

---

WEDI - Strategic National Implementation Process (SNIP)

# Security and Privacy Workgroup



SNIP - Security and Privacy Workgroup  
“White Papers” DRAFT Version 3.1  
**De-Identification**

December 2001

***Workgroup for Electronic Data Interchange***

*12020 Sunrise Valley DR., Suite 100, Reston, VA. 20191*

*(t) 703-391-2716 / (f) 703-391-2759*

# Contents

- Introduction** **2**
- Disclaimer ..... 2
  
- De-identification** **3**
- Background ..... 3
- Definition ..... 3
- Objectives ..... 4
- Implementation Considerations and Issues ..... 5
  - Modification of the NPRM in the Final Regulations ..... 5
  - Processes ..... 6
- Definition of Terms ..... 7
- Costs and Resources ..... 8
- Risk Considerations ..... 9
- Other Sources of Information ..... 10
- Acknowledgements ..... 10

# Introduction

---

## Disclaimer

This document is Copyright © 2001 by The Workgroup for Electronic Data Interchange. It may be freely redistributed in its entirety provided that this copyright notice is not removed. It may not be sold for profit or used in commercial documents without the written permission of the copyright holder. This document is provided “as is” without any express or implied warranty.

While all information in this document is believed to be correct at the time of writing, this document is for educational purposes only and does not purport to provide legal advice. If you require legal advice, you should consult with an attorney. The information provided here is for reference use only and does not constitute the rendering of legal, financial, or other professional advice or recommendations by the Workgroup for Electronic Data Interchange. The listing of an organization does not imply any sort of endorsement and the Workgroup for Electronic Data Interchange takes no responsibility for the products, tools, and Internet sites listed.

The existence of a link or organizational reference in any of the following materials *should not* be assumed as an endorsement by the Workgroup for Electronic Data Interchange (WEDI), or any of the individual Security and Privacy Workgroup members of the Strategic National Implementation Process (SNIP).

## Document is for Education and Awareness Use Only

The HIPAA Security and Privacy requirements are designed to be ubiquitous, technology neutral and scalable from the very largest of health plans, to the very smallest of provider practices. As the Privacy Rule and a majority of the proposed Security Rule relates to policies and procedures, many covered entities will find compliance not an application of exact template processes or documentation, but rather a remediation based on a host of complex factors unique to each organization.

The work in Version 3 was completed before the Privacy Guidance of July 6, 2001, was released by the Department of Health and Human Services. The next version of the white papers will include review of the guidance document.

# De-identification

---

## Background

The HIPAA notice of proposed rule making (NPRM) for Privacy was discussed by a broad cross-section of the health care community during the WEDI SNIP Forum held on September 27-28, 2000. Participants in the Privacy Workgroup's deliberations sought to identify those parts of the proposed rule that could pose significant challenges to effective and timely implementation of the standards. One of the six (6) key issues was the definition of "de-identification." As a result, the de-identification subgroup of the security and privacy workgroup was established.

The task of the de-identification subgroup is to investigate the provisions in the proposed privacy regulation related to the "de-identification" standards. This white paper addresses the issues involved in implementing these standards.

The privacy standards are applicable to "individual identifiable health information" and no longer applicable to health information that has been de-identified. It is important to be able to determine those elements of identifiable information as to remove, code, encrypt or conceal this information in a way to be able to use and disclose the information more freely. The intent was to support greater use of de-identified information provided the use or disclosure of such information would not result in the disclosure of protected individual information.

There is support in the approach to expand the use and disclosure of information through the concept of de-identification of information, as it is an approach used in other sectors.

---

## Definition

The NPRM requires a covered entity and their business partners to use and disclose de-identified protected health information in any way provided *they reasonably believe* that such use or disclosure will not result in the use or disclosure of protected health information. As stated in the proposed NPRM rule [§164.506(d)(2)(ii), (1999 *Federal Register* page 59936)]:

We propose that there be a presumption that if specified identifying information is removed and if the holder has no reason to believe that the remaining information can be used by the reasonably anticipated recipients alone or in combination with other information to identify an individual, then the covered entity is presumed to have created de-identified information.

