Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule

September 4, 2012

OCR gratefully acknowledges the significant contributions made to the development of this guidance by Bradley Malin, PhD, through both organizing the 2010 workshop and synthesizing the concepts and perspectives in the document itself. OCR also thanks the 2010 workshop panelists for generously providing their expertise and recommendations to the Department.

Table of Contents

1. Overview

This document provides guidance about methods and approaches to achieve de-identification in accordance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule. The guidance explains and answers questions regarding the two methods that can be used to satisfy the Privacy Rule's de-identification standard: Expert Determination and Safe Harbor[1]. This guidance is intended to assist covered entities to understand what is de-identification, the general process by which de-identified information is created, and the options available for performing de-identification.

In developing this guidance, the Office for Civil Rights (OCR) solicited input from stakeholders with practical, technical and policy experience in de-identification. OCR convened stakeholders at a workshop consisting of multiple panel sessions held March 8-9, 2010, in Washington, DC. Each panel addressed a specific topic related to the Privacy Rule's de-identification methodologies and policies. The workshop was open to the public and each panel was followed by a question and answer period. More information about the workshop, including a summary, can be found at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/deidentificationworkshop2010.html. A webcast of the workshop can be viewed through streaming video from the website.

1.1. Protected Health Information

The HIPAA Privacy Rule protects most "individually identifiable health information" held or transmitted by a covered entity or its business associate, in any form or medium, whether electronic, on paper, or oral. The Privacy Rule calls this information *protected health information* (PHI).[2] Protected health information is information, including demographic information, which relates to:

- the individual's past, present, or future physical or mental health or condition,
- the provision of health care to the individual, or
- the past, present, or future payment for the provision of health care to the individual,

and that identifies the individual or for which there is a reasonable basis to believe can be used to identify the individual. Protected health information includes many common identifiers (e.g., name, address, birth date, Social Security Number) when they can be associated with the health information listed above.

---

[1] The Health Information Technology for Economic and Clinical Health (HITECH) Act was enacted as part of the American Recovery and Reinvestment Act of 2009 (ARRA). Section 13424(c) of the HITECH Act requires the Secretary of HHS to issue guidance on how best to implement the requirements for the de-identification of health information contained in the Privacy Rule.

[2] Protected health information (PHI) is defined as individually identifiable health information transmitted or maintained by a covered entity or its business associates in any form or medium (45 CFR 160.103). The definition exempts a small number of categories of individually identifiable health information, such as individually identifiable health information found in employment records held by a covered entity in its role as an employer.

For example, a medical record, laboratory report, or hospital bill would be PHI because each document would contain a patient's name and/or other identifying information associated with the health data content.

By contrast, a health plan report that only noted the average age of health plan members was 45 years would not be PHI because that information, although developed by aggregating information from individual plan member records, does not identify any individual plan members and there is no reasonable basis to believe that it could be used to identify an individual.

The relationship with health information is fundamental. Identifying information alone, such as personal names, residential addresses, or phone numbers, would not necessarily be designated as PHI. For instance, if such information was reported as part of a publicly accessible data source, such as a phone book, then this information would not be PHI because it is not related to heath data (see above). If such information was listed with health condition, health care provision or payment data, such as an indication that the individual was treated at a certain clinic, then this information would be PHI.

1.2. Covered Entities, Business Associates, and PHI

In general, the protections of the Privacy Rule apply to information held by covered entities and their business associates. HIPAA defines a covered entity as 1) a health care provider that conducts certain standard administrative and financial transactions in electronic form; 2) a health care clearinghouse; or 3) a health plan.[3] A business associate is a person or entity (other than a member of the covered entity's workforce) that performs certain functions or activities on behalf of, or provides certain services to, a covered entity that involve the use or disclosure of protected health information. A covered entity may use a business associate to de-identify PHI on its behalf only to the extent such activity is authorized by their business associate agreement.

See the OCR website http://www.hhs.gov/ocr/privacy/ for detailed information about the Privacy Rule and how it protects the privacy of health information.

1.3. De-identification and its Rationale

The increasing adoption of health information technologies in the United States accelerates their potential to facilitate beneficial studies that combine large, complex data sets from multiple sources. The process of de-identification, by which identifiers are removed from the health information, mitigates privacy risks to individuals and thereby supports the secondary use of data for comparative effectiveness studies, policy assessment, life sciences research, and other endeavors.

---

[3] Detailed definitions and explanations of these covered entities and their varying types can be found in the "Covered Entity Charts" available through the OCR website, at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/index.html. Discussion of business associates can be found at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/businessassociates.html

The Privacy Rule was designed to protect individually identifiable health information through permitting only certain uses and disclosures of PHI provided by the Rule, or as authorized by the individual subject of the information. However, in recognition of the potential utility of health information even when it is not individually identifiable, §164.502(d) of the Privacy Rule permits a covered entity or its business associate to create information that is not individually identifiable by following the de-identification standard and implementation specifications in §164.514(a)-(b). These provisions allow the entity to use and disclose information that neither identifies nor provides a reasonable basis to identify an individual.[4] As discussed below, the Privacy Rule provides two de-identification methods: 1) a formal determination by a qualified expert; or 2) the removal of specified individual identifiers as well as absence of actual knowledge by the covered entity that the remaining information could be used alone or in combination with other information to identify the individual.

Both methods, even when properly applied, yield de-identified data that retains some risk of identification. Although the risk is very small, it is not zero, and there is a possibility that de-identified data could be linked back to the identity of the patient to which it corresponds.

Regardless of the method by which de-identification is achieved, the Privacy Rule does not restrict the use or disclosure of de-identified health information, as it is no longer considered protected health information.

1.4. The De-identification Standard

Section 164.514(a) of the HIPAA Privacy Rule provides the standard for de-identification of protected health information. Under this standard, health information is not individually identifiable if it does not identify an individual and if the covered entity has no reasonable basis to believe it can be used to identify an individual.

> **§ 164.514 Other requirements relating to uses and disclosures of protected health information.**
> (a) *Standard: de-identification of protected health information.* Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.

Sections 164.514(b) and(c) of the Privacy Rule contain the implementation specifications that a covered entity must follow to meet the de-identification standard. As summarized in Figure 1, the Privacy Rule provides two methods by which health information can be designated as de-identified.

[4] In some instances, other federal protections also may apply, such as those found in Family Educational Rights and Privacy Act (FERPA) or the Common Rule.

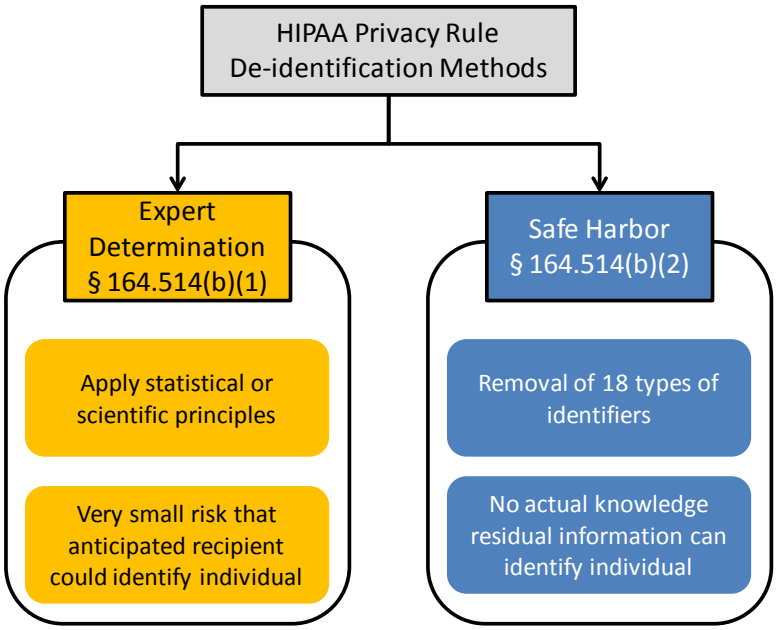

Figure 1. Two methods to achieve de-identification in accordance with the HIPAA Privacy Rule.

