Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

June 1, 2009

EPIDEMIOLOGY

Subject: Health Breach Notification Rulemaking, Project No. R911002

Dear FTC Secretary Donald S. Clark:

Thank you for this opportunity to offer public comments in response to the FTC's solicitation of comments regarding the NPRM for the proposed new Part 318 of 16 CFR for Personal Health Records (PHRs).

I have a number of specific comments to offer regarding the Commission's request for examples of any instances beyond compliance with the HIPAA standard for de-identification as specified in 45 CFR 164.514(b) in which there would be "*no reasonable basis to believe that information is individually identifiable*".

I approach this topic from my perspective as an academic epidemiologist with specific experience and expertise in conducting statistical disclosure control analyses consistent with the requirements of the HIPAA Privacy Rule to assess re-identification risks in healthcare data sets under the statistical de-identification provision [§164.514(b)(1)]. My comments on the proposed new Part 318 of 16 CFR are, therefore, particularly focused on the security and protection of health data within PHRs from statistical re-identification attacks.

I also have several further comments to make about the relationship of de-identification issues to points regarding whether various technologies or methods sufficiently reduce re-identification risks so they may be considered appropriate for rendering individually identifiable health information "unusable, unreadable or indecipherable" and, therefore, no longer considered to be "unsecured". It is the risk of re-identification that obviously causes these issues to be closely intertwined and, because of this, I also offer some important clarifications/corrections to earlier public comments that you have received from others about the estimated re-identification risks for various data set configurations in relationship to the Limited Data Set allowed under HIPAA in 45 CFR 164.514(e).

COMMENTS REGARDING
DE-IDENTIFICATION

Point 1: ***Health information which would qualify as properly de-identified under the HIPAA statistical de-identification provision at 45 CFR §164.514(b)(1) (once having been properly certified as de-identified by a statistician in accordance with this section) should be presumed to have had "no reasonable basis" for belief that the information could be used to identify an individual and, thus, should be considered to not be "PHR identifiable information" after such a determination of statistical de-identification has been properly made.***

In those cases when health information has been breached and such breached information in fact does have a "very small risk" of re-identification as would be determined by a statistician in accordance with 45 CFR §164.514(b)(1), but where no such evaluation or determination has yet been rendered, it would be helpful for the Commission to explicitly clarify that determination of statistical de-identification by an after-the-fact determination of a "very small risk" of re-identification would also constitute an appropriate means of establishing that there was indeed "no reasonable basis" for belief that the information could be used to identify an individual and, therefore, the health information in question would be considered not to have been "PHR identifiable information".[1]

It would clearly be in the best interests of PHR vendors, PHR related entities, and their third-party service providers to have completed any determination of statistical de-identification (for data sets which they wish to have so designated) at the time of data set creation in order to assure their ability to meet the proposed conditions in 16 CFR 318.4 (the Timeliness of Notification provision). However, it would be useful for the purposes of reducing the burden of compliance for PHR vendors to make it clear that an after-the-fact determination of statistical de-identification would also suffice to establish that a data set was indeed considered to be de-identified under Part 318 of 16 CFR.

Recommendation 1:

The FTC, in coordination with the HHS, should issue joint guidance to clarify that that after-the-fact determinations of statistical de-identification would suffice to establish that a data set was indeed considered to be de-identified under Part 318 of 16 CFR at the time of a breach.

Point 2: ***For any data sets which have not already been established as de-identified in accordance with 45 CFR §164.514(b) at the time of a breach, the date on which a breach should be treated as discovered should remain as proposed in section 318(c) (as the day on which the breach was known to the PHR vendor, PHR related entity or third party service provider) in order to assure that PHR vendors provide breach notifications "without reasonable delay".***

---

[1] For further discussion of this point, see Hubbard, M. and Wilson D. "*De-identified Health Information: Legal and Practical Approaches to HIPAA Compliance*". pp.364-366 In Gosfield, A.G., Ed. Health Law Handbook 2006. Thomson/West.