This requirement included two approaches to de-identify information:

- "Reasonableness standard" that entities with sufficient statistical experience and expertise can remove or code a different combination of information as long as the result has a low probability of identification.

- Removal of a list of 19 specified items of information with no reason to believe the remaining information could be used to identify the individual.

The rule requires that each covered entity must:

- Create de-identified information that removes the listed 19 elements but still must remove additional elements if there was reason to believe that the remaining information by itself or in combination with other available information can identify an individual.
- Not allow use or disclosure if there is reason to believe the individual can be re-identified
- Not share re-identification key, algorithm or method with the recipient of the de-identifiable information.

Once information is re-identified the information again falls under the rules for protected health information. A covered entity that re-identifies information for a purpose for which is protected health information cannot be used or disclosed will be deemed in violation. The use and disclosure of de-identified information is regulated by the rules for covered entities but the does not cover the use and disclosure of re-identified information by business partners. Therefore the covered entity must not share information that can be used to re-identify the transferred information.

---

## Objectives

This subgroup, as well as the Security and Privacy Workgroup and WEDI SNIP focused on the specific issue of de-identification as a subset of efforts to implement *all* HIPAA final rules. The task of the de-identification subgroup is to provide guidance regarding how to interpret and implement the requirements to de-identify protected health information to be able to further use and disclose such information.

The subgroup objectives focused on the following issues raised in the NPRM related to de-identifying information:

- There is ambiguity regarding whether particular protected health information is individually identifiable
- There is ambiguity regarding the phrase “no reason to believe” that the de-identified information was unable to be re-identified and identify an individual after the required elements were removed.
- There were no specific guidelines as to the documentation that may be required to demonstrate what principals or methods a statistically experienced person made to demonstrate that the risk was small that the de-identified information could be re-identified.
- The two methods of de-identifying information did allow for the covered entity to have decision-making control about what methods to apply to the de-identify protected information, however to remove all 19 elements as the one method required would result in limited use and value of the information. The covered entity must make a determination as to the balance between risk of identification and the usefulness of the resulting information.

The covered entity must consider how the information might be used in combination with other information and that the standards do not regulate entities that may receive the information and attempt to link information.

- The definition of what is considered individually identifiable is not always clear. As an example, a diagnostic code could be considered “identifiable” in a research report if only one member of the sample group has that particular condition. When releasing information, it is a good idea to evaluate what is to be released on a case-by-case basis as well as establishing standard definitions.
- The ease of removal of identifiers for disclosure would result in less burden and complexity to the smaller healthcare provider who may not have the statistical expertise or technical capability. The removal of specified identifiers reduces the amount of judgment that would be required in the process and is more measurable in limiting liability.
- There are data elements that appear in free text fields that may contain identifiable elements and pose difficulty technologically to find and remove. Clinical data in free text includes results in histories and physicals, nursing notes, progress notes, discharge summaries and other records.

Covered healthcare entities will be required to establish procedures and processes to ensure compliance with the HIPAA Privacy standards. Technology, available resources/skill sets and the complexity of a healthcare entity will influence the approach and solutions used in and meeting compliance. A GAP analysis will determine the current status of processes and procedures used in the use and disclosure of protected health information. Protected health information that is de-identified will not fall under the privacy rules for use and disclosure. Processes and procedures used to de-identify protected information in a manner that follows the implementation specification requirements for de-identification of protected health information must be met.

The final privacy regulations did modify the original NPRM that resulted in further clarification to most of the issues raised by the de-identification workgroup. The result of the subgroup’s efforts was to redirect efforts to the implementation of the final standards of de-identification of information.