The first is the "Expert Determination" method:

(b) *Implementation specifications: requirements for de-identification of protected health information.* A covered entity may determine that health information is not individually identifiable health information only if:

(1) 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; or

The second is the "Safe Harbor" method:

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(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 the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units

<table>
<tr><td colspan="2">containing 20,000 or fewer people is changed to 000</td></tr>
<tr><td colspan="2">(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, 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</td></tr>
<tr><td>(D) Telephone numbers</td><td>(L) Vehicle identifiers and serial numbers, including license plate numbers</td></tr>
<tr><td>(E) Fax numbers</td><td>(M) Device identifiers and serial numbers</td></tr>
<tr><td>(F) Email addresses</td><td>(N) Web Universal Resource Locators (URLs)</td></tr>
<tr><td>(G) Social security numbers</td><td>(O) Internet Protocol (IP) addresses</td></tr>
<tr><td>(H) Medical record numbers</td><td>(P) Biometric identifiers, including finger and voice prints</td></tr>
<tr><td>(I) Health plan beneficiary numbers</td><td>(Q) Full-face photographs and any comparable images</td></tr>
<tr><td>(J) Account numbers</td><td rowspan="2">(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section “Re-identification”]; and</td></tr>
<tr><td>(K) Certificate/license numbers</td></tr>
</table>

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Satisfying either method would demonstrate that a covered entity has met the standard in §164.514(a) above. De-identified health information created following these methods is no longer protected by the Privacy Rule because it does not fall within the definition of PHI. Of course, de-identification leads to information loss which may limit the usefulness of the resulting health information in certain circumstances. As described in the forthcoming sections, covered entities may wish to select de-identification strategies that minimize such loss.

Re-identification

The implementation specifications further provide direction with respect to *re-identification*, specifically the assignment of a unique code to the set of de-identified health information to permit re-identification by the covered entity.

(c) *Implementation specifications: re-identification.* A covered entity may assign a code or other means of record identification to allow information de-identified under this section to be re-identified by the covered entity, provided that:

(1) *Derivation.* 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; and

(2) *Security.* The covered entity does not use or disclose the code or other means of record identification for any other purpose, and does not disclose the mechanism for re-identification.

If a covered entity or business associate successfully undertook an effort to identify the subject of de-identified information it maintained, the health information now related to a specific individual would again be protected by the Privacy Rule, as it would meet the definition of PHI. Disclosure of a code or other means of record identification designed to enable coded or otherwise de-identified information to be re-identified is also considered a disclosure of PHI.

1.5. Preparation for De-identification

The importance of documentation for which values in health data correspond to PHI, as well as the systems that manage PHI, for the de-identification process cannot be overstated. Esoteric notation, such as acronyms whose meaning are known to only a select few employees of a covered entity, and incomplete description may lead those overseeing a de-identification procedure to unnecessarily redact information or to fail to redact when necessary. When sufficient documentation is provided, it is straightforward to redact the appropriate fields. See section 3.10 for a more complete discussion.

In the following two sections, we address questions regarding the Expert Determination method (Section 2) and the Safe Harbor method (Section 3).

2. Guidance on Satisfying the Expert Determination Method

In §164.514(b), the Expert Determination method for de-identification is defined as follows:

> (1) 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

2.1. Have expert determinations been applied outside of the health field?

Yes. The notion of expert certification is not unique to the health care field. Professional scientists and statisticians in various fields routinely determine and accordingly mitigate risk prior to sharing data. The field of statistical disclosure limitation, for instance, has been developed within government statistical agencies, such as the Bureau of the Census, and applied to protect numerous types of data.[5]

2.2. Who is an "expert?"

There is no specific professional degree or certification program for designating who is an expert at rendering health information de-identified. Relevant expertise may be gained through various routes of education and experience. Experts may be found in the statistical, mathematical, or other scientific domains. From an enforcement perspective, OCR would review the relevant professional experience and academic or other training of the expert used by the covered entity, as well as actual experience of the expert using health information de-identification methodologies.

2.3. What is an acceptable level of identification risk for an expert determination?

There is no explicit numerical level of identification risk that is deemed to universally meet the "very small" level indicated by the method. The ability of a recipient of information to identify an individual (i.e., subject of the information) is dependent on many factors, which an expert will need to take into account while assessing the risk

[5] Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology. Report on statistical disclosure limitation methodology. *Statistical Policy Working Paper 22, Office of Management and Budget*. May 1994. Revised by the Confidentiality and Data Access Committee. 2005. Available online: http://www.fcsm.gov/working-papers/wp22.html

from a data set. This is because the risk of identification that has been determined for one particular data set in the context of a specific environment may not be appropriate for the same data set in a different environment or a different data set in the same environment. As a result, an expert will define an acceptable "very small" risk based on the ability of an anticipated recipient to identify an individual. This issue is addressed in further depth in Section 2.6.

2.4. How long is an expert determination valid for a given data set?

The Privacy Rule does not explicitly require that an expiration date be attached to the determination that a data set, or the method that generated such a data set, is de-identified information. However, experts have recognized that technology, social conditions, and the availability of information changes over time. Consequently, certain de-identification practitioners use the approach of time-limited certifications. In this sense, the expert will assess the expected change of computational capability, as well as access to various data sources, and then determine an appropriate timeframe within which the health information will be considered reasonably protected from identification of an individual.

Information that had previously been de-identified may still be adequately de-identified when the certification limit has been reached. When the certification timeframe reaches its conclusion, it does not imply that the data which has already been disseminated is no longer sufficiently protected in accordance with the de-identification standard. Covered entities will need to have an expert examine whether future releases of the data to the same recipient (e.g., monthly reporting) should be subject to additional or different de-identification processes consistent with current conditions to reach the very low risk requirement.

2.5. Can an expert derive multiple solutions from the same data set for a recipient?

Yes. Experts may design multiple solutions, each of which is tailored to the covered entity's expectations regarding information reasonably available to the anticipated recipient of the data set. In such cases, the expert must take care to ensure that the data sets cannot be combined to compromise the protections set in place through the mitigation strategy. (Of course, the expert must also reduce the risk that the data sets could be combined with prior versions of the de-identified dataset or with other publically available datasets to identify an individual.) For instance, an expert may derive one data set that contains detailed geocodes and generalized aged values (e.g., 5-year age ranges) and another data set that contains generalized geocodes (e.g., only the first two digits) and fine-grained age (e.g., days from birth). The expert may certify a covered entity to share both data sets after determining that the two data sets could not be merged to individually identify a patient. This certification may be based on a technical proof regarding the inability to merge such data sets. Alternatively, the expert also could require additional safeguards through a data use agreement.

