



De-Identification Framework

A Consistent, Managed Methodology for the De-Identification of Personal Data
and the Sharing of Compliance and Risk Information

Contents

Preface	3
Introduction.....	4
Defining Categories of Health Information.....	5
Data Definitions	5
Patient Identifiable	6
Patient De-Identified	7
Patient Non-Identifiable	8
Evaluating De-Identification Methodologies	9
Background	9
Program Methodology	10
De-Identification Methodology	12
Section Summary	15
Scoring	15
Methods of De-Identification	15
Two Methods of De-Identification	15
Who has appropriate knowledge and experience?.....	16
Re-Identification Risk	20
Categories of Variables	20
Directly Identifying Variables	21
Indirectly Identifying Variables (quasi-identifiers)	21
Other Variables	21
Measuring Re-identification Probability	21
The Risk of Re-identification for Direct Identifiers	21
The Risk of Re-identification for Quasi-identifiers	22
Equivalence Classes	22
Taking Context into Account	23
What is Acceptable Risk?	23
How to Manage Re-identification Risk.....	24
Conclusion	26
Appendix A: De-Identification—CSF Mapping	27
Appendix B: Scoring	36

Preface

The HITRUST De-Identification Framework™ provides a consistent, managed methodology for the de-identification of personal data and the sharing of compliance and risk information amongst entities and their key stakeholders. However, to provide the level of consistency and repeatability needed for an industry-accepted framework, organizations must fully understand the methodology when de-identifying data in order to apply it appropriately and share de-identified data among entities.

The HITRUST De-Identification Framework:

- Defines the multiple levels of anonymization and de-identification and recommends specific use cases for each variant.
- Develops criteria for evaluating de-identification methodologies, estimating re-identification risks and criteria for certifying expertise in these methodologies.
- Creates a framework for mitigating the risks associated with the use, storage and maintenance of de-identified data, in conjunction with the HITRUST Common Security Framework (“CSF”). The controls will create a baseline security framework for de-identified data and will include controls to mitigate re-identification risks.
- Provides an appendix on risk treatments, which includes some of HITRUST’s views on transference, avoidance, mitigation and acceptance with explanatory examples, including corrective action plan (CAP) prioritization and alternate control risk analysis that will be incorporated into the CSF.

Users of this Framework also can benefit from a basic level of knowledge about de-identification, information security and privacy, risk management and risk analysis, up to and including knowledge commensurate with holders of the International Information Systems Security Certification Consortium (ISC)²™ Health Care Information Security and Privacy Practitioner (HCISPP) certification or the HITRUST Certified CSF Practitioner (CCSFP™) credential. In addition, users of this Framework also would benefit from reviewing the following documentation prior to using this Framework:

- HITRUST RMF Whitepaper
- HITRUST CSF Assessment Methodology
- HITRUST CSF Assurance Program Requirements

Readers may also benefit from reviewing other CSF assessment-related information:

- Comparing the CSF, ISO/IEC 27001 and NIST SP 800-53
- HITRUST MyCSF™ datasheets

Introduction

The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule contemplates various categories of health information: fully-identifiable health data (known as protected health information or PHI), limited data sets (a HIPAA specific term meaning PHI where most direct patient identifiers have been removed), and de-identified data where there is a low risk of re-identifying a particular patient. The healthcare industry also deals with other kinds of health information, including health care data with no personal component, aggregated data, where personal identifiers have been blended into a composite analysis, and commercial data. The core HIPAA privacy framework - involving PHI, limited data sets and de-identified information - is a framework that has proven very important in the categorization of health information and has effectively allowed for use of personal data with reasonable flexibility and with privacy protections that are appropriate to the information that is being used or disclosed. In addition, data that has been de-identified consistent with the HIPAA standard can be used with effectiveness for critical research and public health purposes (along with a wide variety of other beneficial purposes for the health care system), while still protecting patient privacy. But rapidly changing technology and increasingly new and sophisticated uses of health information necessitate a review of this overall framework, and an effort to refine the various categories using a consistent terminology.

The HIPAA de-identification framework is not the only de-identification approach that is available, in the U.S. and around the world, but it provides a useful model for analysis and discussion because it is more detailed and comprehensive than many other potentially applicable approaches. De-identification - whether through the HIPAA model or otherwise - is a useful tool in protecting individual privacy interests, while at the same time enhancing innovation and the improved use of health and other personal information for important public and private purposes. In many situations, de-identified data can be leveraged to provide the same analytical value as using PHI, while fully embracing HIPAA's minimum necessary concept and appropriately protecting individual privacy interests.

Within the healthcare industry, however, there remains confusion and uncertainty as it relates to the precise elements of the HIPAA de-identification process and the use of de-identified data. Compounding these concerns, there is disagreement among experts as to the risks of re-identification of de-identified data. Moreover, there is a lack of qualified experts willing to opine on de-identification methodologies and results.

HITRUST believes clearer guidelines in the form of standards for the uses of de-identified data and managing associated risks are needed. In addition, HITRUST believes that consistent standards for the statistical and scientific methodologies used to arrive at this definition of "de-identified," enhanced protections in the form of appropriate technical, physical and administrative (including contractual) safeguards for de-identified data and standards used to certify experts that evaluate these methodologies and protections are also required.

While much confusion has been alleviated by the 2012 HHS guidance¹, the HITRUST Data De-identification Working Group (“De-IdWG”) has established some additional guidelines to clarify the differences in interpretations of what qualifies as de-identified data, incorporate practical guidance to establishing controls, and remediate general discomfort with reliance upon expert opinions.

Absent this framework of standards and controls, the ability of the healthcare industry to use data to improve the system will be slowed. HITRUST believes, though, that the appropriate use of de-identified data can protect patient privacy while at the same time make the healthcare system work better for everyone by paving the way for innovation and increased public health benefits. HITRUST believes that offering a framework of standards and controls around de-identification will significantly improve the current state and lead to better healthcare for all.

Defining Categories of Health Information

Data Definitions

Under HIPAA, healthcare information or data that is not individually identifiable or otherwise has been de-identified by the Safe Harbor or expert determination methods of de-identification is no longer subject to HIPAA regulations and may be used for any purpose while still protecting patient privacy². While the Safe Harbor method remains a legitimate means for data de-identification, the HITRUST De-IdWG primary focus is on expert determination data de-identification, which affords a controlled, secure de-identification process that preserves the integrity and quality of the data set for secondary uses. Moreover, expert determination data de-identification has proven to be an effective mechanism to ensure individual privacy and the preservation of the quality of the data to support uses beyond direct care services.

There remains, however, continued confusion over the use of expert determination de-identified data and the risks of re-identification, despite overwhelming evidence supporting the expert determination de-identification methodology. This white paper, therefore, serves as the HITRUST De-IdWG effort to establish documented data classification standards for data in various forms and to establish the foundation for a common framework of data designed to improve the overall understanding and use of the data sets that use de-identified data. Establishing such a common vocabulary will enhance the usefulness of the data, improve health care, increase industry and patient understanding, and create a consistent approach to the protection of personal data. The De-IdWG is hopeful that if these goals are met, industry, patient, member, and public concerns regarding de-identified data will be lessened.

HITRUST consulted a wide range of both domestic and international data protection laws, data definitions and other global frameworks to create a common vocabulary. The De-IdWG also took into account levels of risk: from patient data which contains direct identifiers or where indirect identification should be anticipated; data where the likelihood of re-identification is reduced through statistical techniques and other restrictions and safeguards, through to data where there is no possibility that the individual can be re-identified.

1. The guidance is available at <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html>

2. 45 C.F.R. Section 164.502(d)(2).

We identify below various uses that will fit into these different data classification standards. These descriptions are general, and the appropriateness of these Use Cases will depend on both specific data elements that are involved and the governing legal framework. Ultimately, the De-IdWG identified three core data classification standards:

- Patient Identifiable
- Patient De-Identified
- Patient Non-Identifiable

Patient Identifiable

Patient identifiable information is any data set that contains information relating to a natural person that can be used to directly identify that individual, or for which there is a reasonable basis to believe that indirect identification is possible. Under the US health care privacy framework, to be subject to the HIPAA rules, such information also must be created or received by a healthcare provider, health plan, or health care clearinghouse, and be related to an individual's past, present, or future physical or mental health condition or payment for the provision of healthcare to that individual.

