February 23, 2017

Honorable Thomas E. Price, M.D.
Secretary
Department of Health and Human Services
200 Independence Avenue, S.W.
Washington, D.C. 20201

Re: Recommendations on De-identification of Protected Health Information under HIPAA

Dear Secretary Price:

This letter transmits the findings of the National Committee on Vital and Health Statistics (NCVHS) regarding the de-identification standard under the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) and makes recommendations to improve current practices. NCVHS is your advisory committee on health data, statistics, privacy, national health information policy, and the Health Insurance Portability and Accountability Act (HIPAA).

The Privacy Rule was designed to protect individuals by limiting uses and disclosures of individuals' protected health information that they have not authorized. In particular, the Privacy Rule specifies the circumstances under which covered entities may disclose de-identified information. According to the National Institute of Standards and Technology (NIST), "De-identification is a process that is applied to a dataset with the goal of preventing or limiting informational risks to individuals, protected groups, and establishments, while still allowing for the production of aggregate statistics."[1]

There are many important uses for de-identified patient information. Healthcare providers share de-identified data to enable research and evaluate care for quality improvement and cost efficiencies. Population health experts analyze large sets of de-identified data to derive insights about care effectiveness and outcomes. For example, public health departments rely on statistical and trend data to derive patterns that allow them to track the spread of disease.

The Committee held a hearing on "De-identification and HIPAA" on May 24-25, 2016, at which it heard testimony from public and private sector computer science, legal, data analytic, informatics, and privacy experts. Through its hearing and deliberations, the Committee sought to increase awareness of practices involving protected health information (PHI) and consider

[1] Garfinkel, Simson L., *De-identifying Government Datasets,* National Institute of Standards and Technology 800-188 (2d DRAFT) (Dec. 2016), p. 8, *available at* http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf> (visited Feb 23, 2017).

how well the current de-identification standard stands up in light of these practices. The Committee also sought to develop practical recommendations in areas of guidance, research, education, and useful policy change.

EXECUTIVE SUMMARY

Major Findings

De-identification is intended to protect individuals' privacy while enabling important uses for health data. Once de-identified, PHI is transformed into data no longer subject to the provisions of the HIPAA Rules. Even data properly de-identified under the Privacy Rule may carry with it some private information, and, therefore, poses some risk of re-identification, a risk that grows into the future as new datasets are released and as datasets are combined.

Among our panelists, there was general agreement that the standard for de-identification is an essential component of the Privacy Rule and has generally provided a reasonable level of protection. The challenges, however, of protecting privacy using even de-identified health information are far more complex today than when HIPAA was enacted two decades ago or when the Privacy Rule went into effect in 2003. Uses for increasingly complex data are growing exponentially as new powerful tools are able to combine data sets and extract information from large volumes of data. Expert testimony at our hearing suggested that the goals of preserving the individual's right to privacy while fully using digital information to improve health and outcomes may be on a "collision course."[2]

The Committee considered issues relating to:

- the science and practice of de-identification;
- assessing and minimizing the risk of re-identification;
- lifecycle management and stewardship of de-identified data; and
- mitigating harmful re-identification, use, or re-disclosure.

The HIPAA de-identification standard permits two approaches to de-identification: the Expert Determination method[3] and the Safe Harbor method.[4] The Expert Determination method requires a person with appropriate knowledge to make a determination that the risk of re-identifying an individual is "very small."[5] Safe Harbor, the most commonly used approach,

[2] *See* Oral testimony of Daniel Barth-Jones, Hearing before the Subcommittee on Privacy, Confidentiality & Security, NCVHS, "De-Identification and the Health Insurance Portability and Accountability Act (HIPAA)" May 24-25, 2016, Washington, DC (hereafter May 2016 Hearing), *available at* <http://www.ncvhs.hhs.gov/transcripts-minutes/transcript-of-the-may-24-2016-ncvhs-subcommittee-on-privacy-confidentiality-security-hearing/> (visited Feb. 23, 2017).

[3] *See*, 45 CFR § 164.514(b)(1).

[4] *See*, 45 CFR § 164.514(b)(2).

[5] 45 CFR § 164.514(b)(1)(i).

entails removing 18 identifying attributes and having no actual knowledge that an individual in the data set can be re-identified. These methods do not necessarily protect against future uses of data that may result in re-identifying or inferring the identity of individuals, protected groups, or establishments.