2.6. How do experts assess the risk of identification of information?

No single universal solution addresses all privacy and identifiability issues. Rather, a combination of technical and policy procedures are often applied to the de-identification task. OCR does not require a particular process for an expert to use to reach a determination that the risk of identification is very small. However, the Rule does require that the methods and results of the analysis that justify the determination be documented and made available to OCR upon request. The following information is meant to provide covered entities with a general understanding of the de-identification process applied by an expert. It does not provide sufficient detail in statistical or scientific methods to serve as a substitute for working with an expert in de-identification.

A general workflow for expert determination is depicted in Figure 2. Stakeholder input suggests that the determination of identification risk can be a process that consists of a series of steps. First, the expert will evaluate the extent to which the health information can (or cannot) be identified by the anticipated recipients. Second, the expert often will provide guidance to the covered entity or business associate on which statistical or scientific methods can be applied to the health information to mitigate the anticipated risk. The expert will then execute such methods as deemed acceptable by the covered entity or business associate data managers, i.e., the officials responsible for the design and operations of the covered entity's information systems. Finally, the expert will evaluate the identifiability of the resulting health information to confirm that the risk is no more than very small when disclosed to the anticipated recipients. Stakeholder input suggests that a process may require several iterations until the expert and data managers agree upon an acceptable solution. Regardless of the process or methods employed, the information must meet the very small risk specification requirement.

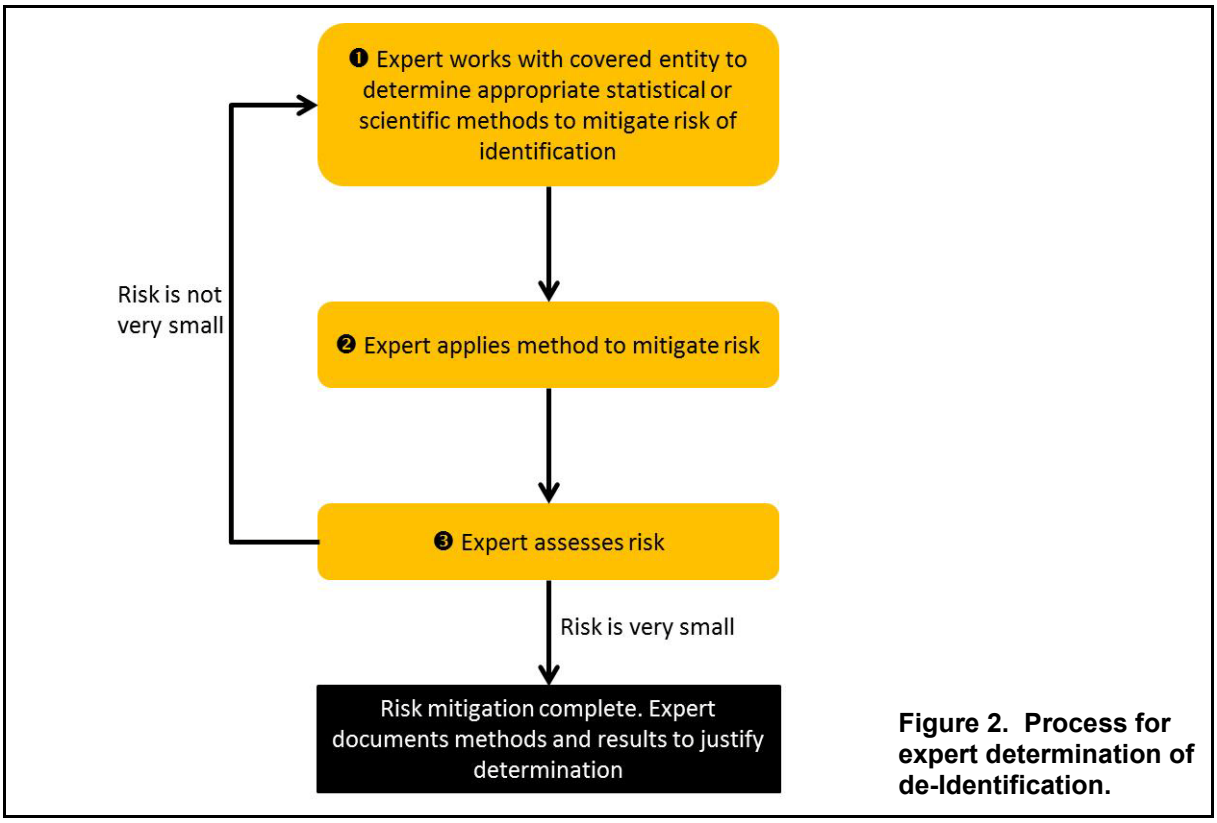

Figure 2. Process for expert determination of de-Identification.

Data managers and administrators working with an expert to consider the risk of identification of a particular set of health information can look to the principles summarized in Table 1 for assistance.[6] These principles build on those defined by the Federal Committee on Statistical Methodology (which was referenced in the original publication of the Privacy Rule).[7] The table describes principles for considering the identification risk of health information. The principles should serve as a starting point for reasoning and are not meant to serve as a definitive list. In the process, experts are advised to consider how data sources that are available to a recipient of health information (e.g., computer systems that contain information about patients) could be utilized for identification of an individual.[8]

Table 1. Principles used by experts in the determination of the identifiability of health information.

Principle	Description	Examples

[6] This table was adapted from B. Malin, D. Karp, and R. Scheuermann. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigative Medicine*. 2010; 58(1): 11-18.

[7] Supra note 3.

[8] In general, it helps to separate the "features," or types of data, into classes of relatively "high" and "low" risks. Although risk actually is more of a continuum, this rough partition illustrates how context impacts risk.

Replicability	Prioritize health information features into levels of risk according to the chance it will consistently occur in relation to the individual.	*Low:* Results of a patient's blood glucose level test will vary
		High: Demographics of a patient (e.g., birth date) are relatively stable
Data source Availability	Determine which external data sources contain the patients' identifiers and the replicable features in the health information, as well as who is permitted access to the data source.	*Low:* The results of laboratory reports are not often disclosed with identity beyond healthcare environments.
		High: Patient name and demographics are often in public data sources, such as vital records -- birth, death, and marriage registries.
Distinguishability	Determine the extent to which the subject's data can be distinguished in the health information.	*Low:* It has been estimated that the combination of *Year of Birth, Gender,* and *3-Digit ZIP Code* is unique for approximately 0.04% of residents in the United States[9]. This means that very few residents could be identified through this combination of data alone.
		High: It has been estimated that the combination of a patient's *Date of Birth, Gender,* and *5-Digit ZIP Code* is unique for over 50% of residents in the United States[10,11]. This means that over half of U.S. residents could be uniquely described just with these three data elements.
Assess Risk	The greater the replicability, availability, and distinguishability of the health information, the greater the risk for identification.	*Low:* Laboratory values may be very distinguishing, but they are rarely independently replicable and are rarely disclosed in multiple data sources to which many people have access.

[9] See L. Sweeney. Testimony before that National Center for Vital and Health Statistics Workgroup for Secondary Uses of Health information. August 23, 2007.

[10] See P. Golle. Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM Workshop on Privacy in the Electronic Society*. ACM Press, New York, NY. 2006: 77-80.

[11] See L. Sweeney. K-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*. 2002; 10(5): 557-570.

		High: Demographics are highly distinguishing, highly replicable, and are available in public data sources.

