she would represent a lower risk to the data custodian. With a higher identifiability threshold, the researcher would also get higher-quality data with fewer distortions due to de-identification.

Having multiple thresholds is consistent with, for example, the US's Health Insurance Portability and Accountability Act's (HIPAA's) Privacy Rule. In this case, the Safe Harbor provision assumes a low threshold for a PUMF disclosure, whereas the Limited Data Set provision assumes a higher threshold for a researcher disclosure scenario.

In practice, more than two disclosure scenarios will exist, so we need a general decision rule for selecting a threshold based on a set of criteria.[7] Data custodians have used a set of criteria informally for at least the past 15 years, and recent research by my team has formalized them; they cover three dimensions.[8]

*Mitigating controls* is the set of security and privacy practices that the data requestor has in place. A recent review has identified a union of practices that large data custodians use and funding agencies and research ethics boards have recommended for managing sensitive health information.[9]

*Invasion of privacy* evaluates the extent to which a particular disclosure would invade patients' privacy (a checklist is available elsewhere[8]) and looks at three considerations:

- the data's sensitivity—the greater its sensitivity, the greater the invasion of privacy;
- the potential injury to patients from an inappropriate disclosure—the greater the potential for injury the greater the invasion of privacy; and
- the appropriateness of the consent obtained for disclosing the data, if any—the less appropriate the consent the greater the invasion of privacy.

*Motives and capacity* considers the data requestor's motives and capacity to re-identify the data, examining issues such as conflicts of interest, the potential for financial gain from a re-identification, and whether the requestor has the skills and financial capacity to re-identify the data (a checklist is available elsewhere[8]).

For example, if we again consider our typical health researcher requesting the renal cancer registry, he or she would likely have some basic security practices in place, the invasion-of-privacy risk would be relatively high, and the researcher wouldn't have the motive to actively re-identify individuals nor the capacity to do so. In addition, he or she would sign the data-sharing agreement. Contrast that with a PUMF, in which we'd assume no mitigating controls existed, and the motives and capacity risks are high because anyone would have access to the data. By considering the three dimensions just mentioned, a data custodian would conclude that the disclosure to the researcher was lower risk and would therefore set a higher identifiability threshold than for disclosing a PUMF.

By considering the risks for each data disclosure, the custodian can assign appropriate thresholds to manage these risks each time. If a data set has identifiability higher than the threshold, then it would be de-identified before disclosure. The custodian will perform less de-identification in lower-risk situations.

After applying this risk-based approach in practice, several things have become evident:

- It provides an incentive for data requestors to improve their security and privacy practices (the mitigating controls dimension) because it results in higher thresholds and hence less distortion to the data they receive.
- It ensures that the amount of de-identification done to the data—and commensurate distortions of the data—is proportionate to the actual disclosure risks.
- It improves the relationship between the data custodian and requestor because the trade-offs are explicit and negotiable.

Therefore, rather than controlling data's identifiability, data custodians should manage the risk of re-identification and be judged on how well they can do so.

## References

1. *Health Informatics: Pseudonymization*, ISO/TS 25237, Int'l Organization for Standardization, 2008.
2. K. El Emam and A. Fineberg, *An Overview of Techniques for De-Identifying Personal Health Information*, tech. report, funded by the Access to Information and Privacy Division of Health Canada, 2009.
3. S. Ochoa et al., *Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study*, tech. report, Massachusetts Inst. of Technology, 2001.
4. L. Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection*, Massachusetts Inst. of Technology, 2001.
5. K. El Emam and F. Dankar, "Protecting Privacy Using k-Anonymitym," *J. Am. Medical Informatics Assoc.*, vol. 15, no. 5, 2008, pp. 627–637.
6. F. Dankar and K. El Emam, "A Method for Evaluating Marketer Re-Identification Risk," *Workshop on Privacy and Anonymity in the Information Society*, 2010; http://portal.acm.org/citation.cfm?id=1754239.1754271.
7. K. El Emam, "Method and Experiences of Risk-Based De-Identification of Health Information," *Proc. Workshop on the HIPAA Privacy Rule's De-Identification Standard*, Dept. of Health and Human Services, 2010; www.hhshipaaprivacy.com.
8. K. El Emam et al. "A Method for Managing Re-Identification Risk from Small Geographic Areas in Canada," *BMC Medical Informatics and Decision Making*, vol. 10, no. 18, 2010; www.biomedcentral.com/1472-6947/10/18.
9. K. El Emam et al., "Evaluating Patient Re-Identification Risk from Hospital Prescription Records," *Canadian J. Hospital Pharmacy*, vol. 62, no. 4, 2009, pp. 307–319.

***Khaled El Emam** is a senior scientist at the Children's Hospital of Eastern Ontario Research Institute, and is a Canada research chair and associate professor in the Faculty of Medicine at the University of Ottawa. His research interests include re-identification risk assessment and developing practical de-identification techniques for health information. El Emam has a PhD in electrical and electronic engineering from King's College, University of London. Contact him at kelemam@uottawa.ca; www.ehealthinformation.ca.*