Patient identifiable information is protected health information (PHI) under the HIPAA Privacy Rule and as such is highly sensitive. PHI should be secured in accordance with the HITRUST CSF as a mechanism to achieve compliance with the HIPAA Privacy and Security Rules.

Example Use Cases:

- For treatment and payment of medical services
- Case and disease management
- Health coaching
- Prescription adherence
- Institutional Review Board and Privacy Board approved studies under HIPAA
- Public health purposes (e.g., FDA Mini-Sentinel Program)

Patient De-Identified

Patient De-identified information is information or data where the risk is very small that the information could be used, alone or in combination with other reasonably available information, to identify an individual who is subject of the information. Patient De-identified information was patient identifiable, but data elements have been removed, generalized or transformed consistent with the applicable legal standard (HIPAA or otherwise) so that there is a very small risk that a patient could be re-identified. The HIPAA de-identification approaches include (1) the “Safe Harbor” approach, where a specific list of identifiers is removed from the data, and (2) information that has been de-identified by the expert determination method, where identifiers are removed and there is an overall calculation that there is a low risk of re-identification, based on administrative, physical and technical safeguards, anticipated recipients and other factors.

Examples of patient de-identified information include:

De-Identified Data: Health information that does not identify an individual because the probability that the information can be used to identify an individual is statistically very small.

Example Use Cases

- Early disease outbreaks detection and geographic tendencies
- Quality and outcomes research analysis
- Reduction of medical errors and improved patient safety
- Provider quality transparency to patients
- Commercial use, including marketing and product development

Longitudinal De-Identified Data: De-Identified data that does not identify an individual, corresponds to multiple transactions related to a person over time, and for which the probability of re-identification is very small taking into account the longitudinal nature of the data.

Example Use Cases

- Comparative Effectiveness Research (CER)
- Health Economics and Outcomes Research (HEOR)
- Other de-identified research
 - Studies analyzing the relationship between medication compliance and outcomes
 - Studies that compare the results of one approach for managing disease to the results of other approaches
- Factors that influence the adherence to certain medication

Patient Non-Identifiable

Patient non-identifiable information is personal information that is processed in such a way that the information obtained cannot be associated with an identified or identifiable person.

Example Use Cases:

- Healthcare cost management/Provider cost comparison data
- Aggregated analytics regarding member/patient use of health care services
- FDA uses of large population database records to identify risk factors
- CDC uses data to perform seasonal tracking of infectious diseases such as influenza
- FDA Risk Minimization Action Plans (RiskMAPS) – drug safety

This common vocabulary for levels of health care data will support the ongoing efforts to provide guidance and clarity to the framework provided by HIPAA for the use of de-identified data.

Evaluating De-Identification Methodologies

Background

The foundation for de-identification methodologies that apply to United States healthcare organizations is the HIPAA Privacy Rule and the September 4, 2012, U.S. Department of Health and Human Services (“HHS”) Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) (Guidance). Yet neither the Rule nor the Guidance provides criteria by which de-identification methodologies may be assessed.

The Guidance does provide, however, some limited discussion regarding the Expert Determination Method. In particular, it provides a “general understanding” that a combination of technical and policy procedures are often applied to the de-identification task. HHS 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 HHS upon request.

The Guidelines continue and provide a three step process by which an expert can determine that data has been de-identified:

- First, the expert evaluates the extent to which the health information can (or cannot) be identified by the anticipated recipients.
- Second, the expert provides guidance on which statistical, scientific or other methods can be applied to the health information to mitigate the anticipated risk.
- 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. This will most likely require several iterations.

While this process is helpful in explaining how to arrive at a de-identified data set, it does not address the holistic requirements necessary for an effective de-identification program and methodology.

After reviewing multiple de-identification programs and methods, including those distributed by agencies in the United States, Canada, and the United Kingdom³, HITRUST believes that no one method is appropriate for all organizations. Instead, HITRUST has identified twelve criteria for a successful de-identification program and methodology that can be scaled for use with any organization. These twelve characteristics are further divided into two general areas: Program and De-identification Methodology. The first set of characteristics represents those for the program and the administrative controls that an organization should have in place to govern de-identification.

3. In addition to the HIPAA Guidance, the De-IdWG reviewed the following standards: Federal Trade Commission Report, 20-22 Protecting Consumer Privacy in an Era of Rapid Change, Recommendations for Businesses and Policy Makers (Mar. 2012); Statistical Policy Working Paper 22, Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology (1999; Second version, 2005); Checklist on Disclosure Potential of Proposed Data Releases, The Interagency Confidentiality and Data Access Group, an Interest Group of the Federal Committee on Statistical Methodology; Best Practice’ Guidelines for Managing the Disclosure of De-Identified Health Information, Health System Use Technical Advisory Committee Data De-Identification Working Group, Canadian Institute for Health Information (Oct. 2010); Anonymization: Managing Data Protection Risk Code of Practice, Information Commissioner’s Office, United Kingdom (Nov. 2012).

The second set represents how the organization can actually arrive at a de-identified data set, either on an ad hoc basis or by instituting a process that will deliver de-identified data sets.

The two areas and twelve criteria are as follows. This paper will address each of the areas and criteria in turn.

Program Methodology

1. Governance
2. Documentation
3. Explicit Identification of the Data Custodian and Recipients
4. External or Independent Scrutiny

De-identification Methodology

1. Re-Identification Risk Thresholds
2. Measurement Of Actual Re-Identification Risks
3. Identification And Management Of Direct Identifiers And Quasi-Identifiers;
4. Identification Of Plausible Adversaries And Attacks;
5. Identification Of Specific Data Transformation Methods And How They Reduce The Risks
6. Process And Template For The Implementation Of Re-Identification Risk Assessment And De-Identification
7. Mitigating Controls To Manage Residual Risk
8. Data Utility

Program Methodology

1. Governance. The organization's de-identification program should address how it will be governed. Organizations should consider addressing at a minimum the following areas:

- Assigning overall responsibility for the de-identification program
- Identifying the area of the organization responsible for data de-identification and data custodians
- De-identification policies and procedures
- Training of individuals responsible for de-identification and use and/or disclosure of de-identified data
- Independent or external review
- Identification of recipients and agreement that re-identification will not be attempted
- Maintaining transparency around de-identification practices

- Standards regarding how often the re-identification risk assessment need to be re-performed
- Regular examination of disclosure of overlapping data sets
- Regulatory change monitoring
- Re-identification/breach response process
- Technical and physical controls to further mitigate re-identification risk

2. Documentation. For an effective methodology, the organization must document the governance program in criterion 1 so that it is repeatable and can be reviewed by an external expert. At a minimum, the documentation should include a description of the organization's program, de-identification policies, procedures where necessary, and training where appropriate. Individuals should be trained on the importance of not re-identifying (or attempting to re-identify) de-identified data.

3. Explicit Identification of the Data Custodian and Recipients. For a de-identification methodology to be effective, the organization must know who the custodian of the data is and who is receiving the data. This is critical, because both internal and external controls are needed to ensure that an organization administratively controls data. In particular:

- There should be a data custodian responsible for internal use and disclosure of the data. A mature organization should have a separate individual or team that is responsible for the de-identified data set and ensures compliance with its de-identification methodology.
- Each recipient of the data must be identified, including all downstream recipients and intended uses. Recipients must enter into appropriate agreements that prohibit re-identification and allow for changes in the event of modifications to the de-identification determination.

4. External or Independent Scrutiny. For a methodology to be effective, it and the final result must be reviewed and approved by an external or independent expert(s), or an appropriate regulatory or standards-setting body. HIPAA, for instance, requires that "a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" determines that the re-identification risk is very small. For an organization to meet this standard, the expert must determine, at a minimum, that the methodology used is appropriate to result in a de-identified data set." At a minimum, this determination should be reviewed informally each year, with a full review every three years.

De-Identification Methodology

1. Re-identification Risk Thresholds. Re-identification risk thresholds should be established taking into account the context of the data, including its recipients, potential re-identification attacks, and controls. If the measured risk in the data is higher than the threshold then the data is considered to be personal information. If the measured risk is below the threshold then the data is considered not to be personal information. Therefore, the definition of thresholds is critical. Once the threshold(s) are known, then it would be possible to make a case that a data set has an acceptably small re-identification risk.