When evaluating identification risk, an expert often considers the degree to which a data set can be "linked" to a data source that reveals the identity of the corresponding individuals. Linkage is a process that requires the satisfaction of certain conditions. The first condition is that the de-identified data are unique or "distinguishing." It should be recognized, however, that the ability to distinguish data is, by itself, insufficient to compromise the corresponding patient's privacy. This is because of a second condition, which is the need for a naming data source, such as a publicly available voter registration database (see Section 2.6). Without such a data source, there is no way to definitively link the de-identified health information to the corresponding patient. Finally, for the third condition, we need a mechanism to relate the de-identified and identified data sources. Inability to design such a relational mechanism would hamper a third party's ability to achieve success to no better than random assignment of de-identified data and named individuals. The lack of a readily available naming data source does not imply that data are sufficiently protected from future identification, but it does indicate that it is harder to re-identify an individual, or group of individuals, given the data sources at hand.

Example Scenario

Imagine that a covered entity is considering sharing the information in the table to the left in Figure 3. This table is devoid of explicit identifiers, such as personal names and Social Security Numbers. The information in this table is distinguishing, such that each row is unique on the combination of demographics (i.e., *Age*, *ZIP Code*, and *Gender*). Beyond this data, there exists a voter registration data source, which contains personal names, as well as demographics (i.e., *Birthdate*, *ZIP Code*, and *Gender*), which are also distinguishing. Linkage between the records in the tables is possible through the demographics. Notice, however, that the first record in the covered entity's table is not linked because the patient is not yet old enough to vote.

Data Considered for Sharing

Age	Zip Code	Gender	Diagnosis
15	00000	Male	Diabetes
21	00001	Female	Influenza
36	10000	Male	Broken Arm
91	10001	Female	Acid Reflux

Voter Registration Records (Identified Resource)

Birthdate	Zip Code	Gender	Name
2/2/1989	00001	Female	Alice Smith
3/3/1974	10000	Male	Bob Jones
4/4/1919	10001	Female	Charlie Doe

Figure 3. Linking two data sources to identity diagnoses.

Thus, an important aspect of identification risk assessment is the route by which health information can be linked to naming sources or sensitive knowledge can be inferred. A higher risk "feature" is one that is found in many places and is publicly available. These are features that could be exploited by anyone who receives the information. For instance, patient demographics could be classified as high-risk features. In contrast,

lower risk features are those that do not appear in public records or are less readily available. For instance, clinical features, such as blood pressure, or temporal dependencies between events within a hospital (e.g., minutes between dispensation of pharmaceuticals) may uniquely characterize a patient in a hospital population, but the data sources to which such information could be linked to identify a patient are accessible to a much smaller set of people.

> *Example Scenario*
> An expert is asked to assess the identifiability of a patient's demographics. First, the expert will determine if the demographics are independently *replicable*. Features such as birth date and gender are strongly independently replicable—the individual will always have the same birth date -- whereas ZIP code of residence is less so because an individual may relocate. Second, the expert will determine which *data sources* that contain the individual's identification also contain the demographics in question. In this case, the expert may determine that public records, such as birth, death, and marriage registries, are the most likely data sources to be leveraged for identification. Third, the expert will determine if the specific information to be disclosed is *distinguishable*. At this point, the expert may determine that certain combinations of values (e.g., Asian males born in January of 1915 and living in a particular 5-digit ZIP code) are unique, whereas others (e.g., white females born in March of 1972 and living in a different 5-digit ZIP code) are never unique. Finally, the expert will determine if the data sources that could be used in the identification process are readily *accessible*, which may differ by region. For instance, voter registration registries are free in the state of North Carolina, but cost over $15,000 in the state of Wisconsin. Thus, data shared in the former state may be deemed more risky than data shared in the latter.[12]

2.7. What are the approaches by which an expert assesses the risk that health information can be identified?

The de-identification standard does not mandate a particular method for assessing risk.

A qualified expert may apply generally accepted statistical or scientific principles to compute the likelihood that a record in a data set is expected to be *unique, or linkable to only one person,* within the population to which it is being compared. Figure 4 provides a visualization of this concept.[13] This figure illustrates a situation in which the records in a data set are not a proper subset of the population for whom identified information is known. This could occur, for instance, if the data set includes patients over one year-old but the population to which it is compared includes data on people over 18 years old (e.g., registered voters).

---

[12] See K. Benitez and B. Malin. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association*. 2010; 17(2): 169-177.

[13] Figure based on Dan Barth-Jones's presentation, "Statistical de-identification: challenges and solutions" from the Workshop on the HIPAA Privacy Rule's De-Identification Standard, which was held March 8-9, 2010 in Washington, DC.

The computation of population uniques can be achieved in numerous ways, such as through the approaches outlined in published literature.[14,15] For instance, if an expert is attempting to assess if the combination of a patient's race, age, and geographic region of residence is unique, the expert may use population statistics published by the U.S. Census Bureau to assist in this estimation. In instances when population statistics are unavailable or unknown, the expert may calculate and rely on the statistics derived from the data set. This is because a record can only be linked between the data set and the population to which it is being compared if it is unique in both. Thus, by relying on the statistics derived from the data set, the expert will make a conservative estimate regarding the uniqueness of records.

> *Example Scenario*
> Imagine a covered entity has a data set in which there is one 25 year old male from a certain geographic region in the United States. In truth, there are five 25 year old males in the geographic region in question (i.e., the population). Unfortunately, there is no readily available data source to inform an expert about the number of 25 year old males in this geographic region.
>
> By inspecting the data set, it is clear to the expert that there is at least one 25 year old male in the population, but the expert does not know if there are more. So, without any additional knowledge, the expert assumes there are no more, such that the record in the data set is unique. Based on this observation, the expert recommends removing this record from the data set. In doing so, the expert has made a conservative decision with respect to the uniqueness of the record.

In the previous example, the expert provided a solution (i.e., removing a record from a dataset) to achieve de-identification, but this is one of many possible solutions that an expert could offer. In practice, an expert may provide the covered entity with multiple alternative strategies, based on scientific or statistical principles, to mitigate risk.

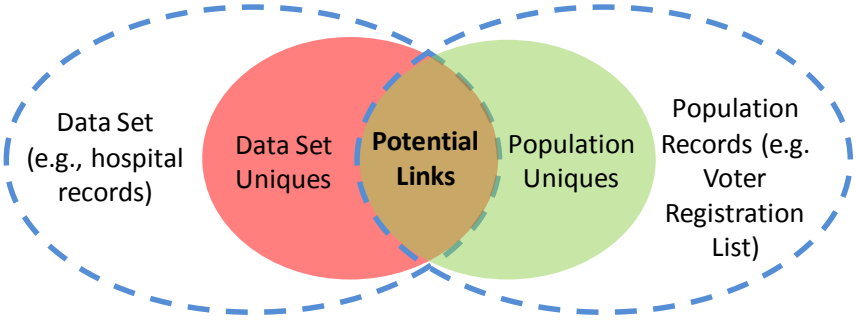

Figure 4. Relationship between uniques in the data set and the broader population, as well as the degree to which linkage can be achieved.

The expert may consider different measures of "risk," depending on the concern of the organization looking to disclose information. The expert will attempt to determine which

[14] Supra note 10.

