Heuristics for De-identifying Health Data

Khaled El Emam
Children's Hospital of Eastern Ontario Research Institute

Demands for secondary use of personal health information (PHI) are increasing. Disease registries and registries to track patients for monitoring drug safety, for example, often capture data already collected for providing care. Secondary uses include academic or commercial research, public health, and policy making. Privacy officers must ensure that PHI is appropriately de-identified before disclosing it for such purposes.

Here I review six de-identification heuristics that privacy analysts commonly use. These rules of thumb are precise, easy to understand, and require no statistical expertise to apply. This simplicity is perhaps one of the key reasons for their frequent use.

Common heuristics

All de-identification heuristics apply to a set of indirectly identifying variables: *quasi-identifiers*. These variables, which might be publicly known, make individuals more unique in the population or define subpopulations that an intruder can target. Quasi-identifiers include age, gender, race, ethnicity, dates (birth, death, admission, discharge), and place of residence. The heuristics' labels reflect a healthcare-specific application[1] and might have different names in other domains.

The data in Table 1 illustrates the various heuristics. It includes hypothetical data describing the cause of death for four census divisions (CDs) in Ontario, Canada. The quasi-identifiers for the data are CD, cause of death, race, gender, and age range.

The *limited data set* heuristic suppresses certain variables from the data set to de-identify it. The best-known limited data set heuristic is the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor list. This heuristic provides a precise set of items to exclude from a data set. According to this list, a privacy officer could disclose the data set in Table 1 because it would be considered de-identified.

The *uniqueness rule* applies to individuals who are unique in a population. These individuals are considered to be at a higher risk of re-identification. Researchers have estimated that between 63 and 87 percent of the US population is unique in their basic demographics. One set of thresholds used in cancer registries suggests that having up to 20 percent unique individuals in the population based on the quasi-identifiers is acceptable if the data will be disclosed to a trusted researcher, but only 5 percent is acceptable if the data will be disclosed to the public. The US Census Bureau deemed uniqueness approaching 10 percent suitable for publicly released data sets. For example, the unique individual in column B of Table 1 would be suppressed.

The *rareness rule* applies to individuals who are rare on some visible and traceable characteristics and thus are easier to re-identify. For example, a journalist or other motivated person might notice an individual with a rare and visible disease, such as some types of endocrine disorders and tumors. Definitions of a rare disease vary from an incidence rate of 0.5 to 0.1 percent. Top and bottom coding are special cases of the rareness rule in that they affect extreme values. National statistical agencies often top-code age at 89+ using a cutoff of 0.5 percent incidence, because old age is also visible. In my example, if we use the 0.5 percent cutoff for the data in Table 1, any deaths of black males within that age range would be considered sufficiently rare within a CD; thus, no data would be disclosed at all.

The *numerator rule* is sometimes known as the "rule of 5" or "rule of 3." It implies that each possible set of values on the quasi-identifiers (also known as an equivalence class) has at least five, or three, records. This heuristic is only useful if a particular individual's presence in the data set is known with certainty. For example, (almost) everyone will have a record in a population registry by definition. However, it's unlikely that it would be known which individuals are in a small random sample. In Table 1, the five B records in the first and third CDs would be suppressed.

Table 1. Example used to illustrate the various heuristics.*

DATA			APPLICATION OF HEURISTICS			
A CENSUS DIVISION POPULATION	B AIDS DEATHS, BLACK MALES, AGE 25–44	C TOTAL DEATHS, BLACK MALES, AGE 25–44	NUMERATOR RULE B < 6	DENOMINATOR RULE 1 A < 100,000	DENOMINATOR RULE 2 C < 30	MISSOURI RULE C – B < 10
75,000	1	100	Suppress	Suppress	Release	Release
60,000	95	100	Release	Suppress	Release	Suppress
150,000	4	8	Suppress	Release	Suppress	Suppress
120,000	6	7	Release	Release	Suppress	Suppress

*The hypothetical data consists of counts of individuals that match the criteria in the columns. (I assume that only record suppression will be used.)