2. Measurement of Re-identification Risks. The organization's methodology must have a way in which it can objectively measure the risk of re-identification in a repeatable objective fashion. In order to measure the risk of re-identification, metrics must be defined that are consistent with the plausible attacks that have been identified and the nature of the data. The metrics should take into account the maximum risk, average risk, and the risk with respect to longitudinal data.

3. Identification and Management of Direct Identifiers and Quasi-identifiers. A key part to any de-identification methodology is the identification of direct identifiers and data elements that increase the likelihood of re-identification, known as "quasi-identifiers." Under HIPAA, this means that at a minimum the following 18 direct and quasi-identifiers should be considered:

- Names;
- All geographic subdivisions smaller than a State;
- All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89;
- Telephone numbers;
- Fax numbers;
- Electronic mail addresses;
- Social security numbers;
- Medical record numbers;
- Health plan beneficiary numbers;
- Account numbers;

- Certificate/license numbers;
- Vehicle identifiers and serial numbers, including license plate numbers;
- Device identifiers and serial numbers;
- Web Universal Resource Locators (URLs);
- Internet Protocol (IP) address numbers;
- Biometric identifiers, including finger and voice prints;
- Full face photographic images and any comparable images; and
- Any other unique identifying number, characteristic, or code (unless otherwise permitted).

The methodology should pay particular attention to dates of service and geo-locators, such as zip codes, hospital referral regions, and metropolitan areas. Other possible quasi-identifiers could include race, income, and various consumer data. Documentation of identifiers and quasi-identifiers must be performed to allow for audit or investigation.

4. Identification of Plausible Adversaries and Attacks. The de-identification methodology must also take into account potential adversaries that could attempt to re-identify the data, and both guidance and steps should be provided to ensure a comprehensive and repeatable outcome. The organization should first ask whether the data is public or nonpublic. Based on this context, it should then take into account scenarios of deliberate attacks to re-identify the data (demonstration attacks, rogue employees, “nosy neighbors,” academic research, business advantage, and the data recipient itself), as well as accidental or inadvertent re-identification of the data (such as a researcher recognizing a co-worker in the office), and the possibility of a data breach that would make nonpublic data public.

5. Identification of Specific Data Transformation Methods and How They Reduce the Risks. The organization will often need to transform the data to make it de-identified. At a minimum, the methodology used here must be irreversible or there should be a way to assess whether it can be easily reversed. (This latter consideration is important because there are many techniques that have been proposed to protect data, and some of them have been found to be vulnerable or only applicable under certain conditions.) Where appropriate, the organization should consider reviewing data transformations on a regular basis and validating them where there is a material change. Finally, the usability of the data must be taken into account post-transformation. For instance, the banding of years of age in a study regarding the effectiveness of a certain drug for newborns may not be helpful.

Some common examples of data transformation include:

- Suppression
- Pseudonymization
- Generalization/Reduction in detail
- Perturbation
- k-anonymity principle
- Substitution, e.g., replacing names
- Elimination of small cell counts
- Summarization
- Derived data items and banding

6. Process and Template for the Implementation of Re-identification Risk Assessment and De-identification.

The methodology must have a repeatable process to quantify the controls based against the risk. This can be on a case-by-case basis, or for certain organizations, the data custodian may be performing de-identification of data feeds on a continuous basis. The implementation of de-identification will be different in each of these two scenarios, but for either one the process, whether by hand or through an automated workflow, must be documented.

7. Mitigating Controls to Manage Residual Risk. The methodology must have in place appropriate administrative, technical, and physical controls to manage residual risk. Examples include implementing the HITRUST Common Security Framework (CSF), a privacy and security program that oversees the organization's compliance with the HIPAA Privacy and Security Rules, and a policy requiring that data users agree that they will not attempt to re-identify the data. Any controls implemented should be auditable.

8. Data Utility. The de-identification methodology must not only consider the risk of re-identification and what quasi-identifiers are in the data, but must also consider what is in the data and its ultimate usefulness. To meet this standard, the organization must take into account:

- The suitability to the recipient of the data;
- The data transformation methods; and
- The form of the data, including whether it is able to be used with standard applications and tools.

Section Summary

HITRUST believes that if an organization's program and de-identification methodology meets the above twelve criteria then it will be consistent with a common framework. If it does not meet these twelve criteria, then the organization is at risk that a review would result in a negative finding.

Scoring

During an assessment of the program and methodology, the twelve criteria can be scored along two dimensions: implementation and capability. The implementation dimension indicates whether a criterion can be satisfied in any way – it is essentially an indicator of existence of a practice or set of practices. The capability dimension indicates how well a criterion can be satisfied. A criterion can be satisfied in a very basic manner (low capability), or in a very convincing manner (high capability). A criterion must be implemented to have any capability. A sample scoring approach is attached as Appendix B.

Methods of De-Identification

Two Methods of De-Identification

The HIPAA Privacy Rule provides two methods by which protected health information can be “de-identified.” The first method – often called the “safe harbor” method, provides that PHI is de-identified when specific enumerated identifiers about the “individual or of relatives, employers, or household members of the individual,” are removed from the PHI (such as name, telephone numbers; electronic mail addresses; Social Security numbers; and medical record numbers (full list set forth at 45 C.F.R. § 164.514(b)(2)). This requirement also includes: “any other unique identifying number, characteristic, or code” except as otherwise permitted by the rule. In addition, for the PHI to be de-identified, the covered entity must “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.” As set forth in the regulation, this method appears to be an “automatic” method of de-identification, although clearly there is some need for at least a “real world” reality check on the “actual knowledge” component of this analysis, but presumably this analysis can be conducted and overseen by a knowledgeable layperson.

The second method is more complicated and involves additional judgment and analysis. This method (the “expert determination” method) permits PHI to be considered de-identified where:

- A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
- 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
- Documents the methods and results of the analysis that justify such determination.

This expert determination method therefore requires (1) “a person with appropriate knowledge of and experience with” (2) “generally accepted statistical and scientific principles and methods for rendering information not individually identifiable;” and (3) application of “such principles and methods”; (4) to reach a conclusion that “the risk is very small” that (5) 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 (6) documentation of the “methods and results” of the analysis to justify the determination.

Who has appropriate knowledge and experience?

So, who are these people “with appropriate knowledge of and experience with” these principles?

The Department of Health and Human Services, in its most recent guidance on these de-identification principles, has made clear that an “expert” in this area does not mean a specific, unique set of skills. In fact, according to HHS,

“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.”

Therefore, it is important that any organization that wishes to de-identify information using this method employ an individual with the appropriate set of academic or substantive skills perspective. This individual also should have experience in the area and a relevant knowledge of how HIPAA operates and how these kinds of de-identification activities will proceed, along with relevant expertise on information safeguards and similar protections specific to a particular arrangement.

HHS has made it clear that the concept of “expert guidance” is not unique to HIPAA. For example, according to HHS,

“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.”

Similarly, it is clear that an expert must recognize – and be able to work with – the fact that there is not a single “level of risk” in this area or any “one size fits all” approach to de-identification. For any de-identification review, context matters. While the concept of “very small” from the HIPAA rules is based on guidance from a variety of sources,

“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 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.”

In addition, experts must be prepared to review and consider a variety of approaches for addressing de-identification issues. No single universal solution addresses all privacy and identifiable data issues. Instead, based on expertise and experience, a combination of technical and policy procedures are often applied to the de-identification task. Therefore, under the HHS Guidance, while HIPAA “does not require a particular process for an expert to use to reach a determination that the risk of identification is very small,” it is crucial to remember that the key to demonstrating HIPAA compliance is that the HIPAA Privacy Rule “does require that the methods and results of the analysis that justify the determination be documented and made available to OCR upon request.”

In utilizing this expertise and experience, the expert often must follow a series of related steps in making an appropriate determination. These should include a methodology that meets the criteria identified above, including:

- Re-identification Risk Thresholds.
- Measurement of Actual Re-identification Risks.
- Identification and Management of Direct Identifiers and Quasi-identifiers.
- Identification of Plausible Adversaries and Attacks.
- Identification of Specific Data Transformation Methods and How They Reduce the Risks. Process and Template for the Implementation of Re-identification Risk Assessment and De-identification.
- Mitigating Controls to Manage Residual Risk.
- Data Utility.

Accordingly, because of this need for flexibility and creativity in addressing the various approaches to de-identification and the related concept of a “very small” likelihood of re-identification, it is clear that an appropriate expert must have a significant array of skills. It is clear that a strong understanding of the HIPAA Privacy and Security rules is relevant and useful. In many situations, it also may be critical to understand specific or business principles at issue, particularly where significant operational controls are in place.