[15] See M. Elliot, C. Skinner, and A. Dale. Special unique, random unique and sticky populations: some counterintuitive effects of geographic detail on disclosure risk. *Research in Official Statistics*. 1998; 1(2): 53-58.

record in the data set is the most vulnerable to identification. However, in certain instances, the expert may not know which particular record to be disclosed will be most vulnerable for identification purposes. In this case, the expert may attempt to compute risk from several different perspectives.

2.8. What are the approaches by which an expert mitigates the risk of identification of an individual in health information?

The Privacy Rule does not require a particular approach to mitigate, or reduce to very small, identification risk. The following provides a survey of potential approaches. An expert may find all or only one appropriate for a particular project, or may use another method entirely.

If an expert determines that the risk of identification is greater than very small, the expert may modify the information to mitigate the identification risk to that level, as required by the de-identification standard. In general, the expert will adjust certain features or values in the data to ensure that unique, identifiable elements no longer, or are not expected to, exist. Some of the methods described below have been reviewed by the Federal Committee on Statistical Methodology[16], which was referenced in the original preamble guidance to the Privacy Rule de-identification standard and recently revised.

Several broad classes of methods can be applied to protect data. An overarching common goal of such approaches is to balance disclosure risk against data utility.[17] If one approach results in very small identity disclosure risk but also a set of data with little utility, another approach can be considered. However, data utility does not determine when the de-identification standard of the Privacy Rule has been met.

Table 2 illustrates the application of such methods. In this example, we refer to columns as “features” about patients (e.g., Age and Gender) and rows as “records” of patients (e.g., the first and second rows correspond to records on two different patients).

Table 2. An example of protected health information.

Age (Years)	Gender	ZIP Code	Diagnosis
15	Male	00000	Diabetes
21	Female	00001	Influenza
36	Male	10000	Broken Arm
91	Female	10001	Acid Reflux

[16] Supra note 5.

[17] See G. Duncan, S. Keller-McNulty, and S. Lynne Stokes. Disclosure risk vs. data utility: the R-U confidentiality map as applied to topcoding. *Chance*. 2004; 3(3): 16-20.

A first class of identification risk mitigation methods corresponds to *suppression* techniques. These methods remove or eliminate certain features about the data prior to dissemination. Suppression of an entire feature may be performed if a substantial quantity of records is considered as too risky (e.g., removal of the ZIP Code feature). Suppression may also be performed on individual records, deleting records entirely if they are deemed too risky to share. This can occur when a record is clearly very distinguishing (e.g., the only individual within a county that makes over $500,000 per year). Alternatively, suppression of specific values within a record may be performed, such as when a particular value is deemed too risky (e.g., "President of the local university", or ages or ZIP codes that may be unique). Table 3 illustrates this last type of suppression by showing how specific values of features in Table 2 might be suppressed (i.e., black shaded cells).

Table 3. A version of Table 2 with suppressed patient values.

Age (Years)	Gender	ZIP Code	Diagnosis
	Male	00000	Diabetes
21	Female	00001	Influenza
36	Male		Broken Arm
	Female		Acid Reflux

A second class of methods that can be applied for risk mitigation are based on *generalization* (sometimes referred to as abbreviation) of the information. These methods transform data into more abstract representations. For instance, a five-digit ZIP Code may be generalized to a four-digit ZIP Code, which in turn may be generalized to a three-digit ZIP Code, and onward so as to disclose data with lesser degrees of granularity. Similarly, the age of a patient may be generalized from one- to five-year age groups. Table 4 illustrates how generalization (i.e., gray shaded cells) might be applied to the information in Table 2.

Table 4. A version of Table 2 with generalized patient values.

Age (Years)	Gender	ZIP Code	Diagnosis
Under 21	Male	0000*	Diabetes
Between 21 and 34	Female	0000*	Influenza
Between 35 and 44	Male	1000*	Broken Arm
45 and over	Female	1000*	Acid Reflux

A third class of methods that can be applied for risk mitigation corresponds to *perturbation*. In this case, specific values are replaced with equally specific, but different, values. For instance, a patient's age may be reported as a random value within a 5-year window of the actual age. Table 5 illustrates how perturbation (i.e., gray shaded cells) might be applied to Table 2. Notice that every age is within +/- 2 years of the original age. Similarly, the final digit in each ZIP Code is within +/- 3 of the original ZIP Code.

Table 5. A version of Table 2 with randomized patient values.

Age (Years)	Gender	ZIP Code	Diagnosis
16	Male	00002	Diabetes
20	Female	00000	Influenza
34	Male	10000	Broken Arm
93	Female	10003	Acid Reflux

In practice, perturbation is performed to maintain statistical properties about the original data, such as mean or variance.

The application of a method from one class does not necessarily preclude the application of a method from another class. For instance, it is common to apply generalization and suppression to the same data set.

Using such methods, the expert will prove that the likelihood an undesirable event (e.g., future identification of an individual) will occur is very small. For instance, one example of a data protection model that has been applied to health information is the *k*-anonymity principle.[18,19] In this model, "*k*" refers to the number of people to which each disclosed record must correspond. In practice, this correspondence is assessed using the features that could be reasonably applied by a recipient to identify a patient. Table 6 illustrates an application of generalization and suppression methods to achieve 2-anonymity with respect to the Age, Gender, and ZIP Code columns in Table 2. The first two rows (i.e., shaded light gray) and last two rows (i.e., shaded dark gray) correspond to patient records with the same combination of generalized and suppressed values for Age, Gender, and ZIP Code. Notice that Gender has been suppressed completely (i.e., black shaded cell).

Table 6, as well as a value of *k* equal to 2, is meant to serve as a simple example for illustrative purposes only. Various state and federal agencies define policies regarding small cell counts (i.e., the number of people corresponding to the same combination of features) when sharing tabular, or summary, data.[20,21,22,23,24,25,26,27] However, OCR does

---

[18] Supra note 11.

[19] See K. El Emam and F. Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*. 2008; 15(5): 627-637.

[20] Arkansas HIV/AIDS Surveillance Section. Arkansas HIV/AIDS Data Release Policy. First published: May 2010. http://www.healthy.arkansas.gov/programsServices/healthStatistics/Documents/STDSurveillance/Datadeissemination.pdf

[21] Colorado State Department of Public Health and Environment. Guidelines for working with small numbers. http://www.cdphe.state.co.us/cohid/smnumguidelines.html

[22] Iowa Department of Public Health, Division of Acute Disease Prevention and Emergency Reponse. Policy for disclosure of reportable disease information. http://www.idph.state.ia.us/adper/common/pdf/cade/disclosure_reportable_diseases.pdf

[23] R. Klein, S. Proctor, M. Boudreault, and K. Turczyn. Healthy people 2010 criteria for data suppression. Centers for Disease Control Statistical Notes Number 24. 2002.

[24] National Center for Health Statistics. Staff Manual on Confidentiality. Section 9: Avoiding inadvertent disclosures through release of microdata; Section 10: Avoiding inadvertent disclosures in tabular data. 2004.

not designate a universal value for *k* that covered entities should apply to protect health information in accordance with the de-identification standard. The value for *k* should be set at a level that is appropriate to mitigate risk of identification by the anticipated recipient of the data set.[28]

Table 6. A version of Table 2 that is 2-anonymized.

