HIPAA Certification of the Provisio De-identified Data Tiers

November 28, 2006

by Fritz Scheuren and Patrick Baier

Contents

1. DECLARATION OF FRITZ J. SCHEUREN AND PATRICK D. BAIER

1. Fritz J. Scheuren is a resident of Alexandria, Virginia, is over the age of 18 and has personal knowledge of the matters set forth in this declaration. Patrick D. Baier is a resident of Washington, D.C., is older than 18 years and has assisted Fritz Scheuren in a number of HIPAA related certifications.

1.1. Qualifications

1.1.1. Fritz Scheuren

2. Fritz Scheuren is the Vice President for Statistics at NORC, a research arm of the University of Chicago, and has been so from 2001 through the present. At the Urban Institute, from 1999 to 2001, he directed the National Survey of America's Families ("NSAF") and made all decisions on how to release NSAF data to the public in a de-identified form. From 1997 to 1999, he served as National Technical Director, Statistical Sampling Economics Group, for Ernst & Young, LLP. From 1994 to 1996, he was a Professor of Statistics at the George Washington University, where he has served as Adjunct Professor of Statistics since 1988. From 1980 to 1994, Dr. Scheuren was Director, Statistics of Income Division, at the Internal Revenue Service. He was Chief Mathematical Statistician of the Social Security Administration from 1978 to 1980.

3. Dr. Scheuren is an expert on the use of generally accepted statistical and scientific principles and methods for de-identifying information - and has even developed some of these methods. He has over thirty years of experience as a mathematical statistician. He received his Ph.D. in Statistics from the George Washington University in 1972. He has published over two hundred applied and theoretical papers, monographs, and books focused on privacy and confidentiality issues, administrative record research, record linkage, survey sample design, and estimation. He is a member of the following professional organisations, among others: American Statistical Association (Vice Presi-

dent, 1999 to 2001, President Elect 2004, President, 2005), International Association of Survey Statisticians (Scientific Secretary, 1997), National Academy of Sciences, Applied and Theoretical Statistics (1994 - 1997), and Washington Statistical Society (President, 1991-92). He served as Associate Editor of the Journal of the American Statistical Association from 1989 to 1996 and remains an Associate Editor of the Survey Methodology Journal, where he has served since 1986.

4. In 1994, Dr. Scheuren was a member of the Federal Committee on Statistical Methodology ("FCSM") and, as such, participated in the drafting of Statistical Policy Working Paper 22, Report on Statistical Disclosure Limitation Methodology. This FCSM volume was published by the Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget in May, 1994, as Statistical Working Paper No. 22 ("Working Paper 22"). [1]

5. Under the privacy regulations issued pursuant to the Health Insurance Portability and Accountability Act of 1996 ("HIPAA"), the Secretary of Health and Human Services has specifically approved the use of FCSM Working Paper 22 as one of two guidelines to "generally accepted statistical and scientific principles and methods for rendering information not individually identifiable." The second, more process-oriented guide mentioned by the HIPAA Privacy Rule is the FCSM "Checklist on Disclosure Potential of Proposed Data Releases." While Dr. Scheuren did not work on development of this second FCSM protection tool, he has used it extensively and commented on its strengths and limitations in several publications. I regularly speak before statistical meetings regarding privacy matters and methods to protect the confidentiality of individual data. In November 2002, for example, I spoke before the FCSM at their annual government-wide meeting, commenting on new methods to afford adequate de-identification protections in federal government data products.

[1] www.fcsm.gov/working-papers/SPWP_22_rev_TOC.pdf

1.1.2. Patrick Baier

6. Dr. Baier holds a D.Phil. in Mathematics and brings in his experience both with statistical programming and the use of cryptography in the HIPAA world. He has been working as a statistician for the National Opinion Research Center (NORC) on a number of projects.

7. In several HIPAA certifications over the past three years for a number of health care companies, many of them operating on a national level, Patrick Baier has assisted Fritz Scheuren in carrying out statistical analyses. Moreover, Dr. Baier has reviewed encryption technologies used to de-identify data, including a review of protocols and source code. In 2006, Dr. Baier has spoken about the HIPAA Rule in an invited session at the Joint Statistical Meeting in Seattle.

2. OVERVIEW

8. The Health Insurance and Accountability Act of 1996 (HIPAA) prohibits the disclosure or dissemination of Protected Health Information (PHI) by a Covered Entity, for example, a health plan, health care provider, or health care clearing house. Protected health information is generally personally identifiable data pertaining to the health or medical treatment of an individual, or the payment of medical services. The precise legal definitions, and exceptions, can be found in 45 CFR §160.103.