There is therefore wide latitude in who is an expert. But for purposes of HITRUST, and based on HHS regulations and guidance, an expert “checklist” can be summarized as follows. For an individual to qualify as an expert, he or she must have:

- Appropriate and relevant professional and academic or other training regarding generally accepted statistical and scientific principles and methodologies for rendering information not individual identifiable. This includes:
 - Professionals and academics, including scientists, mathematicians, health economists, computer scientists, epidemiologists, medical and biomedical informatics, and statisticians in various fields that routinely determine and accordingly mitigate risk with respect to sharing data.
 - This appropriate academic background should be combined with academic or other training in one or more of the following:
 - » Statistical Disclosure Limitation/Control
 - » U.S. Census Data and Demography

- » Statistical Sampling
 - » Privacy Preserving Data Publishing
 - » Medical Informatics/Standard Healthcare Billing Coding
 - » Privacy Preserving Data Mining
 - » Biostatistics/Epidemiology
 - » Health Systems Research
 - » Cryptography
 - » Computer Security
 - » Geographic Information Systems
 - » Quantitative Risk Management Analysis and Methods
- Actual experience using health information de-identification methodologies. This can be satisfied through knowledge of the methodologies discussed above and:
 - Research and academic work involving actual data sets;
 - Professional experience with an established expert;
 - Successfully completing a reputable de-identification methodology course that provides actual experience.
 - The ability to reach a conclusion, using the de-identification methodologies, 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.”
 - A strong understanding of the HIPAA Privacy and Security rules.
 - The willingness and ability to document the “methods and results” of the analysis to justify the determination.

Thus, while not all of these skills may be required in each situation, this checklist helps defines what an expert is. Ultimately, it will be critical for companies to work with experts who have a strong background in most (if not all) of these areas. Companies should be able to document not only the background and experience of the expert, and also the considerations that went into a specific determination of de-identification.

Re-Identification Risk

The issues surrounding the de-identification of personal data are complicated and evolving. While HIPAA provides a precise framework for de-identification, many other legal and regulatory frameworks do not.

It is also clear that in many contexts and across the globe that there are substantial public and private benefits to the broader use of personal data in a wide variety of situations. We are just beginning to understand the opportunities provided by the vast array of data that is gathered. In many of these potentially beneficial situations, data that has been anonymized to such an extent that re-identification is impossible also is data that has been stripped of much of its value.

In the HIPAA setting and other contexts, the concept of a “small risk” of re-identification has become an appropriate component of the legal framework. This “small risk” element permits the vast benefits of this data in a context where associated privacy risk is small. We encourage inclusion of this concept in any new legal frameworks governing privacy and the de-identification of personal health data.

In some situations, this concept of “small risk” can be addressed solely through statistical and technical means. There are data sets where the risk of re-identification can be measured quantitatively; this section focuses on some of these situations and approaches. However, in many situations, the true measure of “small risk” incorporates not only these statistical and quantitative measures, but also additional privacy-protective variables, including legal restrictions, contractual requirements and the development of appropriate technical, physical and administrative safeguards that are adopted to protect the security and confidentiality of personal data when significant identifiers have been removed. The determination of a “small risk,” therefore, must be made in the overall context of how data is used and disclosed, and not just quantitatively from the data.

We now consider the statistical and technical elements of this “small risk” discussion. It highlights - for data suppliers, data users and others - some of the key variables that appropriate personnel should review and evaluate in determining whether there is a small risk of re-identification. These principles will always be a component of any de-identification discussion (for example, a group of patient names will never be considered de-identified simply because of significant safeguards), but these principles reflect only one important component of how a “small risk” of re-identification should be assessed in this evolving data landscape.

To be able to define in a meaningful way what “very small” risk is, we first need to present a few preliminary principles and concepts, such as the different ways to measure re-identification risk.

Categories of Variables

It is useful to differentiate among the different types of variables in a data set. The way the variables are handled during the de-identification process will depend on how they are categorized. We make a distinction among three types of variables⁴:

4. See L. Sweeney, “k-anonymity: a model for protecting privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002; P. Samarati, “Protecting respondents’ identities in microdata release,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

Directly Identifying Variables

Direct identifiers have two important characteristics: (a) one or more direct identifiers can be used to uniquely identify an individual, either by themselves or in combination with other readily available information (such as telephone number linked with name), and (b) they are very often not useful for data analysis purposes. Examples of directly identifying variables include names, email address, and telephone numbers of the data subjects, all three of which are typically unhelpful for research and data analysis.

Indirectly Identifying Variables (quasi-identifiers)

Quasi-identifiers are variables about data subjects that an adversary could use, either individually or in combination, to attempt to re-identify a record, but only if the adversary has appropriate background knowledge of the variable⁵. The manner by which an adversary can obtain such background knowledge will determine which attacks on a data set are plausible. For example, the background knowledge may be available because the adversary knows a particular target individual in the disclosed data set, a data subject has a visible characteristic that is also described in the data set, or the background knowledge exists in a public or semi-public registry. Examples of quasi-identifiers include sex, date of birth or age, locations (such as postal codes, census geography, information about proximity to known or unique landmarks), languages spoken, ethnic origin, total years of schooling, marital status, criminal history, total income, minority status, activity difficulties/reductions, profession, event dates (such as admission, discharge, procedure, death, specimen collection, visit/encounter), codes (such as diagnosis codes, procedure codes, and adverse event codes), country of birth, birth weight, and birth plurality.

Other Variables

These are the variables that are not direct or quasi-identifiers and are not useful for determining an individual's identity. They may be clinically relevant or not. We would generally not worry about these variables.

Individuals can be re-identified because of the directly identifying variables and quasi-identifiers. Therefore, our focus will be on these two types of variables only.

Measuring Re-Identification Probability

Re-identification risk measurement is a topic that has received extensive study over multiple decades. We will examine risk measurement at a conceptual level to illustrate key concepts. This builds on the classification of variables described above.

The Risk of Re-Identification for Direct Identifiers

One must define risk as the probability of re-identifying the record of a data subject. If there is a direct identifier in a data set, then, by definition, it would be considered to have a very high risk of re-identification.

Strictly speaking the probability is not always very high. However in the case where direct identifiers exist, erring on the conservative side and assuming the probability of re-identification is very high allows us to focus our attention on the quasi-identifiers, which is where, in many instances, the most data utility is.

5. In the context of de-identification and potential re-identification, the term "adversary" is often used to denote any individual that might attempt to re-identify a de-identified data set without explicit permission. It is not necessarily pejorative, however, and can apply to researchers, journalists, or others. As we look at determining the risk of re-identification throughout this paper, we assume that such an adversary exists.

For direct identifiers, several methods could be applied to protect them: field suppression, randomization, and pseudonymization. These methods are often referred to collectively as “data masking.” Suppression is simply the removal of the variable. The second method entails replacing data values with randomly generated values. The third method is to create a pseudonym⁶. Pseudonymization is also sometimes called “coding” in the health research literature⁷.

Protecting direct identifiers using these methods does not affect the data utility as direct identifiers are not typically useful for research. Quasi-identifiers, however, cannot be protected using such procedures. This is because the resulting data, in almost all cases, will not be useful for analytic purposes. As such, a different set of approaches are required for measuring the risk of re-identification and de-identifying the quasi-identifiers.

The Risk of Re-identification for Quasi-identifiers

Equivalence Classes

All the records that share the same values on a set of quasi-identifiers are called an equivalence class. Equivalence class sizes for a data concept, such as age, potentially change during de-identification. In general there is a trade-off between the level of detail provided for a data concept and the size of the corresponding equivalence classes, with more detail associated with smaller equivalence classes.

A very common way to measure the probability of re-identification for a record in the data is equal to one divided by the size of its equivalence class. For example, if an equivalence class is at a size of 3, then the probability for each of these records is 1/3. Records that are in equivalence classes of size 1 are called “uniques”.

This probability applies under two conditions: (a) the adversary knows someone in the real world and is trying to find the record that matches that individual, and (b) the adversary has selected a record in the data set and is trying to find the identity of the person in the real world. Both of these types of attacks on health data have occurred in practice, and therefore both perspectives are important to consider. An example of the former is when an adversary gathers information from a newspaper and attempts to find the data subject in the data set. An example of the latter re-identification attempt is when the adversary selects an unusual record in the data set and tries to match it with a record in the voter registration list.