Age (years)	Gender	ZIP Code	Diagnosis
Under 30		0000*	Diabetes
Under 30		0000*	Influenza
Over 30		1000*	Broken Arm
Over 30		1000*	Acid Reflux

As can be seen, there are many different disclosure risk reduction techniques that can be applied to health information. However, it should be noted that there is no particular method that is universally the best option for every covered entity and health information set. Each method has benefits and drawbacks with respect to expected applications of the health information, which will be distinct for each covered entity and each intended recipient. The determination of which method is most appropriate for the information will be assessed by the expert on a case-by-case basis and will be guided by input of the covered entity.

Finally, as noted in the preamble to the Privacy Rule, the expert may also consider the technique of limiting distribution of records through a data use agreement or restricted access agreement in which the recipient agrees to limits on who can use or receive the data, or agrees not to attempt identification of the subjects. Of course, the specific details of such an agreement are left to the discretion of the expert and covered entity.

2.9. Can an Expert determine a code derived from PHI is de-identified?

There has been confusion about what constitutes a code and how it relates to PHI. For clarification, our guidance is similar to that provided by the National Institutes of Standards and Technology (NIST)[29], which states:

---

[25] Socioeconomic Data and Applications Center. Confidentiality issues and policies related to the utilization and dissemination of geospatial data for public health application; a report to the public health applications of earth science program, national aeronautics and space administration, science mission directorate, applied sciences program. 2005. http://www.ciesin.org/pdf/SEDAC_ConfidentialityReport.pdf

[26] Utah State Department of Health. Data release policy for Utah's IBIS-PH web-based query system, Utah Department of Health. First published: 2005. http://health.utah.gov/opha/IBIShelp/DataReleasePolicy.pdf

[27] Washington State Department of Health. Guidelines for working with small numbers. First published 2001, last updated July 2010. http://www.doh.wa.gov/Data/guidelines/SmallNumbers.htm.

[28] See K. El Emam, et al. A globally optimal *k*-anonymity method for the de-Identification of health information. *Journal of the American Medical Informatics Association*. 2009; 16(5): 670-682.

[29] E. McCallister, T. Grance, and K. Scarfone. Guide to protecting the confidentiality of personally identifiable information (pii): recommendations of the National Institute of Standards and Technology. Special Publication 800-122, National Institute of Standards and Technology. 2010.

> "*De-identified information can be re-identified (rendered distinguishable) by using a code, algorithm, or pseudonym that is assigned to individual records. The code, algorithm, or pseudonym should not be derived from other related information* about the individual, and the means of re-identification should only be known by authorized parties and not disclosed to anyone without the authority to re-identify records. A common de-identification technique for obscuring PII [Personally Identifiable Information] is to use a one-way cryptographic function, also known as a hash function, on the PII.*
>
> **This is not intended to exclude the application of cryptographic hash functions to the information."*

In line with this guidance from NIST, a covered entity may disclose codes derived from PHI as part of a de-identified data set if an expert determines that the data meets the de-identification requirements at §164.514(b)(1). The re-identification provision in §164.514(c) does not preclude the transformation of PHI into values derived by cryptographic hash functions using the expert determination method, provided the keys associated with such functions are not disclosed, including to the recipients of the de-identified information.

2.10. Must a covered entity use a data use agreement when sharing de-identified data to satisfy the Expert Determination Method?

No. The Privacy Rule does not limit how a covered entity may disclose information that has been de-identified. However, a covered entity may require the recipient of de-identified information to enter into a data use agreement to access files with known disclosure risk, such as is required for release of a *limited data set* under the Privacy Rule. This agreement may contain a number of clauses designed to protect the data, such as prohibiting re-identification.[30] Of course, the use of a data use agreement does not substitute for any of the specific requirements of the Expert Determination Method. Further information about data use agreements can be found on the OCR website.[31] Covered entities may make their own assessments whether such additional oversight is appropriate.

---

[30] For more information about data use agreements please see the following: Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology. Report on statistical disclosure limitation methodology. Statistical Policy Working Paper 22, Office of Management and Budget. May 1994. Revised by the Confidentiality and Data Access Committee. 2005. Available online: http://www.fcsm.gov/working-papers/spwp22.html.

[31] See http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/research/index.html.

3. Guidance on Satisfying the Safe Harbor Method

In §164.514(b), the Safe Harbor method for de-identification is defined as follows:

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

<table>
<tr><td colspan="2">(A) Names</td></tr>
<tr><td colspan="2">(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 the ZIP code if, according to the current publicly available data from the Bureau of the Census:
(1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
(2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000</td></tr>
<tr><td colspan="2">(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, 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</td></tr>
<tr><td>(D) Telephone numbers</td><td>(L) Vehicle identifiers and serial numbers, including license plate numbers</td></tr>
<tr><td>(E) Fax numbers</td><td>(M) Device identifiers and serial numbers</td></tr>
<tr><td>(F) Email addresses</td><td>(N) Web Universal Resource Locators (URLs)</td></tr>
<tr><td>(G) Social security numbers</td><td>(O) Internet Protocol (IP) addresses</td></tr>
<tr><td>(H) Medical record numbers</td><td>(P) Biometric identifiers, including finger and voice prints</td></tr>
<tr><td>(I) Health plan beneficiary numbers</td><td>(Q) Full-face photographs and any comparable images</td></tr>
<tr><td>(J) Account numbers</td><td rowspan="2">(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section; and</td></tr>
<tr><td>(K) Certificate/license numbers</td></tr>
</table>

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

3.1. When can ZIP codes be included in de-identified information?

Covered entities may include the first three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; or (2) the initial three digits of a ZIP code for all such geographic units

containing 20,000 or fewer people is changed to 000. This means that the initial three digits of ZIP codes may be included in de-identified information *except* when the ZIP codes contain the initial three digits listed in the Table below. In those cases, the first three digits must be listed as 000.

OCR published a final rule on August 14, 2002, that modified certain standards in the Privacy Rule. The preamble to this final rule identified the initial three digits of ZIP codes, or *ZIP code tabulation areas (ZCTAs),* that must change to 000 for release. 67 FR 53182, 53233-53234 (Aug. 14, 2002)).

Utilizing 2000 Census data, the following three-digit ZCTAs have a population of 20,000 or fewer persons. To produce a de-identified data set utilizing the safe harbor method, all records with three-digit ZIP codes corresponding to these three-digit ZCTAs must have the ZIP code changed to 000. *Covered entities should not, however, rely upon this listing or the one found in the August 14, 2002 regulation if more current data has been published*.

The 17 restricted ZIP codes are:

036	692	878
059	790	879
063	821	884
102	823	890
203	830	893
556	831	

The Department notes that these three-digit ZIP codes are based on the five-digit ZIP Code Tabulation Areas created by the Census Bureau for the 2000 Census. This new methodology also is briefly described below, as it will likely be of interest to all users of data tabulated by ZIP code. The Census Bureau will not be producing data files containing U.S. Postal Service ZIP codes either as part of the Census 2000 product series or as a post Census 2000 product. However, due to the public's interest in having statistics tabulated by ZIP code, the Census Bureau has created a new statistical area called the Zip Code Tabulation Area (ZCTA) for Census 2000. The ZCTAs were designed to overcome the operational difficulties of creating a well-defined ZIP code area by using Census blocks (and the addresses found in them) as the basis for the ZCTAs. In the past, there has been no correlation between ZIP codes and Census Bureau geography. Zip codes can cross State, place, county, census tract, block group, and census block boundaries. The geographic designations the Census Bureau uses to tabulate data are relatively stable over time. For instance, census tracts are only defined every ten years. In contrast, ZIP codes can change more frequently. Because of the ill-defined nature of ZIP code boundaries, the Census Bureau has no file (crosswalk) showing the relationship between US Census Bureau geography and U.S. Postal Service ZIP codes.