9. Health information can be made exempt from the HIPAA provisions if it is de-identified so that it is no longer considered PHI. The HIPAA Rule provides several options how to do this, notably the Safe Harbor approach which requires the removal of eighteen (18) types of variables from any data set, or statistical de-identification. The Safe Harbor approach is a simpler, more general technique for rendering data de-identified and is universally applicable to any covered data set. The types of variables to be removed are listed in 45 CFR §164.514(b)(2).

10. Statistical de-identification, as defined in §164.514(b)(1), by contrast requires that

> "A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
>
> (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
>
> (ii) Documents the methods and results of the analysis that justify such determination [...].

11. This document provides a statisticians' certification in the sense of 10. that certain data sets of medical information have a very small risk of unauthorised re-identification. We define the scope of this certification - the data to be certified and the process used to produce these data - in Section 3. We describe the statistical methodology used and present the results of our analysis in Section 4. In Section 5 we provide the results of our analysis of the data, using the methodology of Section 4. We also describe there any modifications to the data that may be necessary to allow us to conclude that the risk of re-identification is small. We incorporate a requirement of certain procedural safeguards into this certification in Section 6. Adherence to these requirements provides a framework within which our statistical models are applicable and our conclusions are valid. In the language of HIPAA, this section defines the "anticipated recipient" with regard to whom we conclude that there is a small risk of re-identification. A summary and conclusion follows in Section 7.

3. SCOPE OF THIS CERTIFICATION

12. Provisio has developed a "de-identification" engine called *iTrials Service Platform* which is made available to Covered Entities for de-identification of PHI. The engine performs two tasks. First, it uses certain identifiers (names, dates of birth, and gender) which, in combination, are highly unique to an individual, and creates a patient linking key. It subsequently removes or modifies these identifiers in order to create a de-identified data set.

13. The linking key is derived using strong cryptographic algorithms to make it very difficult to re-identify a patient from the linking key. The algorithms are designed to resist reverse engineering, that is, the recovery of the input data from which the keys are constructed. This technology has been reviewed separately and found to be adequate to protect the confidentiality of patient data.

14. The removal or modification of identifiers for the purpose of de-identification can be performed in a number of ways called de-identification "tiers." Each tier corresponds to a well-defined set of rules about removal or truncation of data fields. In this certification, we review each tier separately and determine whether the resulting data can be considered de-identified in the sense of HIPAA.

15. Our determination of whether or not a de-identification tier meets the requirements for statistical de-identification under HIPAA is based on

(a) A review of variables included in the output file layout,

(b) A statistical analysis of the data bases created by the de-identification engine, if appropriate based on a representative random sample,

(c) A review of policies and procedures surrounding the creation, use, storage and handling of the data.

16. We document in detail the statistical methodology we employ in Section 4. The

tiers and the file layouts created by each tier are discussed in Section 5, which also shows the results of our statistical analysis, individually for each tier. Required procedural safeguards needed to ensure that the statistical models we use are valid to measure the risk of unauthorised re-identification are discussed in Section 6.

4. STATISTICAL METHODOLOGY

17. In order to be able to estimate the risk of disclosure, we need to model the tools and understand the methods an intruder could reasonably have available to try to re-identify an individual. These depend non only on the information contained in the databases, but also depend strongly on how easy it is for an intruder to gain access to these data, and what risks an intruder needs to take if he wants to break the HIPAA law.

18. We consider it most appropriate to approach the issue of disclosure limitation from several different directions:

- Statistical disclosure limitations (de-identification of the data)
- Technical security measures
- Legal protections
- Procedural safeguards

In our experience it is most secure and most beneficial from a business point of view to implement the procedural safeguards discussed in Section 6.

19. For example, a de-identified data set may be at much higher risk of unauthorised disclosure if it is made available to the general public, say, on the Internet, where anybody can download them and do whatever they want with them, than if they are stored securely on a protected server with access limited to a few authorised users, and where

any illegitimate activity can be monitored and detected. The risk of unauthorised re-identification will also be lower if an intruder risks prosecution or losing their job, than in a situation where no sanctions apply.

20. Of course, it is not possible to mathematically quantify the "risk" impact of certain legal arrangements or technical security measures. Our approach, therefore, is to use legal, technical, and procedural requirements as a framework that ensures that the data are handled in the most restrictive, secure way possible. We then apply quantitative, statistical data analyses which, based on our judgement and experience, provide a valid measure of the risk of unauthorised disclosure, within that legal, technical, and procedural framework.

21. Once we have established that all de-identified data are handled within a controlled framework, with a closed chain of custody, strong physical, electronic, and legal safeguards, we have found the following statistical model to be a valid, conservative benchmark for measuring the risk of re-identification.