A key point here is that the probability of re-identification is not based solely on the uniques in the data set. A record may not be unique, but could still have a high probability of re-identification. Therefore, it is recommended that we consider, and manage, the risk of re-identification for both uniques and non-uniques.

The overall risk for a data set may be calculated as the maximum probability across all records (the “maximum risk”), or the average across all records (the “average risk”). Typically the former is used when data is going to be released publicly, and the latter when the data is going to be shared with a known data recipient and additional controls are going to be imposed.

6. K. El Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O’Reilly, 2013.

7. HIMSS Analytics, “2010 HIMSS Analytics report: Security of patient data,” HIMSS, 2010.

Taking Context into Account

The context of a data disclosure is critical for deciding what an acceptable risk is. For example, the acceptable risk for disclosing a data set publicly would be very different from the acceptable risk for a data set disclosed to data recipients that have implemented significant administrative, security and privacy controls. For data disclosed publicly we assume that the adversary will launch a demonstration attack. There are no other controls that can be put in place. For a non-public data set we consider three types of attacks that cover the universe of attacks: deliberate, inadvertent, and breach⁸.

A deliberate attack transpires when the adversary deliberately attempts to re-identify individuals in the data set. This may be a deliberate decision by the leadership of the data recipient (e.g., the data recipient decides to re-identify individuals to link to another data set) or by a rogue employee associated with the data recipient. This risk depends on the security and privacy controls that the data recipient has in place and the contractual controls that are being imposed as part of the data sharing agreement.

The inadvertent attack transpires when a data analyst working with the data recipient (or the data recipient herself) inadvertently re-identifies someone in the data set. For instance, this could occur when the recipient is already aware of the identity of someone in the data set, such as a friend, relative, or more generally, an acquaintance.

The breach attack transpires when there is a data breach at the data recipient. Examples include hacking or improper disclosure that removes privacy and security controls from the data set.

What is Acceptable Risk?

There are strong precedents for what can be considered an acceptable amount of risk. These precedents have been in use for many decades, are consistent internationally, and have persisted over time as well⁹. It should be noted, however, that the precedents set to date have been for assessments of “maximum risk”. In the HIPAA context, commentary about the de-identification standard in the HIPAA Privacy Rule, HHS notes in the Federal Register that

“the two main sources of disclosure risk for de-identified records about individuals are the existence of records with very unique characteristics (e.g., unusual occupation or very high salary or age) and the existence of external sources of records with matching data elements which can be used to link with the de-identified information and identify individuals (e.g., voter registration records or driver’s license records). [...] an expert disclosure analysis would also consider the probability that an individual who is the target of an attempt at re-identification is represented on both files, the probability that the matching variables are recorded identically on the two types of records, the probability that the target individual is unique in the population for the matching variables, and the degree of confidence that a match would correctly identify a unique person.”

8. K. El Emam, Guide to the De-Identification of Personal Health Information. CRC Press (Auerbach), 2013; K. El Emam and L. Arbuckle, Anonymizing Health Data: Case Studies and Methods to Get You Started. O’Reilly, 2013.

9. K. El Emam, Guide to the De-Identification of Personal Health Information. CRC Press (Auerbach), 2013;

It is clear that HHS considered unique records to have a higher risk of re-identification in certain circumstances, and that non-unique records are more likely to have an acceptably low risk of re-identification. Further, a unique within a data set (sometimes referred to as a sample unique) where there is no corresponding unique in reasonably available external data (sometimes referred to as a population unique) may be acceptable. Yet, uniqueness is not a universal threshold. Historically, data custodians (particularly governmental agencies focused on reporting statistics) have used the “minimum cell size” rule as a threshold for deciding whether to de-identify data. This rule has been applied in a variety of contexts, with different conclusions, often driven by the context and sensitivity of the relevant data¹⁰. It should be recognized though that there is no agreed upon threshold, however, even for what many people would agree is highly sensitive data¹¹. Public data releases have used different cell sizes in different jurisdictions. The variability is due, in part, to different tolerances for risk, the sensitivity of data, whether a data sharing agreement will be in place, and the nature of the data recipient.

In general then, there are a range of values for what is acceptable risk with strong precedents. The question is which one of these values should be used. The answer will depend on the outcome of a risk management exercise, as outlined below.

How to Manage Re-identification Risk

Managing re-identification risk means doing three things: (a) selecting an appropriate risk metric, (b) selecting an appropriate risk threshold, and (c) measuring the risk in the actual data set that will be disclosed. The choice of a metric often is a function of whether the data set will be released publicly or not¹². In the current section we will examine the next step (b): selecting thresholds.

Selecting an acceptable threshold within the range described above requires an examination of the context of the data release. The re-identification risk threshold is determined based on factors characterizing the data recipient and the data itself¹³. These factors have been suggested and have been in use informally by data custodians for at least the last decade and a half¹⁴. They cover three dimensions¹⁵ as illustrated in Figure 1:

-
10. For example, this concept was originally applied to count data in tables (e.g., number of males, aged 30-35, living in a certain geographic region). The most common minimum cell size in practice is 5, which implies that the maximum probability of re-identifying a record is 1/5, or 0.2. Some custodians, such as certain public health offices, use a smaller minimum count, such as 3. See, e.g., A. de Waal and L. Willenborg, “A view on statistical disclosure control for microdata,” *Survey Methodology*, vol. 22, no. 1, pp. 95–103, 1996. Others, by contrast, use a larger minimum, such as 11 (in the United States). See, e.g., Centers for Medicare and Medicaid Services, “BSA Inpatient Claims PUF. 2011;” “2008 Basic Stand Alone Medicare Claims Public Use Files.” [Online]. Available: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/Downloads/2008_BSA_PUF_Disclaimer.pdf.
 11. For example, a minimal count of 3 and 5 were recommended for HIV/AIDS data (Centers for Disease Control and Prevention, “Integrated Guidelines for Developing Epidemiologic Profiles: HIV Prevention and Ryan White CARE Act Community Planning.”) and abortion data (Statistics Canada, “Therapeutic abortion survey,” 2007) respectively.
 12. K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013; K. El Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O’Reilly, 2013.
 13. K. El Emam, “Risk-based de-identification of health data,” *IEEE Security and Privacy*, vol. 8, no. 3, pp. 64–67, 2010.
 14. T. Jabine, “Procedures for restricted data access,” *Journal of Official Statistics*, vol. 9, no. 2, pp. 537–589, 1993; T. Jabine, “Statistical Disclosure Limitation Practices of United States Statistical Agencies,” *Journal of Official Statistics*, vol. 9, no. 2, pp. 427–454, 1993.
 15. K. El Emam, A. Brown, P. Abdel Malik, A. Neisa, M. Walker, J. Bottomley, and T. Roffey, “A method for managing re-identification risk from small geographic areas in Canada,” *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, p. 18, Apr. 2010.

- **Mitigating controls.** This is the set of security and privacy practices that the data recipient has in place. A recent review has identified a union of practices used by large data custodians, and recommended by funding agencies and IRBs for managing sensitive health information¹⁶. A list of such practices and a mapping to the HISTRUST CSF is provided in Appendix A.
- **Invasion of privacy.** This evaluates the extent to which a particular disclosure would be an invasion of privacy to the patients or individuals. There are three typical considerations: (a) the sensitivity of the data: the greater the sensitivity of the data the greater the potential invasion-of-privacy, (b) the potential injury to patients from an inappropriate disclosure: the greater the potential for injury the greater the potential invasion of privacy, and (c) the appropriateness of the basis for use and disclosure of the data: the less appropriate the legal, ethical, or other basis for use and disclosure of the data, the greater the potential invasion-of-privacy. In some cases, for instance, the use and disclosure of data may be based upon the legal framework (e.g., HIPAA or available privacy policies) and in other cases it may be based upon individual consent.
- **Motives and capacity.** This considers the motives and the capacity of the data recipient to re-identify the data, considering issues such as conflicts of interest, the potential for financial gain from a re-identification, and whether the data recipient has the skills and the financial capacity to re-identify the data. In general, many of these elements can be managed through contracts (e.g., a prohibition on re-identification, restrictions on linking the data with other data sets, and disallowing the sharing of the data with other third parties).