---

## Implementation Considerations and Issues

Use of health information offers opportunities to improve healthcare and reduce overall costs. Efficiencies can occur with the use of a foundation of health information as a source for developing health policy, trends, and relationships of cost to outcomes. Aggregation of data is needed in this process and requires the sharing of information from various data sources. This sharing of information however needs to be done responsibly to maintain the privacy of individuals. The privacy Act of 1974 (Public Law 93-579) states, “The right to privacy is a personal and fundamental right protected by the Constitution of the United States.” Individual Privacy has evolved in the public’s view from a status to a right. At the same time that the public is demanding privacy rights, technology has raised both the risk and the expectations of the individual.

The perceived threat of disclosure of protected information is increased with the ease of electronic transmission and the demands for data by entity’s that may not directly be involved in the care of the patient. It is this balance of use and disclosure and the risk of disclosing an individual’s health information that needed to be addressed in the development of the final regulations.

## Modification of the NPRM in the Final Regulations

In 164.514 (a) – (c) several modification were made to the provisions of de-identification. The statutory standard was adopted as the basic regulatory standards for whether health information is individually identifiable health information. Information is not individually identifiable if:

- The individual is not identified or
- If the covered entity has no actual knowledge that can be used to identify the individual.

In the implementation specification, it specifies the two ways in which a covered entity can demonstrate that they have met the standard if:

1. **A person with appropriate knowledge and expertise:**

- Applies generally accepted statistical and scientific principles and methods for rendering information not individually identifiable
- Makes a determination the risk is very small that the information could be used by itself or in combination with other available information by the anticipated recipients
- Document the analysis and results in making determination

2. **Safe Harbor method:**

- Remove all of a list of enumerated identifiers, and
- No actual knowledge that the information that the information can be used alone or in combination to identify a subject of the information

## Processes

1. **What Processes are needed to evaluate whether protected information has been adequately de-identified? What choice of statistical method? Is the risk of disclosure from released data low?**

- Covered entities that release data will need to establish policies to establish whether protected information has been de-identified.
- A covered entity may choose to centralize oversight and review the application of statistical methods
- Centralized oversight may be through a committee that ensure that appropriate de-identification policies and practices are developed and provided to users
- The use of the disclosed data may determine whether the risk of disclosure is low
- The committee may also review to whom this data will be released and for what purpose
- May contract with business associates to perform de-identification

2. **How much information needs to be de-identified?**

Judgment may be determined by the following factors:

- Availability of external files with comparable data
- Resources needed to identify individual's information
- Sensitivity of the data elements
- Number of records within the file
- The study population such as geography, income, gender, minority

3. **What methods can be applied to assure minimal risk?**

- Include data from a sample of the population
- Do not include obvious identifiers
- Limit geographic detail
- Limit number of variables
- Recode into intervals or rounding
- Swap or rank swap
- Add or multiply by random numbers (known as noise)

- Aggregate across small groups and replace with ranges or averages

4. **What if I do not have the statistical expertise to make these determinations?**

The regulations allow application of Safe Harbor Methodology (see definitions) please state where the definitions can be found.

5. **How do I de-identify information in free text?**

This has been an issue in the field of pathology where there has been an interest in deriving information from report archives that involved the conversion of free text into a form that can be used as data within medical datasets. This has required skill in programming and computational linguistics and is an evolving effort. Until there is a cost effective technological solution, data within unstructured free text should be removed and may not have value for analysis due to the lack of analytic structure. If the data is of value, the data should be collected and documented in a way it can be located and de-identified or eliminated.

6. **Can a code be used that will allow for re-identification?**

The covered entity may use a code in order to re-identify information that has been de-identified by the covered entity. Once the information has been re-identified the information is then is protected health information under the regulations.