At this time, the Committee is not recommending that the current standard for de-identification be revised, but it has identified a number of actions that HHS can take to improve the way the current standard is applied. The Committee also recommends a set of actions to formalize research into how the current standard may be revised to take advantage of emerging methods to improve de-identification. Finally, the Committee urges greater focus on potential harms of misuse of de-identified data and a process to make these uses more transparent.

Summary of Recommendations

The twelve recommendations listed below are more fully described on pages 11 – 18.

Recommendation 1: At this time, HHS should reinforce the current standard with sub-regulatory guidance and the other actions outlined in these recommendations, as these will inform possible future revisions to the Privacy Rule.

Recommendation 2: HHS should develop guidance to illustrate and reinforce how the range of mechanisms in the Privacy Rule, such as data sharing agreements, business associate agreements, consent and authorization practices, encryption, security, and breach detection, are used to bolster the management of de-identified data in the protection of privacy. Particular attention should be directed at the way in which business associate agreements should address obligations regarding de-identification and the management of de-identified datasets.

Recommendation 3: HHS should establish an information clearinghouse of de-identification best practices.

Recommendation 4: HHS should develop a written competency guide with educational resources for covered entity practitioners responsible for the de-identification process.

Recommendation 5: HHS should provide guidance on policies and practices for management and disclosure of de-identified data, for assessing the risk of re-identification, and for understanding the implications of risks to individuals and to vulnerable populations.

Recommendation 6: HHS should define the minimal skills and competencies to be considered an "expert" capable of de-identifying data using the Expert Determination method.

Recommendation 7: HHS should require that covered entities and business associates, whether they use the Safe Harbor or Expert Determination method of de-identification,

maintain a description of the method used for de-identification, the assumptions used in re-identification risk assessment, and the results of the risk assessment.

Recommendation 8: HHS should use the vehicle of the Model Notice of Privacy Practices to inform individuals that their data may be de-identified and used for other purposes, and the range of downstream uses for de-identified data.

Recommendation 9: HHS should define and promulgate the responsibilities of recipients of de-identified data sets.

Recommendation 10: HHS should establish a reporting process for use by the public to express concerns about when re-identification threatens harm to individuals, protected groups, or establishments.

Recommendation 11: HHS should investigate the feasibility of requiring covered entities and business associates to track disclosures of de-identified data sets and limited data sets to provide information in response to a data subject's request for an accounting of disclosures. The disclosure obligation should include, at minimum, a summary of the de-identified data sets that include the requester's PHI.

Recommendation 12: HHS should support a research agenda on de-identification methods and on re-identification. The research agenda should include:

- **periodic testing of how well Safe Harbor is working;**
- **study of the value of applying statistical disclosure limitation techniques in concert with Safe Harbor; and**
- **techniques for evaluating risks of re-identification and inference.**

DE-IDENTIFICATION AND HIPAA

The HIPAA Privacy Rule strives to protect the privacy interests of individuals while promoting the effective operation of the health system and the broader health and wellness goals of society. The de-identification provisions of the Privacy Rule are one way to enable data sharing among covered entities, business associates, and entities that are not subject to the Privacy Rule. Once de-identified, PHI is transformed into data no longer subject to the provisions of the HIPAA Rules. Even data properly de-identified under the Privacy Rule may carry with it some private information, and, therefore, poses some risk of re-identification, a risk that grows into the future as new datasets are released and as datasets are combined. Re-identified data are not protected by the HIPAA Privacy or Security Rules, unless they are held by, or come into the custody of, a covered entity. While other federal or state law may address the obligations of non-covered entity data holders, risks particular to personal health information may not be clear, and protections may not be adequate.

Today, downstream uses for de-identified data have expanded exponentially, with many new uses driving innovation and contributing to our understanding of health and wellness. Data aggregators, analytics companies, and health application businesses collect, handle, analyze, and re-disclose de-identified health information. In light of this growing range of uses of data that are not covered by HIPAA and the risk of re-identification, the efficacy of methods for de-identification and the way that de-identified data is managed becomes central in upholding the privacy interests of individuals and protected groups. In recognition of both the growing importance of de-identification within the US Government and the paucity of efforts addressing de-identification as a holistic field, NIST began research in this area in 2015.[6] The Committee also notes that many types of organizations look to the de-identification standard of the HIPAA Privacy Rule in the absence of their own specific standard when they have a business need to de-identify health information, including public health agencies, researchers, and others.

The HIPAA Privacy Rule establishes two approaches for de-identifying PHI: The Expert Determination method[7] and the Safe Harbor method.[8] The Expert Determination method calls for "a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" to apply such principles and methods, to determine that the risk of re-identification is very small, and to document the methods and results of the analysis that justifies the determination.