In the absence of a pre-existing determination of compliance with 45 CFR §164.514(b) in advance of a breach, the day on which the breach became known to the PHR vendor, PHR related entity or third party service provider should continue to serve as the starting clock for the period for the evaluation of "unreasonable delay" and the 60-day maximum period for breach notification. A 60-day period should be fully adequate for concluding any evaluation of compliance safe harbor or statistical de-identification, and such an approach would be consistent with sections 13402(c) and 13402(d)(1) of the HITECH/ARRA act without encouraging any undue delays in notification as having been supposedly justified in order to allow such after-the-fact determinations of statistical de-identification.

Recommendation 2:

The FTC, in coordination with the HHS, should issue joint guidance explicitly stating that the date on which a breach should be treated as discovered will remain as proposed within section 318(c) for any data sets which have not already been established as de-identified in accordance with 45 CFR §164.514(b) at the time of breach.

Point 3: ***HHS has recently received public comments supporting the use of "Data Masking" methods in response to their RFI specifying the technologies and methodologies that render PHI unusable, unreadable, or indecipherable to unauthorized individuals under section 13402 of Title XIII of 2009 HITECH/ARRA Act of 2009. Because data masking, scrambling, or other obfuscation methods can be complex and may fail to protect against re-identification when they are not properly implemented, the use of such methods without appropriate statistical review and certification is not appropriate as an unsupervised technology for rendering PHI unusable, unreadable, or indecipherable. However, such methods can be valuable and important tools for supporting the creation of testing and development data as part of the established statistical de-identification process under 45 CFR §164.514(b)(1) and should be explicitly mentioned as appropriate methods for statistical de-identification within this context.***

The safe harbor de-identification method in 45 CFR §164.514(b)(2)(i) specifically requires the ***"removal"*** of all eighteen types of identifiers specified within the safe harbor provision. The safe harbor provision, therefore, does not anticipate or accommodate the need for data masking, scrambling or obfuscation methods in order to create realistic de-identified data for these PHI data fields for the purposes of testing and developing software systems, database systems, electronic medical records and personal health records. It would be helpful for the FTC, in collaboration with HHS, to issue joint guidance explicitly indicating that such data replacement or data generation methods may be appropriately used if they have been reviewed by a statistician and found to pose no more than a "very small" risk of re-identification in accordance with the statistical de-identification provision at 45 CFR §164.514(b)(1).

Recommendation 3:

The FTC, in coordination with the HHS, should issue joint guidance explicitly stating that data masking, scrambling, and obfuscation methods may be appropriately used only if they have been reviewed by a statistician and found to pose no more than a "very small" risk of re-identification in full accordance with the statistical de-identification provision at 45 CFR §164.514(b)(1).

COMMENTS REGARDING
UNSECURED PHR IDENTIFIABLE
HEALTH INFORMATION

My previous comments within this letter have addressed specific requirements for proper data de-identification so that there would be no reasonable basis that health information is "individually identifiable". Such situations are appropriately contrasted with the circumstance in which "individually identifiable" health information has been secured so as to have been rendered unusable, unreadable or indecipherable to unauthorized persons. The FTC has indicated in the NPRM that, consistent with section 13402(h)(2) of the ARRA/HITECH Act, the new Part 318 of 16 CFR is proposed to contain section 318.2, which indicates that the definition for "unsecured" for PHRs will be specified by the Secretary of HHS. HHS recently received public comments in response to their RFI specifying the technologies and methodologies that render PHI unusable, unreadable, or indecipherable to unauthorized individuals under section 13402 of Title XIII of 2009 HITECH/ARRA Act of 2009. I also made earlier public comments in response to this RFI and will not repeat them in full here, but will briefly mention my earlier recommendations here and also provide an attachment with my complete earlier comments. I include these additional comments in the hope that they will be helpful to the FTC in considering similar issues with regard to the specification of the technologies and methodologies that render individually identifiable PHR health Information unusable, unreadable, or indecipherable to unauthorized individuals as you coordinate and harmonize FTC and HHS rules under ARRA/HITECH.