ZCTAs are generalized area representations of U.S. Postal Service (USPS) ZIP code service areas. Simply put, each one is built by aggregating the Census 2000 blocks, whose addresses use a given ZIP code, into a ZCTA which gets that ZIP code assigned as its ZCTA code. They represent the majority USPS five-digit ZIP code found in a given area. For those areas where it is difficult to determine the prevailing five-digit ZIP code, the higher-level three-digit ZIP code is used for the ZCTA code. For further information, go to: http://www.census.gov/geo/www/gazetteer/places2k.html.

The Bureau of the Census provides information regarding population density in the United States. Covered entities are expected to rely on the most current publicly available Bureau of Census data regarding ZIP codes. This information can be downloaded from, or queried at, the American Fact Finder website (http://factfinder.census.gov). As of the publication of this guidance, the information can be extracted from the detailed tables of the "Census 2000 Summary File 1 (SF 1) 100-Percent Data" files under the "Decennial Census" section of the website. The information is derived from the Decennial Census and was last updated in 2000. It is expected that the Census Bureau will make data available from the 2010 Decennial Census in the near future. This guidance will be updated when the Census makes new information available.

3.2. May parts or derivatives of any of the listed identifiers be disclosed consistent with the Safe Harbor Method?

No. For example, a data set that contained patient initials, or the last four digits of a Social Security number, would not meet the requirement of the Safe Harbor method for de-identification.

3.3. What are examples of dates that are not permitted according to the Safe Harbor Method?

Elements of dates that are not permitted for disclosure include the day, month, and any other information that is more specific than the year of an event. For instance, the date "January 1, 2009" could not be reported at this level of detail. However, it could be reported in a de-identified data set as "2009".

Many records contain dates of service or other events that imply age. Ages that are explicitly stated, or implied, as over 89 years old must be recoded as 90 or above. For example, if the patient's year of birth is 1910 and the year of healthcare service is reported as 2010, then in the de-identified data set the year of birth should be reported as "on or before 1920." Otherwise, a recipient of the data set would learn that the age of the patient is approximately 100.

3.4. Can dates associated with test measures for a patient be reported in accordance with Safe Harbor?

No. Dates associated with test measures, such as those derived from a laboratory report, are directly related to a specific individual and relate to the provision of health care. Such dates are protected health information. As a result, no element of a date (except as described in 3.3. above) may be reported to adhere to Safe Harbor.

3. 5. What constitutes "any other unique identifying number, characteristic, or code" with respect to the Safe Harbor method of the Privacy Rule?

This category corresponds to any unique features that are not explicitly enumerated in the Safe Harbor list (A-Q), but could be used to identify a particular individual. Thus, a covered entity must ensure that a data set stripped of the explicitly enumerated identifiers also does not contain any of these unique features. The following are examples of such features:

> *Identifying Number*
> There are many potential identifying numbers. For example, the preamble to the Privacy Rule at 65 FR 82462, 82712 (Dec. 28, 2000) noted that "Clinical trial record numbers are included in the general category of 'any other unique identifying number, characteristic, or code.'
>
> *Identifying Code*
> A code corresponds to a value that is derived from a non-secure encoding mechanism. For instance, a code derived from a secure hash function without a secret key (e.g., "salt") would be considered an identifying element. This is because the resulting value would be susceptible to compromise by the recipient of such data. As another example, an increasing quantity of electronic medical record and electronic prescribing systems assign and embed barcodes into patient records and their medications. These barcodes are often designed to be unique for each patient, or event in a patient's record, and thus can be easily applied for tracking purposes. See the discussion of re-identification.
>
> *Identifying Characteristic*
> A *characteristic* may be anything that distinguishes an individual and allows for identification. For example, a unique identifying characteristic could be the occupation of a patient, if it was listed in a record as "current President of State University."

Many questions have been received regarding what constitutes "any other unique identifying number, characteristic or code" in the Safe Harbor approach, §164.514(b)(2)(i)(R), above. Generally, a code or other means of record identification that is derived from PHI would have to be removed from data de-identified following the safe harbor method. To clarify what must be removed under (R), the implementation specifications at §164.514(c) provide an exception with respect to "re-identification" by the covered entity. The objective of the paragraph is to permit covered entities to assign certain types of codes or other record identification to the de-identified information so that it may be re-identified by the covered entity at some later date. Such codes or other means of record identification assigned by the covered entity are not considered direct identifiers that must be removed under (R) if the covered entity follows the directions provided in §164.514(c).

3.6. What is "actual knowledge" that the remaining information could be used either alone or in combination with other information to identify an individual who is a subject of the information?

In the context of the Safe Harbor method, actual knowledge means clear and direct knowledge that the remaining information could be used, either alone or in combination with other information, to identify an individual who is a subject of the information. This means that a covered entity has actual knowledge if it concludes that the remaining information could be used to identify the individual. The covered entity, in other words, is aware that the information is not actually de-identified information.

The following examples illustrate when a covered entity would fail to meet the "actual knowledge" provision.

Example 1: Revealing Occupation
Imagine a covered entity was aware that the occupation of a patient was listed in a record as "former president of the State University." This information in combination with almost any additional data – like age or state of residence – would clearly lead to an identification of the patient. In this example, a covered entity would not satisfy the de-identification standard by simply removing the enumerated identifiers in §164.514(b)(2)(i) because the risk of identification is of a nature and degree that a covered entity must have concluded that the information could identify the patient. Therefore, the data would not have satisfied the de-identification standard's Safe Harbor method unless the covered entity made a sufficient good faith effort to remove the ''occupation'' field from the patient record.

Example 2: Clear Familial Relation
Imagine a covered entity was aware that the anticipated recipient, a researcher who is an employee of the covered entity, had a family member in the data (e.g., spouse, parent, child, or sibling). In addition, the covered entity was aware that the data would provide sufficient context for the employee to recognize the relative. For instance, the details of a complicated series of procedures, such as a primary surgery followed by a set of follow-up surgeries and examinations, for a person of a certain age and gender, might permit the recipient to comprehend that the data pertains to his or her relative's case. In this situation, the risk of identification is of a nature and degree that the covered entity must have concluded that the recipient could clearly and directly identify the individual in the data. Therefore, the data would not have satisfied the de-identification standard's Safe Harbor method.

Example 3: Publicized Clinical Event
Rare clinical events may facilitate identification in a clear and direct manner. For instance, imagine the information in a patient record revealed that a patient gave birth to an unusually large number of children at the same time. During the year of this event, it is highly possible that this occurred for only one individual in the hospital (and perhaps the country). As a result, the event was reported in the popular media, and the covered entity was aware of this media exposure. In this case, the risk of identification is of a nature and degree that the covered entity must have concluded that the individual subject of the information could be

identified by a recipient of the data. Therefore, the data would not have satisfied the de-identification standard's Safe Harbor method.

Example 4: Knowledge of a Recipient's Ability
Imagine a covered entity was told that the anticipated recipient of the data has a table or algorithm that can be used to identify the information, or a readily available mechanism to determine a patient's identity. In this situation, the covered entity has actual knowledge because it was informed outright that the recipient can identify a patient, unless it subsequently received information confirming that the recipient does not in fact have a means to identify a patient. Therefore, the data would not have satisfied the de-identification standard's Safe Harbor method.

