

What De-Identified Data Costs You in Testing and How the Dynamic Data Generator™ Can Help

Michael Matteson

Abstract

De-identified data destroys your ability to do high-quality performance testing of your health care data system. The robust and consistent data provided by the Patent-Pending Dynamic Data Generator™ can free you from security problems while allowing you to more quickly, cost-effectively and thoroughly test your system.

De-Identification Process

To comply with the HIPPA Privacy Act, vendors need to use what is termed as a HIPPA *De-identified* data set or manually create data sets. The de-identified datasets are datasets derived from real patient data where the data is then manipulated as to lessen the risk of re-identification. The government in 1996 outlined two approaches to accomplish de-identification. The two methods are as follows:

1. The “safe-harbor” method, requires the removal of every one of 18 identifiers enumerated at section 164.514(b)(2) of the Privacy Rule. [See Table 1] Data that are stripped of these 18 identifiers are regarded as de-identified, unless the covered entity has actual knowledge that it would be possible to use the remaining information alone or in combination with other information to identify the subject.
2. To have a qualified statistician determine, using generally accepted statistical and scientific principles and methods, that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by the anticipated recipient to identify the subject of the information. The qualified statistician must document the methods and results of the analysis that justify such a determination.

The “safe-harbor” does an efficient job at protecting privacy, but using this method causes so much damage to the original data’s integrity that its value for sophisticated performance testing is greatly diminished. The de-identification process not only removes the patient’s relevant information, but requires the removal of any Personally Identifiable Information (PII) found within the doctors’ notes, which is where many de-identification efforts fall short.

Consider a cluster of Zip Codes, with the first same first digits. If the population of that cluster is greater than 20,000, it is converted to the same Zip Code. The residents of that area containing over 20,000 people are now converted to a single Zip Code, containing the correct first three numbers of their Zip Code (see Identifier stipulation b. >i.). In the case where one or more rare diseases are tracked in the de-identified data set, the risk of identification becomes significant. This medical data, cross correlated with freely available data from social web groups, as well as data medical portals, and IP information stored through public health records contains enough information to pinpoint the patient’s identity. The availability of information has grown beyond the scope of what “safe harbor” protection hoped to provide when HIPAA law was enacted 1996. People can and will be identified. Similarly anonymized data

has been re-identified such as the re-identification of the Netflix Prize database by use of the relatively new (in comparison to what was available in 1996) sources of freely available information.¹

The table below (Table 1) shows the identifiers required to be removed by the HIPAA Privacy Act of 1996. The data elements essential to testing many decision making systems have been marked in bold.

Table 1: HIPAA's 18 Identifiers

- a. **Names;**
- b. **All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:**
 - i. **The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and**
 - ii. **The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.**
- c. **All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;**
- d. Telephone numbers;
- e. Fax numbers;
- f. Electronic mail addresses;
- g. **Social security numbers;**
- h. **Medical record numbers;**
- i. **Health plan beneficiary numbers;**
- j. **Account numbers;**
- k. Certificate/license numbers;
- l. Vehicle identifiers and serial numbers, including license plate numbers;
- m. **Device identifiers and serial numbers;**
- n. **Web Universal Resource Locators (URLs);**
- o. **Internet Protocol (IP) address numbers;**
- p. **Biometric identifiers, including finger and voice prints;**
- q. Full face photographic images and any comparable images; and
- r. **Any other unique identifying number, characteristic, or code**

Taken from: <http://irb.jhmi.edu/HIPAA/deidentifieddata.html>

Method 2 inherently requires the removal of the same if not similar elements otherwise it would be very difficult for a statistician would be able to say that the chances are significantly low that the data is safe from re-identification when used alone or in combination with other freely available information. For the purpose of this paper we will discuss only Method 1 seeing that it has cited specific elements to be removed.

Inherent Issues with De-Identification

The removal of key elements results in the inability to adequately test several functions of medical data management systems. Systems whose responsibility is to detect outbreaks based upon

¹ http://userweb.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

geographic/demographic information are a prime example. Clinical decision support systems cannot be fully tested or vetted due to the lack of account information, social security numbers and health plan beneficiary numbers. Systems that monitor possibly failing devices will have no pertinent data with which to test when removing device identifiers such as MAC addresses, URLs and web IP information.