However, the Privacy Rule does not establish minimum qualifications for a person to be considered an expert. The hearing revealed that there is no consensus on what qualifies one as an expert, what constitutes best statistical and scientific methods, nor what it means for a risk

[6] Garfinkel, Simson L., *De-Identification of Personal Information*, National Institute of Standards and Technology Internal Report 8053 (Oct. 2015). This paper is available free of charge at <http://dx.doi.org/10.6028/NIST.IR.8053> (visited Feb. 23, 2017).

[7] *See*, 45 CFR. § 164.51 (b)(1).

[8] *See* 45 CFR. § 164.514 (b)(2).

of re-identification to be "very small." The Office for Civil Rights (OCR) has not issued clarifying guidance on these matters.

Further, organizations that use experts are not required to disclose to the record subjects that an expert determination has been made, nor what that determination concluded, (although they are required to document the methods and analysis that justify the determination[9]). The Committee heard testimony that there is a shortage of experts in these methods available for hire considering the growing demand for sound de-identification of PHI.

The Safe Harbor method of de-identification allows a covered entity (or a business associate, in accordance with an authorization to do so in its business associate agreement) to consider a data set de-identified if a predetermined set of 18 data elements are removed (e.g. names, geographic subdivisions smaller than a state, dates, telephone numbers, account numbers, Internet Protocol (IP) addresses).One of the 18 Safe Harbor data elements is "any other unique identifying number, characteristic, or code,"[10] thus reinforcing the need for critical analysis of a particular dataset, a step that testimony revealed is generally weak or lacking. Nevertheless, Safe Harbor remains the most commonly used method for de-identification.

The Safe Harbor method further requires that the covered entity (or business associate) not "have actual knowledge that the [remaining] information could be used alone or in combination with other information to identify an individual who is the subject of the information."[11] Specific guidance from OCR makes it clear that this requirement does not refer to having general knowledge of studies and methods about re-identification and risks, but, rather, it refers to specific knowledge relating to the particular dataset in question.[12] The Committee heard testimony that this requirement to have no "actual knowledge" that the remaining information could be used to identify an individual is not well-understood or adhered to in current practice. Such representation is increasingly difficult because of growing downstream uses that may involve combining datasets, a future action that could dramatically change a risk assessment made at the time of de-identification.

The Privacy Rule permits a covered entity to assign a code or other means of record identification to allow de-identified information to be re-identified by the covered entity provided certain protections are in place. Of course, once re-identified, the information in the hands of a covered entity is again subject to the HIPAA Privacy and Security Rules. Re-identification by the covered entity is important for uses such as precision medicine that target

---

[9] *See* 45 CFR § 164.514(b)(1)(ii).

[10] 45 CFR. § 164.514(b)(2)(i)(R).

[11] 45 CFR. § 164.514(b)(2)(ii).

[12] Office for Civil Rights, U.S. Dept. of Health and Human Services., *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, available at* <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html> (§§ 3.6 and 3.7 specifically address "actual knowledge")(visited Feb. 23, 2017).

treatments to individuals based on specific genetic markers, environmental factors, or other characteristics that are derived from application of analytics to de-identified data sets. The Committee learned, however, that technology limitations make it difficult for covered entities to re-import into their operational systems or use previously de-identified data for a given patient.

In 2010, HHS hosted a workshop on de-identification and synthesized the input from panelists into an extensive guidance document released in 2012.[13] This is the most current guidance issued by the Department. While only five years old, the volume of data, range of its uses, and increasing sophistication of tools to analyze it, calls for regular, periodic review to extend and enhance the guidance.

Consumers are rarely aware of when their data are being de-identified for a new purpose. Unlike disclosures of PHI, where the Privacy Rule requires that individuals have the right to an accounting of disclosures of their PHI to third parties with certain exceptions, the subjects of de-identified datasets may not know how often their data are disclosed in de-identified form as de-identified data are not subject to the HIPAA Privacy Rule. Even disclosure of a "limited data set," a data set with most, but not all, of the Safe Harbor identifiers removed, is not subject to the requirement for a covered entity to maintain an "accounting of disclosures," permitting an individual, on request, to obtain a list of recipients of their protected health information.

Moreover, it's increasingly common for consumers' medical record data to be combined with non-health-related data, increasing the risk of re-identification. Data analysts are now able to augment health data with geographic, socio-economic, and other public and private information to gain new scientific insight or advance a commercial goal.