I will begin, however, by addressing one point on which I did not offer earlier comment, but which was raised by others in response to the recent HHS Request for Information with regard to the use of cryptographic hash methods.

Point 4: ***One-way hashing should be added to the technologies and methods approved to render health information unusable, unreadable or indecipherable.***

In response to the recent May 22, 2009 HHS RFI, the Markel Foundation's *Connecting for Health* initiative and others have made the argument quite convincingly that one-way hashing is a valuable cryptographic method that should be added to the encryption methods already approved by HHS for the purpose of rendering health information unusable, unreadable or indecipherable. While it is to be clearly acknowledged that one-way hash functions can be subject to various methods of attack (as can all cryptographic methods - including strong encryption methods), one-way hash methods offer important advantages to the random number replacement methods that HHS has clearly approved for dealing with section 164.514(2)(i)(R) ("Any other unique identifying number, characteristic, or code, except as permitted by paragraph 164.514(c)") in order to assure that such randomly generated replacement numbers are clearly consistent with the requirement at 164.514(c) indicating that "The code or other means of record identification ***is not derived from or related to information about the individual*** and is not otherwise capable of being translated so as to identify the individual;". Rather than take time here on arguments which have already be well made by others, I will simply note that the use of one-way hash methods for the purposes of creating longitudinal patient data records has considerable logistic and practical advantages over the currently permitted de-identification methods involving the use of unsecured

PHI by contracted HIPAA business associates performing de-identification on behalf of CEs[2], and add that, with proper implementation, the use of such one-way hash methods would arguably create better protections from unauthorized access than the currently allowed approaches.[3]

Recommendation 4:

The FTC should, in its consultations with HHS, encourage the allowance of one-way hash methods as an approved technology or method to render health information unusable, unreadable or indecipherable.

REITERATION OF CRUCIAL POINTS FROM MAY 22, 2009 PUBLIC COMMENTS TO HHS

As already mentioned, I will not reiterate my complete comments from my May 22, 2009 public comments to the recent HHS request for information. However, because these points are crucial to the protection of health information from statistical re-identification risks, I will briefly summarize my points and associated recommendations here and will attach my previous comment letter for FTC review as an appendix to this letter.[4]

Previous HHS Comments Point 1: ***Encryption of individual data fields within otherwise unencrypted data sets will not necessarily provide secure protection from re-identification for those encrypted fields. Encryption of the full data set is necessary to assure protection from statistical re-identification attacks.***

Previous HHS Recommendation 1:

The FTC should issue joint guidance with HHS clarifying that proper encryption resulting in secure protection, thus rendering electronic PHI unusable, unreadable and indecipherable, would exist only when appropriate strong encryption methods have been applied to the entire set of data elements for each individual.

---

[2] HHS amended Sec. 164.514(e)(3)(ii) to make it clear that a covered entity may engage a BA to create an LDS (in the same way it can use a business associate to create de-identified data). The covered entity may hire the intended recipient of the LDS or de-identified data as a BA for this purpose. That is, the covered entity may provide protected health information, including direct identifiers, to a BA who is also the intended data recipient, to create an LDS appropriate for the BA's subsequent use. (See p. 53237 of the August 14, 2002 Federal Register (Vol 67, No 157)).

[3] See Apfelroth S. Zero-Check: a zero-knowledge protocol for reconciling patient identities across institutions. Arch Pathol Lab Med. 2004 Sep;128(9):954; and author's reply p. 954-6 and Berman JJ. Zero-check: a zero-knowledge protocol for reconciling patient identities across institutions. Arch Pathol Lab Med. 2004 Mar;128(3):344-6 for an insight debate on this topic.