Another version of this rule is that at most 20 percent of the records are in equivalence classes of size five or less.[2]

The *denominator rule* assumes that the data's end-user, say an epidemiologist, will compute some proportion or rate from the data (for example, AIDS death rate). Multiple versions of this rule exist because epidemiologists can use different denominators, such as the underlying population's size. For example, in the HIPAA Safe Harbor List, data custodians can release the first three digits of the zip code only if the zip code region contains more than 20,000 residents. The Census Bureau and Statistics Canada use similar population cutoffs—100,000 residents for the former and 70,000 residents (for the Canadian Community Health Survey). The underlying assumption is that in a smaller population, individuals might be sufficiently rare or unique that they are more easily re-identifiable. Using the 100,000-resident threshold on Table 1, we'd suppress all records from the first and second CDs. Another possible denominator is the number of deaths within the specific subpopulation. If we use a threshold of 30, this heuristic would suppress the 15 C records from the last two CDs in Table 1.

The *Missouri rule* combines the numerator and denominator rules by subtracting the numerator from the denominator. If the difference is less than 10, the records can't be disclosed. In Table 1, this heuristic would suppress the C records from the last three CDs.

Data custodians sometimes apply minor variations on the above heuristics, so this isn't a comprehensive list. In particular, privacy officers and research ethics boards often have the discretion to approve variations.

Applying heuristics

You can apply some of the heuristics before any data is collected. You might want to ensure that the researcher, say, has collected the data anonymously or specify how he or she will perform the de-identification a priori, before collecting any data. This would be the case when developing a study protocol or negotiating a data-sharing agreement.

You can apply the *limited data set* rule when preparing a study protocol. This rule stipulates precisely which variables you must exclude to ensure that the data set is de-identified. An organization can also apply some versions of the denominator rule early in a data-collection effort. For example, you can check the most recent census results to determine which CDs have fewer than 100,000 individuals, say, and exclude these CDs from the data collection or merge adjacent CDs and record only the larger entity.

Despite their apparent simplicity, applying these rules to de-identify health data sets is quite challenging, for several reasons.

The heuristics don't directly consider the distortion to the data (that is, information loss). Experiences applying these rules suggest that the extent of information loss can be severe, resulting in de-identified data sets that are no longer suitable for answering important health questions or informing critical policy decisions.

In practice, the most common techniques data custodians use to implement these heuristics are record suppression, variable suppression, simple generalization (for example, changing age to 5- or 10-year ranges), and geographic area aggregation. They can apply these techniques manually or as part of optimization algorithms (for example, *k*-anonymity algorithms). However, organizations rarely apply optimization algorithms to select the best technique or combination of techniques to minimize information loss. Therefore, they can never be quite sure whether the de-identification technique used to satisfy the heuristic is the best one to use.

As Table 1 illustrates, using alternative rules can produce quite different answers. Because of this lack of consistency, the results of de-identification depend on the subset of heuristics an organization applies. The lack of clarity in how the rules relate

to each other can be disconcerting for practitioners.

In addition, few of these simple heuristics account for external risk factors, such as the data recipients' characteristics. For example, if you were creating a public-use data file, the re-identification risks would be considered high. If you disclose the data set to a trusted party who will sign a data-sharing agreement and consent to audits by the custodian, the risk is considered lower. You should perform less de-identification in the latter case because the recipient is more trusted. Some exceptions exist. For example, HIPAA has two versions of the limited data set for different levels of recipient risk.

When deciding on the appropriate thresholds to use, data custodians must also account for the potential of harm to patients should inappropriate data disclosure occur. With everything else being equal (for example, the data recipient), a data set with information on stigmatized health conditions would cause more harm if inappropriately disclosed, and should therefore undergo more aggressive de-identification. For example, the numerator rule should have a value higher than 5 if the data include ICD-10 codes for stigmatized conditions.

The type and extent of de-identification performed must depend on the number and nature of the quasi-identifiers in question. For example, the denominator rule, in which the denominator is the whole population, is applicable whether the data set has only one variable—say, gender—or 20 quasi-identifiers. This can result in counterintuitive de-identification practices.