CURRENT STATE OF DE-IDENTIFICATION

The challenges of preserving privacy of health information have increased over the past decade for the reasons touched upon earlier. De-identification is at the heart of the debate about how to preserve the individual's right to privacy and derive benefit and value from the use of digital information. De-identifying data affords a significant degree of privacy protection and public surveys indicated continued public support for this method. However, some thought leaders have questioned the effectiveness of de-identification in recent years because there is a growing understanding that the ability to re-identify, which now requires extensive expertise, has the potential to become more widely accessible.

The following discussion of current state issues is organized in the following four areas:

- The science and practice of de-identification
- Assessing and mitigating risk of re-identification
- Lifecycle management and stewardship of de-identified data

[13] *Id.*

- Mitigating harmful re-identification, use, or re-disclosure

The science and practice of de-identification

While the Committee heard testimony on de-identification approaches used by experts, the Hearing was not designed to delve into the mathematical or algorithmic science or tools for de-identification. The testimony reinforced, however, several important tenets about de-identification methods and practice:

First, it is never perfect. Some de-identified datasets contain data elements that, directly or indirectly, can be used to identify individuals, protected groups, or establishments.

Second, de-identification is a temporary, rather than a permanent state. A new dataset may become available, which, when compared, combined, or linked with previously available data, unlocks an identity key permitting re-identification.

Third, different de-identification methods produce different results even when applied to the same dataset. The Safe Harbor may be the more common approach, but it is not standardized, so that applied by different practitioners, may produce different results. Expert Determination uses a range of methods based on analysis of the characteristics of the dataset, each of which may produce a different result that can be considered de-identified.

Finally, de-identification reduces the quality and utility of data, the consequence of which must be judged against the characteristics of the dataset and the intended uses.

In comparing the two methods of de-identification established in the de-identification standard of the privacy Rule, Safe Harbor is largely ”one size fits all,” regardless of the characteristics of the dataset. By contrast, the Expert Determination method has the advantage of fitting the de-identification method to the risks associated with the specific dataset. Despite this increasingly important advantage, Expert Determination is used less frequently than Safe Harbor. One reason is that Expert Determination, while more consultative, is also more expensive, and there are too few experts available for hire. Some of the May 2016 hearing participants called for guidelines and standards for use of the Expert Determination method including transparency regarding methods used and results achieved and minimal standards of competencies and qualification to be an expert.

Whether choosing Expert Determination or Safe Harbor, different categories of information in a medical record present different de-identification challenges. Highly structured information such as most laboratory test results or most ICD codes[14] present lower risks of re-identification because they intrinsically carry no direct or indirect identifiers. Demographic and

[14] International Statistical Classification of Diseases and Related Health Problems (ICD) is a medical classification list developed by the World Health Organization (WHO). Information about ICD codes and the ICD standard may be found at the website of the Centers for Disease Control and Prevention, “Classification of Diseases, Functioning, and Disability,” *available at* <https://www.cdc.gov/nchs/icd/index.htm> (visited Feb. 23, 2017).

socioeconomic information present higher risks, and the Safe Harbor method targets removal of 18 direct and indirect identifiers to avoid revealing this type of information or allowing it to be linked to identifiable information. Unstructured narrative information (e.g. medical histories, descriptions of physical exams, discharge summaries, progress and consulting reports) comprises the majority of information in a medical record, presenting the greatest challenge for de-identification. While promising techniques using rule-based natural language processing have potential to de-identify narrative information to an acceptable level of risk, these are not yet scalable. Genomic data and other specialized test results, while not the specific focus for this hearing, also present unique de-identification challenges.[15]

Given the importance of de-identification to protecting privacy in the health sector, there is a lack of directed and ongoing research on the efficacy of de-identification methods. For example, one panel discussed the value of and impact on risk reduction of adding a wider range of Statistical Disclosure Limitation methods and techniques to Safe Harbor.[16] The lessons from de-identification research are not informing day-to-day practice. Practitioners responsible for de-identification and assessing risk of re-identification in non-research settings are often not adequately trained to apply critically the latest methods and research findings.

Assessing and minimizing the risk of re-identification

Just as there is a science of de-identification, there is also a growing science of re-identification that needs greater illumination. There are economic drivers for re-identification of health data to create enhanced datasets that make it an increasingly important topic. For example, combining healthcare service patterns and personal web search patterns may be valuable for marketing products and services. The probability of re-identification often focuses on the likelihood that an "intruder" would be interested in the particular dataset. Like many assumptions in this fast-changing landscape, current assessments of the types or level of interest, or the possibility for exploitation, may be poor markers for the future.