[4] Or see my Public Submission: (HHS-OCR-2009-0004-0125.1) at www.regulations.gov under Docket # HHS-OCR-2009-004.

Previous HHS Comments Point 2: ***The potential re-identification risks for Limited Data Sets (LDSs) are often very high, and could be as high as 100% under the HIPAA LDS specifications. Because the HIPAA LDS specification is defined only in terms of its exclusion criteria, LDSs could potentially be entirely re-identifiable.***

Previous HHS Recommendation 2:

The demonstrated very high potential re-identification risks for Limited Data Sets do not make this an acceptable technology or method for rendering PHI unusable, unreadable, or indecipherable to unauthorized individuals. HHS should not include LDS among the approved methods for rendering PHI unusable, unreadable, or indecipherable to unauthorized individuals.

DISCUSSION OF RE-IDENTIFICATION RISK ESTIMATES PROVIDED BY OTHER COMMENTORS

Finally, I wish to point out some important omissions/inconsistencies within some public comments that the FTC has already received regarding the estimated re-identification risks for safe harbor de-identified and LDS data configurations which would require clarification/correction. The Health Information Privacy Laboratory at Vanderbilt University has provided the FTC with estimated re-identification risks for Safe Harbor de-identified and several various data configurations for LDSs in their public comment #541358-00087. I have re-organized the findings that they have so kindly provided in the public comments into the table below:

Table 1

ID Set	Public Comment Labels	Actual Quasi-identifiers			Re-ID Risk
1	Safe Harbor (SH) - *Was Incorrectly Labeled*	Gender	YoB	State	**0.0001%**
2	LDS	Gender	DoB	Zip5	68.40%
3	LDS –Year	Gender	YoB	Zip5	0.38%
4	LDS -4Zip	Gender	DoB	Zip4	36.80%
5	LDS-3Zip	Gender	DoB	Zip3	7.50%
6	LDS-2Zip	Gender	DoB	Zip2	0.33%
NCVHS Source[6]	*Safe Harbor* with only Gender YoB and Zip3 evaluated as "***quasi-identifiers***"[5] (i.e. with any other quasi-identifiers such as Race, Martial Status, etc. excluded)	Gender	YoB	Zip3	**0.04%** [6]

[5] Quasi-identifier variables can be characterized as variables which are not direct-identifiers, but which, in combination, have the potential to reveal an individual's identity because an individual is unique with regard to this set of characteristics.

[6] In testimony before the National Committee on Vital and Health Statistics Ad Hoc Workgroup On Secondary Uses Of Health Data on August 23, 2007, Latanya Sweeney, PhD, Carnegie-Mellon University, described a 0.04% chance of re-identifying data when de-identified by removal of the 17 data elements in the HIPAA safe harbor definition of de-identification when compared to voter registration records for a confined population.

First, it should be clarified that the estimate designated as "Safe Harbor" within this public comment was incorrectly labeled and would actually apply to a data configuration that, while compliant with the safe harbor provision, does not represent the maximum risk allowed under safe harbor. The most obvious reason why this is the case is that the safe harbor provision actually allows 3-digit zip codes (with populations greater than 20,000) to be reported.[7] The use of the larger geographic reporting unit (i.e., states) in this estimate results in the reporting of a risk estimate that is approximately 400 times smaller than some recently reported estimates of the safe harbor re-identification risks.[8]

Secondly, it should also be pointed out that the U.S. Census data used to generate these estimates is based on population data for "Zip Code Tabulation Areas" (ZCTAs) rather than "Zip Codes" per se. Because the Census Bureau has purposely excluded from inclusion as ZCTAs approximately 10,000 "Unique" zip codes (which are assigned to a single high-volume address usually for business or large organizations) or Post Office box-only zip codes, these estimates derived from the Census data can be expected to seriously underestimate the potential re-identification risks associated with the various zip code based geographies, unless any such non-ZCTA zip codes within real LDS data have been specially recoded to be eliminated from reporting. Therefore, any risk estimates like these made with the method of Golle[9] using Census ZCTA data should be viewed cautiously as minimal estimates.