(a) We determine that no direct identifiers are on file.

(b) We exclude certain variables from consideration on the basis that, within the controlled environment we have established, they do not provide a reasonable means of re-identification.

(c) The remaining variables are "indirect identifiers." Such variables do not uniquely identify patients, but, taken in combination may have some identifying power. The risk of re-identification from these indirect identifiers needs to be measured and shown to be small. We perform the following analysis on all indirect identifiers, including all additional (medical or non-medical) data which we have reason to believe may be usable by somebody who wishes to re-identify patients.

    (i) We construct a vector of all such indirectly identifying data elements and perform a frequency analysis across the entire database, or a representative

sample of it.

(ii) If we find that at most 1% of the patients are in "small cells," that is, only 1% of the patients or less share their (combination of) indirect identifiers and unique medical or non-medical information an intruder might use with only a small number of other patients, we are satisfied that the risk of unauthorised re-identification is small, and acceptable under HIPAA. The reason is that in the vast majority of cases, this type of information does not uniquely identify a patient. For the purpose of this argument, we define a "small cell" to be a cell of one (1) or two (2) patients. This bound is in keeping with Working Paper 22.

5. STATISTICAL ANALYSIS

22. We have carried out the analysis described above for the de-identification tiers described below and in Subsections 5.1 - 5.7. The following types of data are always removed:

1. Patient names
2. Patient location information other than a zip code (Zip codes are handled on a per-tier basis below)
3. Birth dates are handled on a per-tier basis below
4. Patient Telephone numbers
5. Patient Fax numbers
6. Patient Email addresses
7. Patient Social security numbers
8. Patient Medical record numbers

9. Health plan beneficiary numbers

10. Patient Account numbers

11. Patient certificate or license numbers

12. Patient vehicle ID numbers, including license numbers.

13. Patient device identifiers and serial numbers

14. Patient URLs or other Internet address information regardless of protocol, including IP addresses.

15. Biometric identifiers, including fingerprints, voice prints, etc.

16. Full face photographic images of the patient, or other comparable images that may allow reasonable determination of the identity of the patient.

17. Any other identifying number, characteristic or code.

23. In addition to removing the types of variables listed in §22, the following modifications are made specifically to each tier.

Tier	Description
Tier 1	This is the most secure tier; after the fields above are removed, any patient zip code information is also removed, and all birth dates are set to January 1 of the birth year.
Tier 2	Same as tier 1, except the 3 most significant digits of each patient zip code are preserved.
Tier 2a	Same as tier 2, except 4 significant digits of the patient zip code are preserved.
Tier 2b	Same as tier 2, except all 5 digits of the patient zip code are preserved.
Tier 3	Same as tier 2, except the quarter the patient was born in is assigned (Jan 1, Apr 1, Jul 1, or Oct 1 of that year)
Tier 3a	Same as tier 3, except 4 significant digits of the patient zip code are preserved.
Tier 3b	Same as tier 3, except all 5 significant digits of the patient zip code are preserved.

5.1. Tier 1

24. We have reviewed a representative random sample of 99,997 records pertaining to 41,554 distinct patients. The following variables are on the output file created by the de-identification engine.

Variable	Description	Comment
ID	A sequential serial number	1
HashCode	The linking key created by the de-identification engine	2
Gender	Gender (=F/M)	3
ZipCode	Always blank (NULL) in this tier	4
YoB	Year of Birth (YYYY)	3
DoB	Either blank (NULL) or set to 1900-01-01	4
DateOfService	date of service, DDMMYYYY	5
ICD9Code1-4	Diagnostic information	5
CPTCode	Diagnostic information	5
Facility Zip	Facility Zip Code	6
POSCode	Diagnostic information	5
SourceRef	System information	1

25. Not all of these variables are relevant in our analysis of the risk of unauthorised re-identification. Specifically, we determine as follows:

1 Serial numbers and reference numbers that are generated randomly or sequentially, without any relation to the individual, and that are solely needed for proper functionality of the data base, have no risk of being used to re-identify patients. We therefore exclude them from consideration.

2 The linking code is generated by a secure cryptographic algorithm, reviewed separately. We are satisfied with the design of this algorithm to ensure that unauthorised re-identification is very unlikely.

3 Gender and year of birth are indirect identifiers studied in more detail below.

4 Fields that are deleted in this tier need not be considered.

5 Diagnostic and treatment data such as ICD9 and other medical codes, and treatment dates may be highly unique to a patient. However, such data are not reasonably available to an intruder for re-identification because they typically reside only with the covered entity and are already protected under HIPAA. We therefore do not consider them a more than minimal risk of allowing re-identification.