The drive to create longitudinal databases illustrates this point. A dataset with diabetes or blood pressure readings from a single visit to the physician or a single hospital stay is of less value and therefore presents less risk for re-identification, than a longitudinal database tracking diabetes or hypertension patients over time, linking their clinical paths and treatments to socioeconomic, lifestyle or employment data.

Anticipating the risk and estimating the likelihood of re-identification under certain circumstances are important considerations in determining how best to de-identify a data set. Covered entities and business associates are obligated to document that they have considered the risks to, and likelihood of re-identification of individuals in determining the data content to

[15] *See*, Testimony of Rubenstein, Malin, and Barth-Jones, *Panel I – Policy Interpretations of HIPAA's De-identification Guidance*, May 2016 Hearing.

[16] Prepared Statement of Ira Rubinstein, May 2016 Hearing, at 3, *available at* <http://www.ncvhs.hhs.gov/wp-content/uploads/2016/04/RUBENSTEIN.pdf> (visited Feb. 23, 2017).

be de-identified before release of a dataset. This assessment should inform what additional steps can be taken to safeguard information including, for example, stipulations to be included in data use agreements.

In addition to considering re-identification risks, experts who testified before the Committee suggested that inference risks should also be considered. Inference risks are the potential for others to learn about individuals from the inclusion of their information in a dataset, or from their membership in, or association, or perceived association, with the group studied, even if the individual's actual data was not included in the data set.

Re-identification, whether it produces harm or diminishes trust, is a topic that demands more attention than recently has been given in the health sector. Starting with a workable definition, greater focus would expand our understanding of the risks and opportunities to mitigate them. In addition to improved training in de-identification methods, practitioners responsible for de-identification, data scientists, policy analysts, and researchers need training in understanding the context in which a data set was produced and the attendant risks of re-identification or inappropriate inference. A risk-based approach to data release policy is contextual and contingent on the specific use.

NIST identified seven variables to consider in assessing risk: data volume, data sensitivity, type of data recipient, data use, data treatment technique, data access controls, and consent and consumer expectations.[17] HHS offers a set of principles in its de-identification guidance limited to characteristics of the data that experts use in assessing the risk of identification.[18] Data custodians could benefit from expanded principles and illustrations of how to assess both the data set and the context when determining how best to de-identify a particular dataset. For example, covered entities and business associates might consider intended uses or the security and access controls used by recipients of a particular de-identified data set, in addition to considering the attributes of the data set.

Lifecycle management of de-identified data

Information should be managed across its lifecycle from data origination to archive or destruction. De-identification practices are part of a covered entity's or business associate's information disclosure management process, essential for protecting the privacy of Protected Health Information (PHI) and mitigating associated risk. The current regulatory paradigm for de-identified data permitting data holders to release and forget may no longer be sufficient. The

[17] National Institute of Standards and Technology, Joint Task Force Transformation Initiative, Special Publication 800-30, *Guide for Conducting Risk Assessments*, Sept. 2012, *available at* <http://dx.doi.org/10.6028/NIST.SP.800-30r1> (Feb. 23, 2017).

[18] *See* Office for Civil Rights, U.S. Dept. of Health and Human Services., Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, (Nov. 26, 2012), at pp. 12-15, *available at* <https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf> (visited Feb. 23, 2017).

covered entity or business associate data holder may achieve a level of de-identification that carries low risk, but a downstream user may change that risk level by adding or combining with new data. Currently the covered entity has no obligation for downstream uses or inadvertent or malicious re-identification or re-disclosure.

Similar to managing data privacy or security more generally, the Committee heard testimony calling for de-identification to be handled as part of a reasonable management process that spans the lifecycle of information. Managing de-identification requires people, policy, technology, and governance practices. Specific lifecycle sub-processes include acquiring and assessing the data to be shared, analyzing the risk of disclosure, limiting the data to be released to that which is minimally necessary, using the most effective methods and technology for de-identification, and developing and putting in place a monitoring, accountability, and breach response plan.

De-identification, like other aspects of information and disclosure management, demands an oversight and governance process to ensure that the policy and processes are reasonable and are consistently carried out. NIST recommends that government agencies set up a Disclosure Review Board or Data Release Board to set and oversee organization-wide policy.[19] This recommendation is also applicable to covered entities and business associates who should set policies and practices for disclosure review and release management as part of the organization's information governance authority.