Furthermore, it should be also be clarified that for all the risks labeled as "LDS" within this public comment, these risk estimates would actually represent minimal risks, applicable only for any LDSs where the estimates have been correctly made using only datasets containing just the quasi-identifiers included in the three columns following the labels (i.e., marked as "Actual Quasi-identifiers"). As is mentioned by the commenter, these estimates do not, for example, include race as a quasi-identifier, which exists both as part of many standard billing data record sets and as a matter of public record within voter registration lists. Examples of other such quasi-identifiers that are found within standard health information record sets and also as matters of public record include marital status, and date of death.

As also noted in my attached previous public comments to HHS, because LDSs are defined only in terms of exclusion criteria, an LDS could potentially contain an extensive number of additional quasi-identifiers. In addition to allowing quasi-identifier variables which were not listed within the safe harbor exclusion list (e.g., race, gender, martial status, etc.), it is explicitly stated by HHS that an LDS can include the following identifiable information which must be removed under the safe harbor criteria for de-identification: admission, discharge, and service dates; full date of birth and full date of death; age expressed in months, days, or hours (including ages of 90 years or over); five-digit zip codes and any geographic subdivision other than street address (including the equivalent geocodes for such geographic subdivisons). HHS specifically mentions the following geographic subdivisions within the context of geographies: state, county, city, census tract, precinct, and neighborhood. All of these geographies have important research utility which

---

[7] 45 CFR §164.514(b)(2)(i)(B).

[8] See footnote #6 and the last row in Table 1.

[9] Golle, P. (2006) Revisiting the Uniqueness of Simple Demographics in the U.S. Population. http://crypto.standford.edu/pgolle/papers.census.pdf.

validates HHS's allowance of such geographies, if properly justified by a need for such data for a research, public health or health care operation purpose. Therefore, the risk estimates provided in this public comment must clearly be seen as minimal estimates which would only be accurate in a small set of closely constrained LDSs and not generally applicable to the permitted range of LDS quasi-identifier sets that are allowable under 45 CFR §164.514(e).

Thank you for the opportunity to comment on these important issues. Please contact me if further clarification of any of these points would be helpful.

Sincerely,

Daniel C. Barth-Jones, M.P.H., Ph.D.
Assistant Professor of Clinical Epidemiology

Department of Epidemiology
Mailman School of Public Health,
Columbia University

APPENDIX

Public Submission HHS-OCR-2009-0004-0125.1
under Docket # HHS-OCR-2009-004
at www.regulations.gov

Guidance Specifying the Technologies and Methodologies That Render Protected Health Information Unusable, Unreadable, or Indecipherable to Unauthorized Individuals for Purposes of the Breach Notification Requirements Under Section 13402 of the HITECH Act; Request for Information

American Recovery and Reinvestment Act of 2009: Guidance Specifying Technologies and Methodologies that Render Protected Health Information Unusable, Unreadable or Indecipherable

Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

May 21, 2009

EPIDEMIOLOGY

Dear Secretary Sebelius:

Thank you for this opportunity to offer comments in response to the Guidance and Request for Information specifying the technologies and methodologies that render Protected Health Information (PHI) unusable, unreadable, or indecipherable to unauthorized individuals under section 13402 of Title XIII in the ARRA /HITECH Act of 2009.

I have several comments regarding the published guidance for technologies and methodologies that render protected health information unusable, unreadable, or indecipherable to unauthorized individuals as published in the Federal Register on April 27, 2009, as well as comments concerning some of the other public comments that you have received to date on this guidance.

I approach this topic from my perspective as an academic epidemiologist with specific expertise in conducting statistical disclosure control analyses consistent with the requirements of the HIPAA Privacy Rule to assess re-identification risks in healthcare data sets under the statistical de-identification provision [§164.514(b)(1)]. My comments, therefore, are focused mostly on issues pertaining to electronic "data at rest", and the security and protection of such data from statistical re-identification attacks.