These elements have been removed to protect the privacy of the patient; but what is protecting the patient when the software developed to house and use that information has not been properly tested? Without the use of demographic, financial and personal information, systems using this data cannot be properly vetted to ensure the highest quality of healthcare and healthcare related decisions are being made.

Prior to de-identification there are still many concerns using a set of production data. Determining what is in the data being the first concern. Data is often chosen and then a lengthy and costly truthing process ensues to ensure the tester knows what is/can be tested. Very often this is done using manual and automated verification; to ensure enough confidence in what is known about the data set triple verification is used. This can be extremely cost prohibitive to accomplish.

Another issue is the quality of the de-identification process. PII can often be found within the doctor's SOAP notes, providing adequate information to determine the actual identity of the patient. Data sets that haven't been properly scrubbed of this PII introduce an unacceptable risk for both the data provider and system developer.

Using de-identified data the following medical systems cannot be properly tested:

- Clinical Decision Support Systems
- Medical Benefits Decision Support Systems
- Computerized Provider Order Entry systems
- Medical Outbreak Response Systems
- Fraud detection systems

To convert this data into to statistically meaningful information to verify proper testing would require supplanting the missing or creating data sets that fulfill these requirements without violating HIPAA or other privacy regulations.

The Dynamic Data Generator™ - Freedom to Test Completely

The DDG creates fully synthetic records. This means that the 'objects' generated, be it people, whole households, or taxable businesses/households, are completely fabricated based upon models making the data completely devoid of any privacy or confidentiality concerns.

Modeling outbreaks is critical to test systems the monitor outbreaks, and hopefully to limit or prevent a massive pandemic by the proper response from the medical community. Without the availability of adequate demographic information for testing this can't be done. It is inherent in an outbreak situation to have epicenters, regions that are affected and people affected. Without proper testing using similar information, early detection systems will be fundamentally flawed. Consider the following outbreaks:

H1N1 Outbreak², West Nile Virus Outbreak³, Milwaukee Cryptosporidium outbreak⁴. All of these examples used demographic information to ascertain the root cause of the pandemic, educate the public and eventually minimize or eradicate the threat.

Medicare and Medicaid fraud is costing the US Billions annually. "According to Steven Malanga of the Manhattan Institute, experts estimate that "abuses of Medicaid (alone) eat up at least 10 percent of the program's total cost nationwide -- a waste of \$30 billion a year."⁵

The lack of proper testing for the financial aspect of Health IT systems increases the difficulty to have a significant effect on Medicare and Medicaid fraud. In fact, the expense will increase despite the government's increasing investment into systems and incentives for meaningful use⁶.

Adequate testing, including that of financial information, is difficult to impossible with current standards of de-identified data. This inadequacy severely limits the proper vetting of PII systems, which in turn, makes it difficult to minimize that fraud that costs our country billions of dollars each year.

Conclusion

A system that is designed to use/leverage personal information, demographic information or financial information must be tested with such information and cannot be done well when only using de-identified information. How would you test a calculator's functionality if prior to packaging it you couldn't use numbers due to legal issues? The short answer is you can't. De-identified data is, at best, incomplete, non-exhaustive and marred with legal concerns.

Data that mimics real world, production data addresses and circumvents these complications. The DDG also provides unlimited flexibility when it comes to variations to simulated data and/or anomalies built into the data. This ensures that every reasonable eventuality a system will be faced with is in your data model, and thus test data can be built with controls exposed to the user to allow for complete customization to meet any and all testing needs.

² H1N1: <http://www.nejm.org/doi/pdf/10.1056/NEJMe1004468>

³ West Nile: <http://www.contentnejmorg.zuom.info/cgi/content/full/344/24/1807>

⁴ Milwaukee Cryptosporidium: http://en.wikipedia.org/wiki/Milwaukee_Cryptosporidium_outbreak

⁵ <http://www.examiner.com/law-enforcement-in-national/billions-medicare-medicaid-lost-to-fraud-abuse>

⁶ <http://www.emrandhipaa.com/emr-and-hipaa/2010/08/09/emr-stimulus-meaningful-use-checklist/>