Additional management steps may include limiting reuse or downstream re-identification by contract.

Mitigating harmful re-identification, use, or re-disclosure

De-identified data is not subject to HIPAA. So, while other laws may apply to adjudicate harmful use, there are no penalties or processes for remediation under HIPAA for actions by data holders that harm data subjects, protected groups, or other entities attributable to re-identification, use, or re-disclosure.

The Committee heard testimony addressing whether restrictions or penalties should be imposed on data holders and if so, what types. A three-staged approach emerged that has potential to support progress and innovation while protecting data subjects and others from harm. The first stage calls for far greater transparency about the actual uses being made of de-identified health information. Disclosure about uses would support greater public dialogue that in turn could help shape sensible policy, practice, new technology, and law. Of particular interest is disclosure when de-identified data sets are merged to create longitudinal and enhanced new data that change the re-disclosure risk assessment.

[19] Garfinkel, *De-identifying Government Datasets*, p 30.

Stage two entails the development of a matrix of re-identification and inference risks and harms to data subjects, protected groups, or other entities. There is a broad range of potential harms, but it is important to begin defining those for which specific remediation is needed. Again, such a matrix could help to shape policy, practice, technology, and law.

The third stage involves regulatory action to set forth a complaint process which may lead to sanctions and penalties when data holders fail to conduct and document a risk assessment at or below an acceptable level of risk, their actions cause certain harms, or fail to take reasonable steps to protect data.

RECOMMENDATIONS

De-identification is the enabling bridge transforming patient data from PHI under HIPAA to data that is no longer covered by the main federal Privacy or Security Rules that apply to health information. While long used by medical and health system researchers, the use of de-identified health data has expanded significantly in recent years across public and private health delivery, health plans, and the life sciences. In just a few short years, a young but robust commercial health data industry has emerged. Done well, de-identification can protect the confidentiality of data subjects while promoting the use of data for a variety of public and private uses such as monitoring health system performance, enabling population health initiatives, or advancing payment policy. Done poorly, de-identification can expose individuals, protected groups, and establishments to risk of harm to physical well-being, personal dignity, reputation, or financial position.

For the most part, the current HIPAA de-identification standard remains useful and relevant. The standard has been in place for over a decade, and there is little evidence that widespread re-identification or inappropriate inference is actually harming data subjects. However, the threats are increasing rapidly, and there are weaknesses that can be strengthened with the Department's leadership. The Committee underscores the need for attention to these recommendations.

There are multiple significant risks that the de-identification standard does not address. For example, first, the de-identification standard is too often executed with inadequate attention to the unique characteristics of the dataset to which it is applied and its intended uses. Second, data subjects have little information about how their data are used or about the risk of re-identification or inference. Third, the overemphasis on de-identification ignores the management, governance and other practices and processes that must also be in place to minimize the risk of data sharing. Finally, there is no mechanism to impose penalties for harmful uses of de-identified data.

There is a body of de-identification research and a network of experts across the computer science, legal, data analytic, informatics, and privacy fields. There are also valuable policy resources to guide a path forward.

The following recommendations are designed to address shortcomings identified by the Committee and improve practices associated with de-identification. They are intended as short term and practical actions that can impact the current state of de-identification of health information.

Recommendation 1 reiterates the importance of the de-identification standard as an integral part of the HIPAA Privacy Rule. Recommendations 2 through 5 address shortcomings in current de-identification practice as carried out by many covered entities and business associates. The Committee believes that greater emphasis on education and training along with an expanded range of resources could elevate the standard of practice and compliance with the Privacy Rule. Recommendation 6 addresses the availability of de-identification experts and their minimum qualifications.

Recommendations 7 through 11 address improvements to transparency and accountability for steps taken to de-identify data, mitigate re-identification risks, and reduce harm. As discussed earlier, this group of actions forms a strong foundation that could be enhanced in the future. Recommendation 12 sets out subjects for an expanded research agenda that would help promote and advance the science of de-identification.

> **Recommendation 1: At this time, HHS should reinforce the current standard with sub-regulatory guidance and the other actions outlined in these recommendations, as these will inform possible future revisions to the Privacy Rule.**

Concerns about the potential for re-identification of data are real and they are increasing. As technology for re-identification evolves, this part of HIPAA will continue to be under pressure. At the same time, de-identification does provide a significant degree of protection and survey data show that most Americans derive a considerable comfort from having data de-identified even though many are coming to realize that it is not a complete protection.