Point 1: ***Encryption of individual data fields within otherwise unencrypted data sets will not necessarily provide secure protection from re-identification for those encrypted fields – Encryption of the full data set is necessary to assure protection from statistical re-identification attacks.***

I offer the suggestion that it will be quite important for HHS to issue further guidance clarifying that encryption must be applied to all of the data fields regarding a specific individual within a data set, particularly in the case of "data at rest", in order for the PHI associated with that individual to be considered unusable, unreadable and indecipherable. While strong encryption methods as specified in NIST standards referenced within the guidance will often provide a powerful means of assuring that there will be "a low probability of assigning meaning without use of a confidential process or key", it is widely appreciated within the academic fields of statistical disclosure control and privacy-preserving data methods that encrypted data fields can easily be re-identified simply by virtue of strong statistical associations with unencrypted data fields in those cases where encryption has not been applied to the entire set of data fields to be protected. Unfortunately, when strong statistical associations exist between encrypted and unencrypted (the "known" or "disclosed") data fields, no breach of either the encryption algorithm or the confidential key is required for the true values of these encrypted fields to be revealed simply on the basis of these strong statistical associations with the known data values.

In some of the public comments that you have received, it was suggested that "if all of the 18 identifiers specified for de-identification listed in 45 CFR 164.514(b) for the individual and the individual's relatives, household members, and employers, have been secured", then entire record should meet the standard of unusable, unreadable, or indecipherable. Unfortunately, this is not at all the case. This approach was advocated by the commenter because "strong encryption of specific fields within database tables" would avoid "the huge performance degradation that would be caused by encrypting all fields". However, unless such partially encrypted files have been carefully reviewed by a statistician to assure that there are no strong statistical associations between the encrypted and unencrypted variables, this approach can result in a breach of the encrypted data without either the encryption algorithm or the encryption key having been compromised.

For example, consider the extremely strong statistical association that exists between encrypted dates of birth for infants (having been encrypted in order to protect the birth dates, which are easily accessible matters of public record) and unencrypted maternal dates of hospital admission (having been left unencrypted because such dates of service are typically not matters of public record). In such a circumstance, this strong statistical association can easily be exploited by simple maximum likelihood methods (or perhaps even without formal statistical methods) to reveal the actual dates for the encrypted date of birth values with near perfect confidence. Furthermore, if the same encryption scheme and key has been used for protecting other dates within the same (or other) data set(s), these other encrypted dates may also be systematically revealed via such a statistical attack on this maternal/infant date association. Likewise, encrypted patient geographic locations may often be revealed solely through their geostatistical associations with unencrypted provider locations (e.g., patient's physicians or pharmacies).

Because of this, it will be important for future HHS guidance to make clear that encryption methods should only be deemed to provide secure protection rendering electronic PHI unusable, unreadable and indecipherable when appropriate strong encryption methods have be applied to the entire data record. Data users wishing to employ encryption of only selected data elements within an otherwise unencrypted data set should be aware that this is only likely to be feasible within "statistically de-identified" data sets, where a qualified statistician has reviewed the statistical associations between the encrypted and unencrypted data elements and certified that any risks of re-identification are "very small".

Recommendation 1:

HHS should clarify that proper encryption resulting in secure protection, thus rendering electronic PHI unusable, unreadable and indecipherable, would exist only when appropriate strong encryption methods have be applied to the entire set of data elements for each individual.

