GUEST EDITORIAL PREFACE

ICD-10-CM and the Risk of Re-Identification

Liam O'Neill, School of Public Health, University of North Texas Health Science Center, Fort Worth, TX, USA

Barring another false start or last-minute postponement, the United States will finally make the switch to ICD-10-CM on October 1, 2015. (International Classification of Diseases, 10th Revision, Clinical Modification). Overnight, the number of diagnosis and procedure codes will increase almost ten-fold, from about 17,800 under the current system (ICD-9-CM) to more than 141,000 (Meyer, 2011). Many have touted the potential benefits of the new system for such functions as fraud detection, patient safety, and health services research. Yet few have considered the impact that the switch could have on the privacy of patient's health information.

Because the information contained in ICD-10-CM is highly specific, this could increase the risk of re-identification. Moreover, because the medical record often contains one's most personal secrets, its disclosure could cause significant harm. For example, both ICD-9-CM and ICD-10-CM contain multiple codes for high-risk sexual behavior and abortion. Yet ICD-10-CM provides additional detail, such as sexual orientation and abortion trimester. Thus, an "anonymized" hospital database could be used (or misused) to determine the zip codes with the highest rate of third-trimester abortions or high-risk sexual activity.

To illustrate the ease with which an individual hospital record could be re-identified, consider the following scenario. Suppose an adversary knew of a co-worker who fractured his clavicle in a bicycle accident that required an overnight hospital stay. Starting with a database of 500,000 hospital records, the adversary identified 76 records that contained any of the 300 ICD-10-CM codes for bicycle accidents. Of these, seven involved a fracture of the left clavicle and only one record was from a local hospital. Further confirmation was provided by matching the patient's age, gender, race, and zip code. The record contained other information from the patient's medical record unrelated to the reason for admission. In this case, the adversary discovers his co-worker had been treated for depression, opioid addiction, and is HIV-positive. Knowing only a few facts about his co-worker, the adversary was able to narrow down the relevant hospital records search to a sub-group of size one.

The above example is not just hypothetical. Sweeney and colleagues (2013) showed how newspaper accounts of various accidents and assaults may contain sufficient detail to identify individuals from a published database of hospital records. They were able to re-identify thirty-five individuals from the Washington state hospital database, which they obtained for $50. The ensuing publicity quickly led to the passage of a new state regulation to make Washington's hospital databases more secure.

More than 60 percent of ICD-10-CM codes pertain to injuries, and some of these have only marginal clinical relevance (Chute, Huff, Ferguson, Walker, & Halamka, 2012). A marketing campaign by a consulting company highlighted some of the more memorable of these codes in a series of parody videos. Among the codes included were: sucked into a jet engine (V97.33XA), burn due to jet ski on fire (V91.07XA), and struck by orca (W56.22XA). These new codes may offer significant public health benefits, such as for improving aviation or boating safety. However, as these events are exceedingly rare, many such codes are likely to pertain to one or fewer hospital records. A widely used anonymization framework is called "k-anonymity" (Sweeney, 2002). To achieve a given level of security, records cannot be broken down into sub-groups with fewer than k members. Hence a sub-group with only one member has a corresponding re-identification risk of (1/k), i.e., one-hundred percent.

The Federally-mandated policy for de-identifying health information is known as the HIPAA "safe harbor" method, as defined in the HIPAA privacy rule. In order to meet this standard, all "Protected Health Information" (PHI) must be removed from the database. PHI consists of eighteen attributes, such as name, phone number, and social security number (Zhang, O'Neill, Das, Cheng, & Huang, 2012). Health information that meets the safe harbor standard is neither tracked nor regulated, a practice known as "release-and-forget" (Ohm, 2010). While state agencies are the primary source of hospital databases, state data collection is exempt from HIPAA. Thus privacy practices can vary widely across states, and few state databases actually meet the safe harbor standard.

The assumption that underlies the HIPAA standard is that information that is not "personally identifiable" cannot be used to re-identify individuals. Numerous studies have revealed the holes in this argument by demonstrating that re-identification is possible under certain circumstances from "anonymized" databases (Narayanan & Shmatikov, 2010; Ohm, 2010). Since 1997, when Latanya Sweeney first identified the Governor of Massachusetts from public health records, re-identification techniques have grown in sophistication and have raised serious questions about the wisdom of the current practice of "release and forget" (Sweeney, 1997).

Some states have also used various trigger conditions to suppress certain information, while leaving the rest of the database intact. (For example, for zip codes with less than 30 patients, truncate the last two digits.) From the researcher's perspective, this is analogous to buying a jigsaw puzzle and receiving 87 out of 100 pieces. Moreover, some of the missing pieces have little to do with patient privacy, but may instead reflect political priorities, such as by redacting physician or hospital identifiers.