When deciding on the appropriate thresholds to use, data custodians must also account for the potential of harm to patients should inappropriate data disclosure occur.

De-identification as risk management

Data custodians should view de-identification as a risk-management exercise. This means that the objective isn't minimizing re-identification risk to its lowest possible level, but choosing a risk that the custodian is willing to take for a particular disclosure, or type of disclosures, and managing it. Managing risk means making trade-offs among de-identification, procedural constraints, and contractual obligations. The highest risk would be a publicly accessible data file (for example, downloadable over the Internet or through an online query system or data released through an access-to-information request). In this case, the custodian has no control over who has access to the data and what they'll do with it: there are no procedural or contractual levers. Such a disclosure warrants stringent de-identification. On the other hand, when providing data to an epidemiologist in an academic institution, the custodian can reduce the extent of de-identification by, for example, also having a data-sharing agreement and regularly auditing the institution's record-management practices.

Implementing a meaningful risk-management regime requires several actions. First, data custodians must clearly define what constitutes a quasi-identifier. This will let them select the variables to which they need to apply the rules I've presented. Some quasi-identifiers might not be obvious. For example, in Canada, intruders could use information about a patient's profession for re-identification because some public registries can be linked together to construct personal profiles of members of certain professions.[3] Custodians should also consider variables that can be used to infer quasi-identifiers. For example, intruders can use profession-specific third-party payers in insurance claims (such as in the case of some civil servants and teachers) to infer profession, autopsy date to infer date of death, billing codes indicating birthing procedures to infer date of birth, diagnostic codes and lab tests (for example, mammograms and PSA tests) to infer gender and age range, and graduation date to infer age.

Custodians must also establish concrete guidelines for setting thresholds. These thresholds should depend on the situation's specifics. For example, they'd need to consider potential intruders' characteristics (such as motivation and availability of resources), what controls can be put in place, and the potential for harm to the patients if an inadvertent disclosure occurred. Even three thresholds for different levels of high/medium/low risk for internal data users, trusted external data users, and the public would be a good start.

Many of the rules depend on knowing the population's distribution on the quasi-identifiers within geographic areas. This kind of census data is rarely readily available at a sufficient level of detail, especially for the smaller geographic areas. Summary information about the population is needed to facilitate better disclosure control decisions.

Finally, data custodians should use more optimization techniques. Where appropriate, for example, as with the numerator rule, data custodians should consider the use of readily available optimization algorithms that will balance information loss with re-identification risk.[4]

Current privacy practices for secondary use of PHI leave custodians uncomfortable as to

whether they've exercised sufficient due diligence to protect patient anonymity, and users frustrated at the extent to which data has been suppressed and perturbed before they see it. We therefore can't underestimate the benefit of having usable de-identification heuristics that are acceptable to both data custodians and data users. ☐

References

1. B. Rudolph et al., "Small Numbers, Disclosure Risk, Security, and Reliability Issues in Web-Based Data Query Systems," *J. Public Health Management Practice*, vol. 12, no. 2, 2006, pp. 176–183.
2. H. Howe, A. Lake, and T. Shen, "Method to Assess Identifiability in Electronic Data Files," *Am. J. Epidemiology*, vol. 165, no. 5, 2007, pp. 597–601.
3. K. El Emam et al., "Evaluating Common De-Identification Heuristics for Personal Health Information," *J. Medical Internet Research*, vol. 8, no. 4, 2006, p. e28.
4. K. El Emam and F. Dankar, "Protecting Privacy Using *k*-Anonymity," *J. Am. Medical Informatics Assoc.*, Sept./Oct. 2008, to appear.

Khaled El Emam *is a senior scientist at the Children's Hospital of Eastern Ontario Research Institute and is a Canada Research Chair and an associate professor in the Faculty of Medicine, 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.*

Interested in writing for this department? Please contact editors E. Michael Power (michael.power@ssha.on.ca) or Roland L. Trope (roland.trope@verizon.net).