7. **Are there technological methods to de-identify information?**

There are currently generic engine frameworks for linguistic treatment of text applications. Potential applications of these engines include: syntax checking, and terminology extraction. These were originally designed to provide developers and researchers with common development architecture for the open and seamless integration of linguistic services. In the context of these research projects that involved biology and computer science laboratories, computer system using a linguistic approach combined with conceptual graph management tools to perform information extraction from scientific text databases are being developed. In the future these tools may be available for information extraction.

The regulations also allow covered entities to have information de-identified by a business associate.

## Definition of Terms

**Disclosure** means the release, transfer, provision of access to, or divulging in any other manner of information outside the entity holding the information. [Ref: Part 164 Subpart E 164.501]

**De-identified Information** is health information that does not identify and individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable information.[Ref: Part 164 Subpart E 164.514]

**Safe Harbor Methodology** is the method of removing a list of enumerated identifiers so that information can be used for many purposes with a very small risk of privacy violation and is intended to involve a minimum of burden: [Ref: Part 164 Subpart E 164.514]

**Allowed information:**

- Age with dates limited to the year
- Age over 90 and over must be aggregated to 90+
- Aggregated zip codes in the form of initial 3 digit zip codes to include at least 20,000 people

- Gender, race, ethnicity, marital status

Covered entities may use codes or similar means of marking records so they can be linked or later identified if the code does not contain subject information and the code is not used for any other purpose. However is prohibited from disclosing the mechanism for re-identification.

**Not Allowed** information that must be removed (including those of the individual, relatives, employers or household members of the individual):

- Names;
- All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip codes if the geographic unit of combining all the same three initial digits contains more than 20,000 people;
- If zip contains < 20,000 then changed to 00;
- 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 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
- 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.

---

## Costs and Resources

Within the final Privacy regulations, on page 1164, starting with Table C are listed the initial and ongoing annual cost estimates for various entity sizes and classification for meeting privacy compliance. These are the cost estimates related to de-identification.

<b>Average annual ongoing cost for de-identification per industry sector</b>	
<b>Industry</b>	<b>Cost</b>
Drug store and proprietary stores	\$3,751,011
Accident and Health Insurance and Medical Service Plans	\$3,614,697
Medical Equipment Rental and Leasing	\$174,710
Offices and Clinics of Doctors of Medicine	\$15,421,311
Offices and Clinics of Doctors of Dentists	\$6,913,022
Offices and Clinics of Doctors of Osteopathy	\$591,452
Offices and Clinics of Other Health Practitioners	\$3,392,310
Nursing and Personal Care	\$3,584,567
Hospitals	\$14,153,321
Medical and Dental Labs	\$1,004,715

Home Health Services	\$1,769,402
Misc. Health and Allied Health	\$997,304
<b>Total</b>	<b>\$55,367,822</b>

**Covered entities must establish process and procedures for de-identifying information** – An entity may decide to outsource this process to a business associate or retain the responsibility through oversight by a committee or staff member with statistical experience and training. Regardless of who is assigned the responsibility, processes and procedures must be established that ensure information that is de-identified is at a minimal risk for re-identification or disclosure of an individual’s information. These processes and procedures must be documented as written policies and procedures.

**Approaches for De-identification** – The approach used to remove identifiers from data may be scaled as determined by the size of the entity. Small entities can use the “safe harbor” process where larger entities may choose to use the statistical approach. The regulation will also allow a covered entity to contract with business associates to perform this function.

**Training Staff** – Staff will need to be trained to be aware that there are processes and procedures that must be followed for the use and disclosure of de-identified information. (This may be better handled by a centralized process that would function to minimize the risk that uninformed staff will not follow procedure.) This will need to be a business decision that will be influenced by the size and staff resources/skill level.

**Security** – The protection of individually identifiable health information requires the information to remain secure unless it is no longer protected health information. HIPAA privacy regulations require proper security mechanisms and policies be put in place to maintain the privacy as required. Advisement and review by a committee of both technical and non-technical staff will assist in controlling the use and disclosure of protected health information.