Point 2: ***The potential re-identification risks for Limited Data Sets (LDSs) are often very high, and could be as much as 100% under the HIPAA LDS specifications. Because the HIPAA LDS specification is defined only in terms of its exclusion criteria, LDSs could potentially be entirely re-identifiable.***

As stated in the guidance, LDSs are not considered de-identified by HHS. They are considered PHI which may be released under a controlled set of conditions for a select set of purposes[1]. These three purposes (research, public health and health care operations) all clearly produce important societal benefits, which HHS acknowledges as important considerations in their decision to approve the conditions for the LDS. LDSs require "facial" de-identification (i.e., the removal of direct identifiers). Such "facial" or direct identifiers appear to have been implicitly defined as those data elements which would presumably be capable of revealing an individual's identity on the basis of simple "look-up" operations against additional data sources. It should be pointed out that the criteria established for the conditions of an LDS are those of exclusion (i.e., a set of specified direct identifiers must be removed). The removal of such direct identifiers presumably reduces the likelihood of any inadvertent or unintentional re-identification of individuals within the LDS. The LDS specifications were intentionally designed by HHS to not delineate the data that could be released through an LDS. Once all of the required elements have been removed, any other data elements which are justified as being necessary for the purposes of the research, public health or health care operations could be included within an LDS.

Because an LDS is defined only in terms of exclusion criteria, it could potentially contain an extensive number of quasi-identifiers[2]. In addition to allowing quasi-identifier variables which were not listed within the safe harbor exclusion list (e.g., race, gender, etc.), it is explicitly stated by HHS that an LDS can include the following identifiable information which must be removed under the safe harbor criteria for de-identification: admission, discharge, and service dates; full date of birth, full date of death; age expressed in months, days, or hours (including ages of 90 years or over); five-digit zip codes and any geographic subdivision other than street address (including the equivalent geocodes for such geographic subdivisons). HHS specifically mentions the following geographic subdivisions within the context of geographies: state, county, city, census tract, precinct, and neighborhood. All of these geographies have important research utility that justifies HHS's allowance of any geographic subdivision other than street address. It is also important to note that the re-identification risks associated with the data in an LDS must be justified by a need for such data for a research, public health or health care operation purpose. The minimum necessary clause applies to LDSs, so they may not include unlimited sets of data elements any of which are not clearly justified as necessary for the permitted purposes. Still, if all of the data elements are appropriately justified by research, public health or health care operation needs, ***there is no limit to re-identification risks associated with an LDS***. As HHS is quite aware from the citation within the guidance, recent work by statistical disclosure control researchers has indicated nearly two thirds (63%) of the U.S. population would be re-identifiable using data in a very limited LDS containing only three data elements (date of birth, 5-digit zip code and gender)[3]. Although re-identification risks may indeed be small in some LDSs without such high risk variables as dates for matters of public record or detail geography such as zip codes, all of the individuals within an LDS could be potentially re-identifiable given HHS's open-ended specifications for the data elements which are allowable within an LDS.

---

[1] HHS provides a lengthy explanation of their rationale for modifying the proposed privacy rule to allow the use of Limited Data Sets (LDS) for the purposes of research, public health or health care operations on pages 53234-53238 of the August 14, 2002 Federal Register (Vol 67, No 157).

[2] Quasi-identifier variables can be characterized as variables which are not direct-identifiers, but which, in combination, have the potential to reveal an individual's identity because an individual is unique with regard to this set of characteristics.

[3] Golle, P. (2006) Revisiting the Uniqueness of Simple Demographics in the U.S. Population. http://crypto.standford.edu/pgolle/papers.census.pdf.

It should also be noted that, due to the open-ended specifications for the LDS data set, it would simply not be possible to fix the re-identification risks at any particular "safe" level through the simple removal of month or day information from within birth dates or by eliminating the last three digits of a 5 digit zip code. By virtue of how they are defined, LDSs would always have the potential to pose very significant risks of identification for the individuals within an LDS because they may also include an essentially unlimited number of other demographic quasi-identifier variables (e.g., race, marital status) that are matters of public record (or are otherwise "reasonably available" on the internet or from commercial data vendors) and which can be used to link the LDS to such publicly available data. HHS explicitly states in previous guidance that they "*agree that the limited data set is not de-identified information, as retention of geographical and date identifiers measurably increases the risk of identification of the individual through matching of data with other public (or private) data sets*".[4] HHS also directly stated that they "*believe that the limitations on the specific uses of the limited data set, coupled with the requirements of the data use agreement, will provide sufficient protections for privacy and confidentiality of the data*".[5] It is clear that HHS believes that the LDS Data Use Agreement (DUA) conditions that they have stipulated should successfully control any potential risks associated with re-identification attempts involving LDSs. Because LDSs could clearly have high re-identification risks if re-identification were to be attempted, it is also clear that HHS views the probability of any re-identification attempts under the conditions of an LDS DUA as small enough to justify the social benefits resulting from their allowance of the LDS provision.