26. Following the analysis above, we are satisfied that no direct identifiers are on the output of Tier 1. We have analysed the frequencies of all combinations of indirect identifiers, in this case year of birth and gender. We find as follows: Only four (4) patients out of 41,544 are in "small cells" of size < 3, corresponding to 0.01% of the sample. This is a minimal fraction, and we are satisfied that the data can be considered de-identified.

5.2. Tier 2

[To be completed.]

5.3. Tier 2a

[To be completed.]

5.4. Tier 2b

[To be completed.]

5.5. Tier 3

[To be completed.]

5.6. Tier 3a

[To be completed.]

5.7. Tier 3b

[To be completed.]

6. PROCEDURAL SAFEGUARDS

27. The HIPAA Rule requires that the risk of re-identification be small with respect to "an anticipated recipient." By imposing these restrictions, we limit access to the data

and limit the ability of an intruder to link the data to external data sources.

6.1. Data Checking

28. When large databases are created and some of the information is manually entered, there is always a risk that data is entered in the wrong place. Hence it may happen that a variable that is intended to store a treatment code may actually store a name or birthday. Software that is written to standardise and de-identify the file may not recognise this, and identifying information may leak in this way. Since we have not inspected the identifiable source data from which the sample databases were created, we are not in a position to estimate precisely how large this risk is. However, it is necessary that both automatic consistency checks be performed (on all deliverables) and occasional reviews by a qualified individual be conducted, especially in the early phases.

29. A qualified individual, for the purpose of § 28, would be, for example, an experienced data base programmer who has already been working independently with the data, or a manager who is familiar with the meaning of the data. Such reviews should be performed more frequently initially, but can be reduced to occasional random checks, perhaps once a month, once the process has been established and found to be working properly.

30. A consistency check would entail reviewing a small number of records to determine that each variable contains the information it is intended to contain. For example, there should be no birth dates in the ZIP code field, and no names should appear in place of a hashed linking code, etc. Moreover, it should be verified that the rules of the de-identification tiers have been applied as required, that is, the necessary deletions and truncations have been performed. Should identifiable information be found on a file by error, the database software needs to be reviewed and measures taken to prevent the occurrence of such events.

6.2. Electronic Protections for Data Transfer and Storage

31. Data transmission is initiated by Covered Entities via secure, electronic means (e.g., secure FTP) between servers which sit behind firewalls. The transmission initiation and file receipt is logged, transmission monitored and an automated notification system emits status of transmission, size of file and notification of failures to the System Administrators and the managers of that data feed. Alternatively, data may be transmitted by secure carriers with a closed chain of custody (such as FedEx, UPS, DHL, etc.).

32. The requirements summarised above are typically satisfied by using common industry standard solutions, such as secure file transfer protocols and commercially available strong encryption. We concur that they meet the requirements for securing the electronic transfer of HIPAA sensitive materials.

6.3. Legal Arrangements

Provisio has the following legal arrangements in place to protect patient data. Provisio has included in its contractual agreements with the recipients of Provisio's de-identified data or products derived from such data (i.e., Provisio's clients) a legal requirement for the client

(a) not to attempt to re-identify any of the data provide

(b) not to deliver any micro-data (patient level data) to third parties, unless such parties are bound by the same legal obligations

6.4. Employee Training and Screening

33. Provisio has in place (or is committed to putting in place, prior to delivering any data) written policies, both on Provisio's and its clients' side, for handling the data files. Only trained employees are able to access or see HIPAA protected data. The procedures require familiarity with the HIPAA rule, employee screening, and training on a regular

basis. Employees have to sign a written commitment to protect patient data they are entrusted with.

6.5. Data Access

34. Access to the data on the recipient's side is on a strict need-to-know basis to prevent unauthorised use or theft of the data. Access to the data warehouse is restricted to a limited number of HIPAA trained, authorised users (via user ID and password combinations) and only from certain authorised workstations.

6.6. Physical and Safeguards

35. The recipient of any data provided by Provisio is contractually bound to maintain the data in a physically secure location, and ensure the data are appropriately protected.

7. CONCLUSION

36. Based on our statistical analysis a sample of data created by the iTrials de-identification engine, and on our review of the procedural framework in place at Provisio and Covered Entities providing Provisio with de-identified data, we are satisfied that there is only a *de minimis* risk of unauthorised re-identification. We therefore concur that the data created under Provisio's Tier 1 are compliant with the standard for statistical de-identification required by HIPAA.

Fritz Scheuren, Ph.D.

P. Baier

Patrick Baier, D.Phil.

Washington, DC, 11/28/06