The covered entity does not use or disclose the code or other means of record identification for any other purpose and does not disclose the mechanism for re-identification. Any code or other means of record identification is not be a derivative of an individual’s information and is not capable of being translated as to identify the individual, for example, a subpart of a social security number.

---

## Risk Considerations

A health care entity’s will need to decide the amount of risk they are willing to accept and make business decisions accordingly. The results of a GAP analysis will assist in determining the level of risk as an enterprise as well as at a department level. This will also provide an evaluation of the current data flow and the release of protected health information to both business associates and non-covered entities. Future strategic business decisions that involve the sharing of protected health information should involve those who will be given the responsibility of oversight. New business relationships should outline the limitations that regulate the sharing of protected health information and whether de-identified information will meet the purpose.

- A. **Size of the Enterprise:** The more complex and large the healthcare enterprise the larger is the risk that information may be released without proper de-identification. Education of risk and policy and procedure will become critical. Smaller providers will be held to the same standards, but may not have access to educational information and knowledge.
- B. **Educated Resources:** Some entities may have statisticians on staff to assist in policy and procedure development, oversight and determinations on what will constitute de-identified protected information based on statistic methods. Those entities that do not have such resources

- may determine to use “safe harbor” methods as their standard. (or hire an outside vendor to do policies and procedures?)
- C. **De-identification Criteria**: Each covered entity will have to develop policies and procedures to define what is de-identified protected health information for the business entity. Documentation of the criteria and method used for this process will be required.
  - D. **Decision Making**: Covered entities have different management teams that release information from their various departments. The decision-making to de-identify protected information may be allowed by the various managers with oversight of a committee or may need to assure such requests are forwarded to the responsible committee. Regardless of where the decision-making lies, the management team will need to be fully trained as to the requirements and risk surrounding the disclosure of protected health information and what is necessary to deem it de-identified.
  - E. **Documentation**: A covered entity needs to document all of the initiatives taken to comply with the de-identification re-identification standards. These initiatives should include process and procedure development, training programs and attendance, implementation and compliance monitoring.
  - F. **State Regulations**: With certain limited exceptions, the privacy provisions preempt state law if: 1) the HIPAA provisions are more stringent than the state laws; 2) it would be impossible for a covered entity to comply with both the state law and the HIPAA regulations; or 3) the state law creates an obstacle to accomplishment of the goals of HIPAA. Covered entities must be aware of state laws which are more stringent than the HIPAA provisions and must meet the higher standards in every instance. (This is being addressed by the Preemption workgroup.)
  - G. **Cost**: The expected costs for compliance are in relation to resources within the organization from various departments. The costs however need to be weighed against the costs associated with disclosure of individual’s protected health information. Not only are there costs associated with the penalties there are publicity costs as well.

---

## Other Sources of Information

*Any other URLs, papers or organizations that would be a resource for this subject need to be identified and included in the paper.*

Statistical Policy Working Paper-Reports on Statistical Disclosure [http://www.fcsm.gov/working\\_papers](http://www.fcsm.gov/working_papers)

U.S. Census Bureau’s Suggestions Concerning Census 2000 <http://www.ipums.org>

American Health Information Management Association – <http://www.ahima.com>

HHS Administrative Simplification Web Site – <http://aspe.os.dhhs.gov/admsimp>

CPRI – [www.cpri-host.org](http://www.cpri-host.org)

Georgetown Health Privacy Project – [www.healthprivacy.org](http://www.healthprivacy.org)

HIPAAAdvisory – [www.hipaadvisory.com](http://www.hipaadvisory.com)

Health Information Management – [www.himinfo.com](http://www.himinfo.com)

---

## Acknowledgements

### White Paper Authors

WEDI/SNIP would like to express its appreciation to the authors for their efforts in preparing this White Paper:

Donna Steele, De-identification Sub-group Chairman  
Comdisco Healthcare Group

Michelle Chaudry, RHIA  
S3Networks, LLC