No Silver Bullet: 
Identifying Security Vulnerabilities in Anonymization Protocols for Hospital Databases

Nan Zhang, Department of Computer Science, George Washington University, Washington, DC, USA

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

Gautam Das, Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA

Xiuzhen Cheng, Department of Computer Science, George Washington University, Washington, DC, USA

Heng Huang, Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA

ABSTRACT

In accordance with HIPAA regulations, patients’ personal information is typically removed or generalized prior to being released as public data files. However, it is not known if the standard method of de-identification is sufficient to prevent re-identification by an intruder. The authors conducted analytical processing to identify security vulnerabilities in the protocols to de-identify hospital data. Their techniques for discovering privacy leakage utilized three disclosure channels: (1) data inter-dependency, (2) biomedical domain knowledge, and (3) suppression algorithms and partial suppression results. One state’s inpatient discharge data set was used to represent the current practice of de-identification of health care data, where a systematic approach had been employed to suppress certain elements of the patient’s record. Of the 1,098 records for which the hospital ID was suppressed, the original hospital ID was recovered for 616 records, leading to a nullification rate of 56.1%. Utilizing domain knowledge based on the patient’s Diagnosis Related Group (DRG) code, the authors recovered the real age of 64 patients, the gender of 83 male patients and 713 female patients. They also successfully identified the ZIP code of 1,219 patients. The procedure used to de-identify hospital records was found to be inadequate to prevent disclosure of patient information. As the masking procedure described was found to be reversible, this increases the risk that an intruder could use this information to re-identify individual patients.

Keywords: Data Analytic Processing, Data Privacy, Diagnosis Related Group (DRG), Health Insurance Portability and Accountability Act (HIPAA), Re-Identification

DOI: 10.4018/jhisi.2012100104
1. INTRODUCTION

The health care sector has made significant progress over the last decade toward securing the privacy and confidentiality of patient data. Yet due to a number of factors, the issue of patient data security has once again been moved to the front burner. With the passage of the Stimulus Bill in 2009 and the Affordable Care Act of 2010, significant public funds have been dedicated to increase adoption of Electronic Health Records (EHRs). As EHRs become more widespread, health care data have become less costly, more accessible, and with improved clinical detail. Yet the proliferation of these databases has posed additional risks for consumers. Health care data contains personal information that could significantly harm patients if it were used improperly, such as in hiring decisions or to deny health insurance coverage.

The standard protocol to de-identify health data is known as the “safe harbor standard,” as defined by the Health Insurance Portability and Accountability Act (HIPAA) (El Emam, Jonker, Arbuckle, & Malin, 2011). To comply with the standard, eighteen data elements must be removed or generalized (Table 1). Personally identifying information (PII) are attributes that can uniquely identify an individual, such as name or social security number. Quasi-identifiers, such as zip code and birth date, can be used to link the anonymized dataset to other datasets. Once the data have been properly de-identified, the risk of re-identification is thought to be minimal. The safe harbor standard has also been selectively adopted in other countries, such as Canada.

There have been numerous high-profile incidents in which individuals have been re-identified based on weak “release-and-forget” anonymization protocols. In 2006, AOL released the web search history of 650,000 users over a three-month period. Some AOL customers could be uniquely identified based on their web search histories, resulting in a class action lawsuit and a public relations disaster (Barbaro & Zeller 2006). In another case, Sweeney demonstrated how to re-identify an individual (e.g., the governor of Massachusetts) by cross linking the date of birth, gender, and zip code information in a published patients’ data set with the voter registry of Cambridge, Massachusetts (Sweeney, 2000, 2002). The results show that birth date alone can uniquely identify the name and address of 12% of records, with a combination of birth date and gender up to 29%, birth date and 5-digit ZIP code up to 69%, and full postal code and birth date up to 97%. Critics argue that many companies’ privacy policy is based on the mistaken assumption that “personally identifiable” information is a fixed set of attributes that, once removed, effectively “inoculate” the data against re-identification attacks (Narayanan & Shmatikov, 2010). Given the rapid increase in the amount of publicly available data about individuals, they argue that

<table>
<thead>
<tr>
<th>Personally Identifiable Information (PII),</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Name; 2) Geographic information except state, subject to restrictions</td>
</tr>
<tr>
<td>3) Any dates, year allowed. e.g., Birthdate, Admit Date; 4) Phone #; 5) Fax #;</td>
</tr>
<tr>
<td>6) E-mail address; 7) Social Security Number; 8) Medical record #; 9) Insurance #;</td>
</tr>
<tr>
<td>15) Internet Address; 16) Biometric ID; 17) Full face photos; 18) Any other unique ID #</td>
</tr>
</tbody>
</table>

Table 1. These 18 elements that must be removed or generalized according to the HIPAA Privacy Rule, Safe Harbor Standard
E-patients Empower Healthcare: Discovery of Adverse Events in Online Communities
[www.igi-global.com/chapter/patients-empower-healthcare/46680?camid=4v1a](www.igi-global.com/chapter/patients-empower-healthcare/46680?camid=4v1a)