However, under the conditions of a breech by an unauthorized person, it would need to be presumed that such unauthorized persons accessing the LDS data would not in any way be bound by the protective covenants within the LDS DUA which were deemed by HHS protect such PHI from any re-identification attempts. In fact, in most cases, it is probably only reasonable to assume that such unauthorized persons accessing data during a security breech would have malicious intent that would be consistent with their being likely to undertake re-identification. Additionally, it is also clear that, under the conditions of a security breech, there are no social benefits whatsoever being provided either to the persons who have incurred such re-identification risks, or to society as a whole.

LDSs are extremely valuable for the socially beneficial purposes of research, public health and healthcare operations, so it is quite clear that they could not be considered to be "unusable, unreadable or indecipherable". It is precisely because LDSs are usable, readable and decipherable that they have their considerable value. HHS has, however, permitted their use only for these three socially beneficial purposes, and only under contractual conditions that effectively balance and control any risks of re-identification associated with their use for these specific purposes.

I would also suggest there should be little concern regarding the possible issues of undue administrative and legal burdens or purported chilling effects on the use of LDS for their intended purposes which could supposedly result from HHS not permitting the LDS as an acceptable technology or method to render PHI unusable, unreadable or indecipherable. There are two very effective avenues permitted under HIPAA and HITECH which would allow users of LDSs to easily address any concerns that they might have regarding the breach notification requirements. They may either encrypt any LDSs that they possess, or they may have such LDSs certified as

---

[4] Pages 53234-53238 of the August 14, 2002 Federal Register (Vol 67, No 157).
[5] Ibid.

statistically de-identified when the LDS's re-identification risks would justify a determination of very small risk under the statistical de-identification provision [§164.514(b)(1)].

Recommendation 2:

The demonstrated very high potential re-identification risks for Limited Data Sets do not make this an acceptable technology or method for rendering PHI unusable, unreadable, or indecipherable to unauthorized individuals. HHS should not include LDS among the approved methods for rendering PHI unusable, unreadable, or indecipherable to unauthorized individuals.

Point 3: ***Health information is a public good which benefits society in numerous ways. Re-identification of Limited Data Sets or de-identified data endangers both the privacy of individuals and the use of such health information as a public good.***

Finally, in order to better address the long-term societal interests of protecting the availability of health information as a public good while also protecting the privacy individuals with regard to their personal health information, HHS should encourage Congress to make the act of re-identifying individuals from data contained in either LDSs or de-identified data sets illegal. The linkage of any of the sixteen direct identifiers listed within the LDS exclusion list to data within either an LDS or de-identified data sets under the HIPAA Privacy Rule by anyone other than a Covered Entity who already possesses such information as PHI should be prohibited by law in order to better protect such health information as a vital public good.

Recommendation 3:

HHS should encourage Congress to prohibit by law the act of re-identifying individuals within Limited Data Sets or de-identified data by anyone other than the Covered Entities associated with such individuals.

Thank you for the opportunity to comment on these important issues. Please contact me if any clarification of these points would be helpful.

Sincerely,

Daniel C. Barth-Jones, M.P.H., Ph.D.
Assistant Professor of Clinical Epidemiology

Department of Epidemiology
Mailman School of Public Health,
Columbia University