The Committee believes that there are a number of practical actions that HHS can take now to strengthen how the standard is being implemented. The Committee also believes that HHS should put in place a process to regularly and systematically monitor the status of implementation, evolving technology, and relevant research to determine how best to keep this standard current and relevant.

> **Recommendation 2: HHS should develop guidance to illustrate and reinforce how the range of mechanisms in the Privacy Rule, such as data sharing agreements, business associate agreements, consent and authorization practices, encryption, security, and breach detection, are used to bolster the management of de-identified data in the protection of privacy. Particular attention should be directed at the way in which business associate agreements should address obligations regarding de-identification and the management of de-identified datasets.**

De-identification is not a stand-alone process. Covered entities have a range of tools that can be leveraged to help safeguard against re-identification, and this recommendation is designed to illustrate when and how mechanisms such as data use agreements and business associate agreements can and should be used in conjunction with proper de-identification to limit uses that increase the risk of re-identification, limit re-disclosure of de-identified data, and require proper management, including security, for de-identified data.

Recommendation 3: HHS should establish an information clearinghouse of de-identification best practices.

Most covered entities do not have access to experts in statistical and scientific methods and approaches for de-identification. It is important that they have the opportunity to learn from others with strong practices. De-identification practices are not static and practitioners need ready access to resources on evolving practice. A clearinghouse of de-identification case studies would help covered entities and business associates assess the adequacy of their own practices as they learn from others.

Recommendation 4: HHS should develop a written competency guide with educational resources for covered entity practitioners responsible for the de-identification process.

A competency guide would be a resource to define, for both covered entities and de-identification practitioners, the capabilities, proficiencies, and aptitudes that practitioners should have in order to properly apply de-identification methods to PHI, and assess the risk of its re-identification. Covered entities would be better informed when hiring an expert, and practitioners or prospective practitioners would know what it takes to succeed. This recommendation calls on HHS to develop a guide and promulgate the competencies and resources that can help practitioners accelerate learning to elevate practice. Further, a written competency guide could be a first step in defining the competencies necessary for any voluntary certification program.

Recommendation 5: HHS should provide guidance on policies and practices for management and disclosure of de-identified data, for assessing the risk of re-identification, and for understanding the implications of risks to individuals and to vulnerable populations.

This recommendation addresses the need for guidance and resource tools to improve the assessment of re-identification risk by covered entities and business associates so they may carry out effectively the obligations under the Privacy Rule. As noted earlier, this is an area of weak compliance. This recommendation also calls for guidance for covered entities and business associates regarding managing the sharing and release of de-identified data as part of an organization's information governance authority.

Covered entities should carefully delineate business associate responsibilities regarding de-identification and de-identified data and take greater care to monitor that the obligations are being carried out in accordance with the contractual agreement. Covered entities should also require that data custodians such as registries use sound practices and technologies for access control, security, storage and archive. The use of Certificates of Confidentiality[20] such as those required for data sharing research projects may also outline additional precautions to safeguard de-identified information.[21]

Recommendation 6: HHS should define the minimal skills and competencies to be considered an "expert" capable of de-identifying data using the Expert Determination method.

As noted earlier, expertise is developed through a variety of educational and experiential channels. While it is not possible to say, for example, that an expert must have a certain degree or years of experience, there is currently no consensus on a defined set of minimal competencies that an expert should possess to be qualified to make an expert determination. This recommendation relates specifically to the need to describe this expertise more fully so that covered entities and business associates engaging experts can be more assured of hiring a competent, qualified professional. Done well, it may lead those with the interest and statistical, mathematical, or analytic skills to advance their expertise in de-identification and risk assessment techniques as applied to health information. Given the current shortage of experts, this would be highly advantageous.

Recommendation 7: HHS should require that covered entities and business associates, whether they use the Safe Harbor or Expert Determination method of de-identification, maintain a description of the method used for de-identification, the assumptions used in re-identification risk assessment, and the results of the risk assessment.

The current de-identification standard requires that when employing the Expert Determination method, an agency document "the methods and results of the analysis that justify [the] determination."[22] This should be expanded to the Safe Harbor method in order to demonstrate compliance and support the ability to judge the adequacy of work performed. Such an expansion would ensure that there is greater transparency and accountability, and the availability of the documentation would contribute to organizational learning.

---

[20] Office of Extramural Research, National Institutes of Health, U.S. Dept. of Health and Human Services, Certificates of Confidentiality, *available at* https://humansubjects.nih.gov/coc/index (visited Feb. 6, 2017).