For example, if the mitigating controls are low, which means that the data recipient has poor security and privacy practices, then the re-identification threshold should be set at a lower probability level. This will result in more de-identification being applied. However, if the data recipient has very good security and privacy practices in place, then the threshold can be set at a higher probability. If the sponsor is disclosing the data through an online portal, then the sponsor has control of many, but not all, of the mitigating controls. This provides additional assurances to the sponsor that a certain subset of controls would be implemented to their satisfaction.

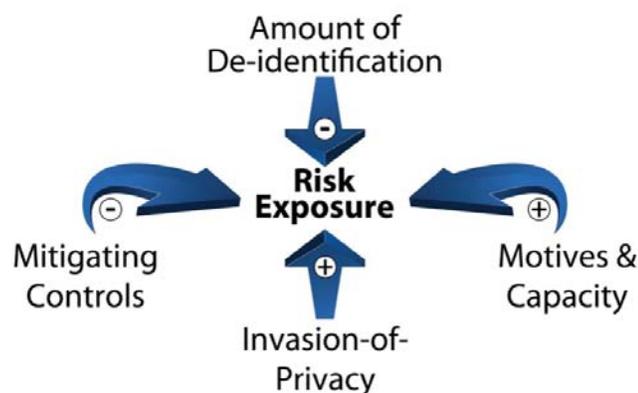


Figure 1: Factors to consider when deciding on an acceptable re-identification risk level.

16. K. El Emam, Guide to the De-identification of Personal Health Information. CRC Press (Auerbach), 2013; K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk, "Evaluating Patient Re-identification Risk from Hospital Prescription Records," Canadian Journal of Hospital Pharmacy, vol. 62, no. 4, pp. 307-319, 2009.

Once a threshold is determined, the actual probability of re-identification is measured in the data set. If the probability is higher than the threshold, then transformations need to be performed on the data. Transformations can include increasing cell sizes, changing geography, eliminating certain data, perturbation, obfuscation, or other methods. Otherwise, the data cannot be declared to have a very small risk of re-identification.

The point here is that the amount of data transformation that needs to be applied will be a function of these context factors.

For example, if the data recipient has good security and privacy practices in place, then the threshold chosen will be higher, which means that the data will receive less transformation. The security and privacy practices of the data recipient can be manipulated through contracts to assure that data privacy is maintained. The contract that the data recipient signs can impose a certain list of practices that must be in place. These practices were the basis for deciding on the threshold. Therefore, they need to be in place by the data recipient to justify the level of transformations that have been performed on the data. These practices are described in Appendix A, but also include prohibiting re-identification, requiring audits, not allowing further disclosure, providing for injunctive relief, and the like.

For public data releases there are no contracts and no expectation that any mitigating controls will be in place. In that case, the lowest probability thresholds are used.

Conclusion

The HITRUST De-Identification Framework aims to ensure that PHI data is handled appropriately, and meets the requirements set forth in federal and state laws and regulations for the use, storage and maintenance of the associated of healthcare data.

As the use of PHI data continues to increase in the healthcare space along with the need to share that data; the ability to assure data privacy throughout its lifecycle will only grow.

The HITRUST De-Identification Framework™ provides a consistent, managed methodology for the protection of healthcare information through de-identification, and a level of consistency and repeatability needed for organizations to fully understand and apply it appropriately. Additionally it provides an appendix on risk treatments, which includes some of HITRUST's views on transference, avoidance, mitigation and acceptance with explanatory examples, including corrective action plan (CAP) prioritization and alternate control risk analysis that will be incorporated into the HITRUST CSF.

De-identification is a useful tool in protecting individual patient privacy and organizational compliance while creating more efficient use of healthcare data and enhancing innovation for the organization and its healthcare partners. Through secure and appropriately handled de-identified information, future advances in technology and treatment will eventually be expedited to the organization and the patient.

Appendix A: De-Identification—CSF Mapping

One of the requirements for HITRUST's De-Identification Methodology includes the implementation of mitigating controls designed to manage the risk from the use of Patient De-Identified data. This involves objectively evaluating the extent to which the data recipient has implemented security and privacy controls – the so-called Mitigating Controls. If a data recipient has extensive mitigating controls then fewer modifications may be required to the data set itself. Conversely if the data recipient has weak mitigating controls in place, then more extensive modification likely are needed to the data to ensure that the risk of re-identification is very small.

For that purpose the De-identification Controls Assessment (DCA) checklist was developed, and this is presented in the current appendix. The DCA can be used to score the practices of the data recipient. Because the framework pertains to data that has been de-identified, the DCA covers a subset of the CSF. The mapping between the DCA and the CSF is provided below.

The DCA is based on a checklist originally developed and reported in a Guide to the De-identification of Personal Health Information¹⁷. The original checklist was developed empirically by examining the practices of large data custodians and professional association guidance for sharing health information. It has also been practically used for the last 5 years and mapped to contemporary standards. Therefore, it has strong face validity. This checklist was then mapped to the HITRUST CSF control requirements as they are specified for assessment in HITRUST's MyCSF governance, risk and compliance tool.

Accompanying the DCA is a scoring scheme. Assessment and scoring of the DCA is performed in accordance with standard HITRUST CSF Assurance methodologies. For more information on these methodologies, refer to the Risk Analysis Guide for HITRUST Organizations and Assessors¹⁸.

This checklist should be used to establish a control risk score which then should be used within the HITRUST de-identification framework as one of the key inputs to decide how much to perturb the actual data set.

17. K. El Emam, Guide to the De-identification of Personal Health Information. CRC Press (Auerbach), 2013.

18. B. Cline, "Risk Analysis Guide for HITRUST Organizations and Assessors," Aug-2014. [Online]. Available: <https://hitrustalliance.net/content/uploads/2014/10/RiskAnalysisGuide.pdf>. [Accessed: 18-Mar-2015].

The following table provides the DCA category and CSF control mapping, along with the associated CSF assessment requirement statement from the MyCSF. Note bolded text indicates changes to an existing requirement or a new requirement, including a new de-identification industry segment for those requirements considered specific to the DCA.

DCA Category Name	CSF Control Name	CSF Level	CSF Assessment Requirement Statement
Access Control: General	01.a Access Control Policy	1	Logical and physical access control rules and rights for each user or group of users for each application are considered together and clearly defined in standard user access profiles (e.g., roles) based on need-to-know, need-to-share, least privilege and other relevant requirements.
Access Control: Access Policies	01.c Privilege Management	2	Role-based access control is implemented and capable of mapping each user to one or more roles, and each role to one or more system functions.
Data Sharing Agreements (DSAs): DSAs	09.t Exchange Agreements	1	Exchange or data sharing agreements specify the minimum set of controls on responsibility, procedures, technical standards and solutions, as well as organizational policies.
Non-disclosure and Confidentiality (NDAs/CAs): Policy	02.c Terms and Conditions of Employment	2	The organization maintains a list of all authorized signed non-disclosure agreement (NDA) forms; and the list is kept up to date to reflect personnel or other workforce member changes and departures.
Aggregated Data: Disclosure Policy	13.I Minimum Necessary Use	1	The covered entity or business associate (i) understands that health information is not identifiable (i.e., de-identified) only when there is no reasonable basis to believe that the information can be used to identify an individual and meets federal requirements for de-identified data; (ii) only creates and uses information that is not individually identifiable (i.e., de-identified) when a code or other means of record identification designed to enable coded or otherwise de-identified information to be re-identified is not disclosed; and (iii), if the de-identified information is subsequently re-identified, the covered entity only uses or discloses such re-identified information as permitted or required for PHI.
Aggregated Data: Disclosure Policy	13.I Minimum Necessary Use	De-ID Segment	The covered entity or business associate only publishes or discloses data that is de-identified for the intended context (environment), unless otherwise permitted by law.
Retention: Data Retention Policy	06.c Protection of Organizational Records	2	The organization establishes a formal record's document retention program.

