The MVA Attack:
Re-Identification Risk in HIPAA Safe Harbor De-Identified Datasets

Victor A. Janmey, Daniel R. Schlegel, Christopher Crowner, Peter L. Elkin, MD
Department of Biomedical Informatics, University at Buffalo, SUNY, Buffalo, NY, USA

Abstract

We present a re-identification attack, the MVA attack, that uses indirect (non-HIPAA) identifiers to target a vulnerable subset of records de-identified to the HIPAA Safe Harbor standard, those involving motor vehicle accidents (MVAs). The attack is demonstrated through a case report involving re-identification of a patient from a de-identified dataset. Remediation strategies to prevent this type of attack are also discussed.

Introduction

The MVA attack is a method of re-identifying free-text patient records de-identified to the HIPAA Safe Harbor standard that utilizes the presence of indirect identifiers, data elements that can be used individually or in combination with other identifiers (direct/HIPAA or indirect) along with outside resources to re-identify a record. It requires that the records contain references to motor vehicle accidents (MVAs) involving the patient or other HIPAA covered entities (e.g. relatives). Re-identification occurs when one or more direct identifiers, obtained from an outside resource like the voter registry used in Sweeney's re-identification attack, are associated with a patient record[1].

Prior Work

Sweeney's seminal 2002 re-identification attack was performed on pre-HIPAA de-identified data and used demographic data such as ZIP codes from the Cambridge, MA voter registry to re-identify individuals from records that contained ZIP codes[1]. Recognizing the risks of such an attack but also the need for records to be easily de-identified, the HIPAA Safe Harbor standard was designed by the Department of Health and Human Services to create an "easy to follow, 'cookbook' approach to de-identification", that involved removing 18 identifiers, including ZIP code (except, in general, the first 3 digits)[2]. HIPAA Safe Harbor de-identification has generally been considered to be effective, as a study by Kwok *et al*. in 2010 showed only 2 of 15,000 individuals could be re-identified from a dataset de-identified to the HIPAA Safe Harbor standard[3]. Of 6 re-identification attacks on healthcare data discussed by El Emam *et al.* in a 2011 review, only the prior study was on HIPAA de-identified data[4].

MVA Attack Demonstration/Case Report

An individual HIPAA Safe Harbor de-identified free-text patient record containing the HPI section exported from an EMR was selected by searching all such records for the keyword "MVA." The following indirect identifiers were extracted from the record: year of encounter, location of encounter (first 3 digits of ZIP), patient age, patient gender, year of accident, details of accident (e.g. injury sustained). As the first 3 digits of ZIP were "142," indicating the Buffalo area, the Buffalo News archives for the year of accident were Boolean searched for "(accident or crash or collision) not (plane or train or boat or ski or pilot or flight or craft or stock)" to yield 477 results. These were narrowed down by age of victim (adjusted for difference between encounter and accident years), gender of victim, and details of accident until a unique match was made with an article that contained the full name of the accident victim.

Remediation

The key indirect identifiers for the MVA attack are those that pertain to the year of accident and location of accident. Scrubbing all of these data elements, however, would potentially cause the loss of a significant amount of valuable clinical information. Instead, we recommend scrubbing all identifiers that could give the location of accident, as these are less likely to be useful clinically than the year of accident/the record year (given that these can give information about available treatments and treatment guidelines available at the time of accident, for example). With these location identifiers scrubbed, the search space for the attack becomes "all MVAs resulting in injury in the U.S. for a given year." Given that in 2007 about 3.2 million people sustained nonfatal injuries that resulted from traffic accidents, this search space is sufficiently large to mitigate the risk of re-identification[5].

References

1. L Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–70, 2002.

2. D McGraw. Building public trust in uses of health insurance portability and accountability act de-identified data. *J Am Med Inform Assoc*, 20(1):29–34, 2013.
3. P Kwok and D Lafky. Harder than you think: A case study of re-identification risk of hipaa-compliant records.
4. K El Emam, E Jonker, L Arbuckle, and B Malin. A systematic review of re-identification attacks on health data. PLoS One, 6(12):e28071, 2011.
5. L Zhao, J Lucado, and C Stocks. Emergency department visits associated with motor vehicle accidents, 2006.