[21] *See,* National Institutes of Health, U.S. Dept. of Health and Human Services, Genomic Data Sharing Policy, *available at* <https://gds.nih.gov/03policy2.html> (visited Feb. 6, 2017). For a PDF copy of this policy, visit <https://gds.nih.gov/PDF/NIH_GDS_Policy.pdf>.

[22] 45 C.F.R. § 164.514(b)(1)(ii).

Recommendation 8: HHS should use the vehicle of the Model Notice of Privacy Practices to inform individuals that their information may be de-identified and used for other purposes, and the range of downstream uses for de-identified data.

The Office for Civil Rights, in collaboration with the Office of the National Coordinator for Health IT, recently released a Model Notice of Privacy Practices incorporating the 2013 updates to the HIPAA Privacy and Security Rules. The Model is silent on de-identification. While de-identified data falls outside of HIPAA, it is reasonable to use the Notice of Privacy Practices to create greater transparency for individuals about how their health information in de-identified form is used and the steps required by HIPAA to maintain its confidentiality.

Recommendation 9: HHS should define and promulgate the responsibilities of recipients of de-identified data sets.

HHS should define and describe the stewardship responsibilities, based on principles of fair information practices, of those who hold de-identified data including responsibilities relating to security, lifecycle management, and protection from re-identification. These responsibilities should be widely disseminated to raise awareness about the risks and potential consequences of misuse — including re-identification — to individuals, protected groups, and establishments.

Recommendation 10: HHS should establish a reporting process for use by the public to express concerns about when re-identification threatens harm to individuals, protected groups, or establishments.

As part of its monitoring of the issues relating to de-identification, HHS needs a way to gather the concerns of individuals, protected groups, or establishments regarding the use of de-identified data or risks to their privacy rights through re-identification or inappropriate inference. This recommendation calls for a reporting process for information gathering, rather than a complaint process in which each complaint must be formally investigated. It would be helpful to have this stream of information to monitor concerns and changing circumstances to guide the evolution of policy and continued consumer education.

Recommendation 11: HHS should investigate the feasibility of requiring covered entities and business associates to track disclosures of de-identified data sets and limited data sets to provide information in response to a data subject's request for an accounting of disclosures. The disclosure obligation should include, at minimum, a summary of the de-identified data sets that include the requester's PHI.

HIPAA ensures that individuals have the right to an accounting of disclosures of their PHI to third parties with certain exceptions. The current accounting for disclosures guidance from HHS does not reflect changes to HIPAA under the 2013 Omnibus Rule, and should be a priority for update and revision. This recommendation is in line with Recommendation 8 calling for acknowledgement of practices with respect to de-identification in the Notice of Privacy

Practices. Greater transparency of these practices is aligned with the growing public expectation that individuals are informed of how their data is being used.[23]

Recommendation 12: HHS should support a research agenda on de-identification methods and on re-identification. The research agenda should include:

- **periodic testing of how well Safe Harbor is working;**
- **study of the value of applying statistical disclosure limitation techniques in concert with Safe Harbor; and**
- **techniques for evaluating risks of re-identification and inference.**

Given the many important uses for de-identified data, the range of approaches and technologies for de-identification, and the real risk of re-identification as more data become available, a funded research agenda is vital for continual improvement in practice. Such a research agenda would include investigation into the technology to support the functionality of electronic health records to access previously de-identified data. In addition to usual channels for publication of research results, findings with implications for practice can be incorporated into the education and training channels recommended above.

The Department has just recognized the 20-year anniversary of the HIPAA law. The Privacy Rule's de-identification standard is an essential foundation for the rapid advancements to a 21st century information-driven learning health system. Now is the time to strengthen how this standard is implemented and create the environment that will allow it to evolve and take full advantage of new learning to keep pace with the rapidly changing data environment. The NCVHS looks forward to discussing the recommendations and perspectives laid out in this letter with you and HHS staff, and to working with the Department to shape future guidance and priorities for advancing this work.

Sincerely,
/s/
William W. Stead, M.D., Chair
National Committee on Vital and Health Statistics

Cc: HHS Data Council Co-Chairs

[23] Changes to the recently updated "Common Rule," reflect this expectation and broadening practice. *See* Final Rule: Federal Policy for the Protection of Human Subjects, 82 FED. REG. 7149 (Jan 19, 2017), *available at* <https://www.gpo.gov/fdsys/pkg/FR-2017-01-19/pdf/2017-01058.pdf> (visited Feb. 23, 2017).