DCA Category Name	CSF Control Name	CSF Level	CSF Assessment Requirement Statement
Retention: Data Retention Policy	06.c Protection of Organizational Records	1	The organization's formal policies and procedures, other critical records and disclosures of individuals' protected health information made are retained for a minimum of 6 years; and, for electronic health records, the organization retains records of disclosures to carry out treatment, payment and health care operations for a minimum of 3 years.
Retention: Data Retention Policy	09.aa Audit Logging	2	Audit records are retained for 90 days and archived for 1 year.
Disposal: Data Destruction Procedures	08.l Secure Disposal or Re-Use of Equipment	1	Surplus equipment is stored securely while not in use, and disposed of or sanitized when no longer required.
Disposal: Data Destruction Procedures	08.l Secure Disposal or Re-Use of Equipment	1	Electronic and physical media containing covered information is securely sanitized or, if it cannot be sanitized, destroyed prior to reuse or disposal.
Extra-national Access Restrictions: General	09.e Service Delivery	De-ID Segment	If required in the applicable jurisdiction, health information including de-identified data is not accessed from off-shore; nor is such data received, stored, processed or disposed via information technology systems located off-shore. Otherwise the entity must justify the off-shore disclosure.
Third-party Assurance: General	05.k Addressing Security in Third Party Agreements	1	The organization maintains written agreements (contracts) that includes (i) an acknowledgement that the third party (e.g., a service provider) is responsible for the security of the data and requirements to address the associated information security risks associated.
Risk Assessments (of Information Systems): Assessments	03.b Performing Risk Assessments	1	The organization performs risk assessments in a consist way and at planned intervals, or when there are major changes to the organization's environment, and reviews the risk assessment results annually.
Risk Assessments (of Information Systems): Assessments	03.a Risk Management Program Development	De-ID Segment	The organization (i) documents and implements a privacy risk management process that assesses privacy risk to individuals resulting from the collection, sharing, storing, transmitting, use, and disposal of personally identifiable information (PII); and (ii) conducts Privacy Impact Assessments (PIAs) for information systems, programs, or other activities that pose a privacy risk in accordance with applicable law, OMB policy, or any existing organizational policies and procedures.
Governance: General	05.a Management Commitment to Information Security	3	An information security management committee is chartered and active.

DCA Category Name	CSF Control Name	CSF Level	CSF Assessment Requirement Statement
Data Storage: General	06.d Data Protection and Privacy of Covered Information	1	The confidentiality and integrity of covered information at rest is protected using an encryption method appropriate to the medium anywhere it is stored, or documentation is maintained; or, where the organization chooses not to encrypt covered information, provides a documented rationale for not doing so.
Data Storage: General	07.a Inventory of Assets	1	An inventory of assets is maintained.
Data Storage: General	06.d Data Protection and Privacy of Covered Information	1	The organization specifies where covered information can be stored.
Privacy and Security Program: General	0.a Information Security Management Program	1	The organization has a formal information protection program based on an accepted industry framework that is reviewed and updated as needed.
Privacy and Security Program: General	06.d Data Protection and Privacy of Covered Information	1	An organizational data protection and privacy program is developed, implemented, and communicated to all persons involved in the processing of covered information, supported by management structure and control, IAW relevant legislation and regulations.
Privacy and Security Training: General	02.e Information Security Awareness, Education, and Training	1	Employees, and contractors, and other workforce members, including collaborating organizations, receive documented initial (as part of their onboarding within 60 days of hire), annual and ongoing training on their roles related to security and privacy.
Privacy and Security Training: Records	02.e Information Security Awareness, Education, and Training	3	The organization maintains a documented list of each individual who completes the on-boarding process and maintains all training records for at least 5 years.
Sanctions: General	02.f Disciplinary Process	1	The organization employs a formal sanctions process for personnel failing to comply with established information security policies and procedures, including license, registration, and certification denial or revocation and other disciplinary actions, and notifies defined personnel (e.g., supervisors) within a defined time frame (e.g., 24 hours) when a formal sanction process is initiated, identifying the individual sanctioned and the reason for the sanction.
Privacy Officer (and/or Data Stewardship Committee): General	06.d Data Protection and Privacy of Covered Information	1	The organization has formally appointed a data protection officer responsible for the privacy of covered information and ensuring awareness of specific data protection principles in accordance with relevant legislation and regulations.
Data Breach Response: General	11.a Reporting Information Security Events	2	The organization adheres to the HIPAA Omnibus requirements for responding to a data breach (of covered information) and reporting the breach to affected individuals, media and federal organizations.

DCA Category Name	CSF Control Name	CSF Level	CSF Assessment Requirement Statement
Data Breach Response: General	11.a Reporting Information Security Events	De-ID Segment	Entities receiving de-identified data notify the providing organization's data custodian of breaches involving Patient De-identified data as required by law for breaches of Patient Identifiable data, so that the providing organization can determine the appropriate response.
Privacy Reviews/Audits (Internal/External): General	05.h Independent Review of Information Security	1	An independent review of the organization's information protection program is initiated by management to ensure the continuing suitability, adequacy, and effectiveness of the organization's approach to managing information security and privacy.
Privacy Reviews/Audits (Internal/External): General	05.h Independent Review of Information Security	2	An independent review of the information security management program and information security controls is conducted at least annually or whenever there is a material change to the business practices that may implicate the security or integrity of records containing personal information.
Identification and Authentication: Authentication Policy	01.q User Identification and Authentication	1	Shared/group and generic user IDs are only used in exceptional circumstances where there is a clear business benefit, when user functions do not need to be traced and additional accountability controls are implemented, and after approval by management.
Identification and Authentication: Authentication Policy	01.q User Identification and Authentication	1	Non-organizational users (all information system users other than organizational users, such as patients or customers), or processes acting on behalf of non-organizational users, determined to need access to information residing on the organization's information systems, are uniquely identified and authenticated.
Identification and Authentication (System-level): Authentication Policy	01.q User Identification and Authentication	1	The organization ensures that redundant user IDs are not issued to other users and that all users are uniquely identified and authenticated for both local and remote accesses to information systems.
Identification and Authentication (Application-level): Authentication Policy	01.c Privilege Management	1	Privileges are formally authorized and controlled, allocated to users on a need-to-use and event-by-event basis for their function role, and documented for each system product/element.
Remote Access: Applicability	01.j User Authentication for External Connections	1	Strong authentication methods such as multi-factor, Radius or Kerberos (for privileged access) and CHAP (for encryption of credentials for dialup methods) are implemented for all external connections to the organization's network.
Remote Access: Applicability	01.j User Authentication for External Connections	1	Remote access by vendors and business partners (e.g., for remote maintenance) is disabled/deactivated when not in use.
Remote Access: Applicability	01.j User Authentication for External Connections	1	Remote access to business information across public networks only takes place after successful identification and authentication.

DCA Category Name	CSF Control Name	CSF Level	CSF Assessment Requirement Statement
Anti-malware: General	09.j Controls Against Malicious Code	1	Protection against malicious code is based on malicious code detection and repair software, security awareness, and appropriate system access and change management controls.
Anti-malware: General	09.j Controls Against Malicious Code	1	Anti-virus and anti-spyware are installed, operating and updated on all devices to conduct periodic scans of the system to identify and remove unauthorized software.
Audit Logging/Monitoring: General	09.aa Audit Logging	1	A secure audit record is created for all activities on the system (create, read, update, delete) involving covered information.
Audit Logging/Monitoring: General	09.aa Audit Logging	1	Audit records include the unique user ID, unique data subject ID, function performed, and date/time the event was performed.
Audit Logging/Monitoring: General	09.ab Monitoring System Use	2	Monitoring includes privileged operations, authorized access, unauthorized access attempts, and system alerts or failures.
Audit Logging/Monitoring: Aberrant and Inappropriate Use	09.ab Monitoring System Use	3	Alerts are generated for technical personnel to analyze and investigate suspicious activity or suspected violations.
Transmission Encryption: Applicability	06.d Data Protection and Privacy of Covered Information	De-ID Segment	Covered information is encrypted in transit whether internal or external to the organization's network, and, if not encrypted in transit, the organization must document its rationale.
Transmission Encryption: Policies	01.n Network Connection Control	2	Transmitted information is secured and, at a minimum, encrypted over open, public networks.
Transmission Encryption: Policies	09.m Network Controls	2	The organization ensures information systems protect the confidentiality and integrity of transmitted information, including during preparation for transmission and during reception.
Physical Security: General	08.b Physical Entry Controls	2	A restricted area, security room, or locked room is used to control access to areas containing covered information, and is controlled accordingly.
Physical Security: General	01.h Clear Desk and Clear Screen Policy	1	Covered or critical business information is not left unattended or available for unauthorized individuals to access including on desks, printers, copiers, fax machines, and computer monitors.
Physical Security: General	08.g Equipment Siting and Protection	1	Information assets handling covered information are positioned and the viewing angle restricted to reduce the risk of information being viewed by unauthorized persons during their use, and storage devices secured to avoid unauthorized access.
Physical Access: Identification Policy	08.b Physical Entry Controls	2	Visible identification of employees, visitors, contractors and third parties is required to clearly identify the individual.