3.7. If a covered entity knows of specific studies about methods to re-identify health information or use de-identified health information alone or in combination with other information to identify an individual, does this necessarily mean a covered entity has *actual knowledge* under the Safe Harbor method?

No. Much has been written about the capabilities of researchers with certain analytic and quantitative capacities to combine information in particular ways to identify health information.[32,33,34,35] A covered entity may be aware of studies about methods to identify remaining information or using de-identified information alone or in combination with other information to identify an individual. However, a covered entity's mere knowledge of these studies and methods, by itself, does not mean it has "actual knowledge" that these methods would be used with the data it is disclosing. OCR does not expect a covered entity to presume such capacities of all potential recipients of de-identified data. This would not be consistent with the intent of the Safe Harbor method, which was to provide covered entities with a simple method to determine if the information is adequately de-identified.

3. 8. Must a covered entity suppress all personal names, such as physician names, from health information for it to be designated as de-identified?

No. Only names of the individuals associated with the corresponding health information (i.e., the subjects of the records) and of their relatives, employers, and household members must be suppressed. There is no explicit requirement to remove the names of

---

[32] K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk. Evaluating the risk of re-identification of patients from hospital prescription records. *Canadian Journal of Hospital Pharmacy*. 2009; 62(4): 307-319.

[33] G. Loukides, J. Denny, and B. Malin. The disclosure of diagnosis codes can breach research participants privacy. *Journal of the American Medical Informatics Association Annual Symposium*. 2010; 17(3): 322-327.

[34] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*. 2004; 37(3): 179-192.

[35] L. Sweeney. Data sharing under HIPAA: 12 years later. A presentation at the Workshop on the HIPAA Privacy Rule's De-Identification Standard. Washington, DC. March 8-9, 2010.

providers or workforce members of the covered entity or business associate. At the same time, there is also no requirement to retain such information in a de-identified data set.

Beyond the removal of names related to the patient, the covered entity would need to consider whether additional personal names contained in the data should be suppressed to meet the *actual knowledge* specification. Additionally, other laws or confidentiality concerns may support the suppression of this information.

3.9. Must a covered entity use a data use agreement when sharing de-identified data to satisfy the Safe Harbor Method?

No. The Privacy Rule does not limit how a covered entity may disclose information that has been de-identified. However, nothing prevents a covered entity from asking a recipient of de-identified information to enter into a data use agreement, such as is required for release of a *limited data set* under the Privacy Rule. This agreement may prohibit re-identification. Of course, the use of a data use agreement does not substitute for any of the specific requirements of the Safe Harbor method. Further information about data use agreements can be found on the OCR website.[36] Covered entities may make their own assessments whether such additional oversight is appropriate.

3.10. Must a covered entity remove protected health information from free text fields to satisfy the Safe Harbor Method?

PHI may exist in different types of data in a multitude of forms and formats in a covered entity. This data may reside in highly structured database tables, such as billing records. Yet, it may also be stored in a wide range of documents with less structure and written in natural language, such as discharge summaries, progress notes, and laboratory test interpretations. These documents may vary with respect to the consistency and the format employed by the covered entity.

The de-identification standard makes no distinction between data entered into standardized fields and information entered as free text (i.e., structured and unstructured text) -- an identifier listed in the Safe Harbor standard must be removed regardless of its location in a record if it is recognizable as an identifier.

Whether additional information must be removed falls under the *actual knowledge* provision; the extent to which the covered entity has actual knowledge that residual information could be used to individually identify a patient. Clinical narratives in which a physician documents the history and/or lifestyle of a patient are information rich and may provide context that readily allows for patient identification.

Medical records are comprised of a wide range of structured and unstructured (also known as “free text”) documents. In structured documents, it is relatively clear which fields contain the identifiers that must be removed following the Safe Harbor method.

---

[36] Supra note 28.

For instance, it is simple to discern when a feature is a name or a Social Security Number, provided that the fields are appropriately labeled. However, many researchers have observed that identifiers in medical information are not always clearly labeled.[37,38] As such, in some electronic health record systems it may be difficult to discern what a particular term or phrase corresponds to (e.g., is 5/97 a date or a ratio?). It also is important to document when fields are derived from the Safe Harbor listed identifiers. For instance, if a field corresponds to the first initials of names, then this derivation should be noted. De-identification is more efficient and effective when data managers explicitly document when a feature or value pertains to identifiers. Health Level 7 (HL7) and the International Standards Organization (ISO) publish best practices in documentation and standards that covered entities may consult in this process.

Example Scenario 1
The free text field of a patient's medical record notes that the patient is the Executive Vice President of the state university. The covered entity must remove this information.

Example Scenario 2
The intake notes for a new patient include the stand-alone notation, "Newark, NJ." It is not clear whether this relates to the patient's address, the location of the patient's previous health care provider, the location of the patient's recent auto collision, or some other point. The phrase may be retained in the data.

[37] D. Dorr, W. Phillips, S. Phansalkar, S. Sims, and J. Hurdle. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of Information in Medicine*. 2006; 45(3): 246-252.
[38] O. Uzuner, Y. Luo, and P. Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*. 2007; 14(5): 550-563.

4. Glossary (*note, some of these terms are paraphrased from the regulatory text; Please see the HIPAA Rules for actual definitions*)

Business Associate	A person or entity that performs certain functions or activities that involve the use or disclosure of protected health information on behalf of, or provides services to, a covered entity. A member of the covered entity's workforce is not a business associate. A covered health care provider, health plan, or health care clearinghouse can be a business associate of another covered entity.
Covered Entity	Any entity that is • a health care provider that conducts certain transactions in electronic form (called here a "covered health care provider"). • a health care clearinghouse. • a health plan.
Cryptographic Hash Function	A hash function that is designed to achieve certain security properties. Further details can be found at http://csrc.nist.gov/groups/ST/hash/
Disclosure	A "disclosure" of Protected Health Information (PHI) is the sharing of that PHI outside of a covered entity. The sharing of PHI outside of the health care component of a covered entity is a disclosure.
Hash Function	A mathematical function which takes binary data, called the message, and produces a condensed representation, called the message digest. Further details can be found at http://csrc.nist.gov/groups/ST/hash/
Health Information	Any information, whether oral or recorded in any form or medium, that: (1) Is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (2) Relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual.
Individually Identifiable Health Information	Information that is a subset of health information, including demographic information collected from an individual, and: (1) Is created or received by a health care provider, health plan, employer, or health care clearinghouse; and (2) Relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision

	of health care to the individual; and (i) That identifies the individual; or (ii) With respect to which there is a reasonable basis to believe the information can be used to identify the individual.
Protected Health Information	Individually identifiable health information: (1) Except as provided in paragraph (2) of this definition, that is: (i) Transmitted by electronic media; (ii) Maintained in electronic media; or (iii) Transmitted or maintained in any other form or medium. (2) Protected health information excludes individually identifiable health information in: (i) Education records covered by the Family Educational Rights and Privacy Act, as amended, 20 U.S.C. 1232g; (ii) Records described at 20 U.S.C. 1232g(a)(4)(B)(iv); and (iii) Employment records held by a covered entity in its role as employer.
Suppression	Withholding information in selected records from release.