In testing the anonymization method used by one state agency, we showed that PHI, such as patient zip codes, could be derived from non-PHI through disclosure channels, such as data dependency and domain knowledge (Zhang, O'Neill, et al., 2012). For example, county can be used to determine zip code, and Diagnosis Related Group (DRG) can reveal the patient's age and gender. Other PHI, such as date of admission, could be gleaned from codes for natural disasters, such as floods, tornadoes, and hurricanes.

As some experts have noted, for a single record selected at random from a properly de-identified database, the risk of re-identification is quite small (El Emam, Jonker, Arbuckle, & Malin, 2011). Yet even a method that is 99.99 percent effective is not secure enough for databases containing millions of records. For the moment, ICD-9-CM and ICD-10-CM codes are not considered PHI. This may change in the future, however, because what is considered "personally identifiable" is an ever-increasing category. Some data elements that were once thought to be sufficiently generic (e.g., zip code and date of birth) have since been re-classified as PHI. Any future changes in privacy protocols will unable to protect databases that have already been released.

What is needed is a more comprehensive approach that considers both the risk of re-identification and the likely harm if a particular ICD-10-CM code was associated with an individual. While the majority of diagnosis codes are likely to be of minor import, other codes could cause significant harm to individuals, such as those related to sexuality, mental health, abortion, and HIV status. Yet there is currently no multi-stage model of privacy (e.g., "confidential," "highly confidential,") for ICD-9-CM or ICD-10-CM codes. Yet such a system of classifications and clearances has been used for years (e.g., by the NSA) to protect state secrets.

While the risk of re-identification can never be reduced to zero, there are ways to manage and mitigate such risks. For example, Decision Support Systems are currently in development that will automatically perform a "privacy attack" on a given health care database by identifying data dependencies and unbalanced suppressions as points of vulnerability. This will provide immediate feedback to hospital managers and state officials on the effectiveness of their suppression rules.

During the 1990s and early 2000s, while EHR adoption remained low, the HIPAA law was barely adequate to protect patient privacy. In recent years, however, an infusion of government spending and private capital has re-shaped the health IT landscape, while improving the quality, quantity, and accessibility of "big data" resources. Yet this rapid progress has also exposed the weaknesses of a HIPAA standard based on yesterday's technology.

Hospitals have spent millions to comply with the HIPAA privacy rule, and millions more will be spent on the transition to ICD-10-CM. Yet scant funds have gone into rigorous testing of the HIPAA standard itself at both the Federal and state level. Moreover, hospitals have few legal or economic incentives to go far beyond what is required by HIPAA. A more proactive approach is needed that employs the latest re-identification techniques from computer science. This could greatly reduce the risk of re-identification, while preserving the utility of the data for research and other legitimate purposes.

Liam O'Neill
Guest Editor
IJHISI

REFERENCES

Chute, C. G., Huff, S. M., Ferguson, J. A., Walker, J. M., & Halamka, J. D. (2012). There are important reasons for delaying implementation of the new ICD-10 coding system. *Health Affairs (Project Hope)*, *31*(4), 836–842. doi:10.1377/hlthaff.2011.1258 PMID:22442180

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A Systematic Review of Re-Identification Attacks on Health Data. *PLoS ONE*, *6*(12), e28071. doi:10.1371/journal.pone.0028071 PMID:22164229

Meyer, H. (2011). Coding complexity: US health care gets ready for the coming Of ICD-10. *Health Affairs (Project Hope)*, *30*(5), 968–974. doi:10.1377/hlthaff.2011.0319 PMID:21555481

Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of 'personally identifiable information.'. *Communications of the ACM*, *53*(6), 24–26. doi:10.1145/1743546.1743558

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review. University of California, Los Angeles. School of Law*, *57*, 1701–1777.

Sweeney, L. (1997). Guaranteeing Anonymity when Sharing Medical Data, the Datafly System. In D. Masys (Ed.), *Proceedings, American Medical Informatics Association* (pp. 51–55). Nashville, TN: Hanley & Belfus, Inc.

Sweeney, L. (2002). K-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, *10*(5), 557–570. doi:10.1142/S0218488502001648

Sweeney, L. (2013). Matching Known Patients to Health Records in Washington State Data. Harvard University. Data Privacy Lab. http://thedatamap.org/1089-1.pdf

Zhang, N., O'Neill, L., Das, G., Cheng, X., & Huang, H. (2012). No Silver Bullet: Identifying Security Vulnerabilities In Anonymization Protocols for Hospital Databases. *International Journal of Healthcare Information Systems and Informatics*, *7*(4), 48–58. doi:10.4018/jhisi.2012100104