DCA Category Name	CSF Control Name	CSF Level	CSF Assessment Requirement Statement
Physical Access: Identification Policy	08.b Physical Entry Controls	2	The organization ensures onsite personnel and visitor identification (e.g., badges) are revoked, updated when access requirements change, or terminated when expired or when access is no longer authorized, and all physical access mechanisms, such as keys, access cards and combinations, are returned disabled or changed.
Physical Access: Inappropriate Use	08.b Physical Entry Controls	2	An audit trail of all physical access is maintained.
Visitor Access: Policy	08.b Physical Entry Controls	1	Visitor and third party support access is recorded and supervised unless previously approved.
Visitor Access: Incidents	11.a Reporting Information Security Events	De-ID Segment	Visitor-related incidents are tracked and corrective actions are taken when they occur.
Perimeter Security (Alarms): General	08.b Physical Entry Controls	3	Doors to internal secure areas lock automatically, implement a door delay alarm, and are equipped with electronic locks.
Perimeter Security (Alarms): General	08.b Physical Entry Controls	3	Intrusion detection systems (e.g., alarms and surveillance equipment) are installed on all external doors and accessible windows, the systems are monitored, and incidents/alarms are investigated.
Perimeter Security (Alarms): General	08.b Physical Entry Controls	3	The organization actively monitors unoccupied areas at all times and sensitive and restricted areas as appropriate for the area and investigates alarms.
Perimeter Security (Alarms): Testing	08.b Physical Entry Controls	3	Alarms are regularly tested to ensure proper operation.
Perimeter Security (Alarms): Testing	02.e Information Security Awareness, Education, and Training	3	Workforce members are trained on how to properly respond to perimeter security alarms (see 08.b, level 3).
Perimeter Security (Alarms): Logging	08.b Physical Entry Controls	3	The organization maintains an electronic log of alarm system events and regularly reviews the logs no less than monthly.
Storage (Minimal Locations Authorized): Policy	06.d Data Protection and Privacy of Covered Information	1	Covered information storage shall be kept to a minimum.
Storage (Minimal Locations Authorized): Implementation	02.e Information Security Awareness, Education, and Training	2	The organization trains its workforce to ensure covered information is stored in organization-specified locations.
Storage (Minimal Locations Authorized): Implementation	06.d Data Protection and Privacy of Covered Information	2	The organization implements technical means to ensure covered information is stored in organization-specified locations.
Public Access to Sensitive Areas: General	08.a Physical Security Perimeter	2	Security perimeters, such as any boundaries where security controls are in place to protect assets from unauthorized access, are clearly defined, and the siting and strength of each of the perimeters depend on the security requirements of the assets within the perimeter (public, sensitive and restricted areas).

DCA Category Name	CSF Control Name	CSF Level	CSF Assessment Requirement Statement
Public Access to Sensitive Areas: General	08.c Securing Offices, Rooms, and Facilities	2	Critical facilities are sited to avoid access by the public; and, for particularly sensitive and restricted facilities, buildings are unobtrusive and give minimum indication of their purpose.
Video Surveillance: General	08.c Securing Offices, Rooms, and Facilities	2	Video cameras or other access control mechanisms are securely implemented and protected ; the output/results are reviewed regularly and correlated with other entries and access control information; and the information is stored for at least six (6) months IAW the retention policy.
Physical and Environmental Security: General	08.g Equipment Siting and Protection	1	The organization plans the location or site of the facility where the information system resides with regard to physical and environmental hazards and for existing facilities, and considers the physical and environmental hazards in its risk mitigation strategy.
Physical and Environmental Security: General	12.b Business Continuity and Risk Assessment	1	Information security aspects of business continuity are (i) based on identifying events (or sequence of events) that can cause interruptions to the organizations critical business processes (e.g. equipment failure, human errors, theft, fire, natural disasters and acts of terrorism); (ii) followed by a risk assessment to determine the probability and impact of such interruptions, in terms of time, damage scale and recovery period; (iii) based on the results of the risk assessment, a business continuity strategy is developed to identify the overall approach to business continuity; and (iv) once this strategy has been created, endorsement is provided by management, and a plan created and endorsed to implement this strategy.
Physical and Environmental Security: General	12.c Developing and Implementing Continuity Plans Including Information Security	1	The organization can recover and restore business operations and establish an availability of information in the time frame required by the business objectives and without a deterioration of the security measures.
Physical and Environmental Security: General	12.c Developing and Implementing Continuity Plans Including Information Security	2	Alternative storage and processing sites are identified (permanent and/or temporary), at a sufficient distance from the primary facility and configured with security measures equivalent to the primary site, and the necessary third party service agreements are established to allow for the resumption of information systems operations of critical business functions within the time-period defined (e.g. priority of service provisions) as defined by a risk assessment (see 12.b), including Recovery Time Objectives (RTO), IAW the organization's availability requirements.

DCA Category Name	CSF Control Name	CSF Level	CSF Assessment Requirement Statement
Physical and Environmental Security: General	12.c Developing and Implementing Continuity Plans Including Information Security	2	The alternate processing site is configured so that it is ready to be used as the operational site supporting essential missions and business functions of the organization and the specific type of configuration is defined by a risk assessment (see 12.b), acceptable alternatives for which include cold, warm, hot and mobile sites.
Accountable Individuals: General	05.a Management Commitment to Information Security	1	A senior-level information security official is appointed and is responsible for ensuring security processes are in place, communicated to all stakeholders, and consider and address organizational requirements.
Accountable Individuals: General	05.c Allocation of Information Security Responsibilities	1	The organization's senior-level information security official coordinates, develops, implements, and maintains an organization-wide information security program and assigns specific responsibilities, coordinated and aligned with internal and external partners.
Accountable Individuals: Policies	06.d Data Protection and Privacy of Covered Information	1	The organization makes the names and contact information for its senior-level information security and privacy officer(s) publically available.
Security Points of Contact: General	05.a Management Commitment to Information Security	1	Security contacts are appointed by name for each major organizational area or business unit.
Transparency: General	05.j Addressing Security When Dealing with Customers	1	The public has access to information about its privacy activities and is able to communicate with its senior privacy official.
Complaints: Policy	13.b Rights to Protection and Confidentiality	1	The covered entity provides for individual complaints concerning the covered entity's privacy policies and procedures or its compliance with such policies and procedures.
Secondary Use: General	13.e Authorization or Opportunity Not Required	1	The covered entity uses or discloses PHI for secondary use (e.g., research) only if approved by a valid IRB or privacy board or other oversight body and receives appropriate representations from the secondary user (e.g., researcher) regarding the appropriate uses and disclosures necessary for research purposes.
Audit and Monitoring: Auditing	13.i Required Uses and Disclosures	De-ID Segment	Audits of the use and disclosure of covered information are regularly conducted and any identified issues are remediated.
Audit and Monitoring: Monitoring	13.i Required Uses and Disclosures	De-ID Segment	The use or disclosure of covered information is monitored, and such monitoring is supported by automated alerting and response plans.
Data Stewardship: General	13.e Authorization or Opportunity Not Required	1	Documentation for a use or disclosure permitted for research or other secondary use based on approval of an alteration or waiver shall contain a signed, dated statement from the IRB, or privacy board or other oversight body that confirms the necessary conditions for use or disclosure.

Appendix B: Scoring

Area	Assessment	Notes
Program		
1. Governance		
2. Documentation		
3. Explicit Identification of the Data Custodian and Recipients		
4. External or Independent Scrutiny		

De-identification Methodology		
1. Re-Identification Risk Thresholds		
2. Measurement Of Actual Re-Identification Risks		
3. Identification And Management Of Direct Identifiers And Quasi-Identifiers		
4. Identification Of Plausible Adversaries And Attacks		
5. Identification Of Specific Data Transformation Methods And How They Reduce The Risks		
6. Process And Template For The Implementation Of Re-Identification Risk Assessment And De-Identification		
7. Mitigating Controls To Manage Residual Risk		
8. Data Utility		

For more information on



go to www.HITRUSTalliance.net