Deidentification Guide

This is a guide to resources and general principles for data deidentification in the context of the HeartShare project. Nothing in this document should be considered legal advice.

Version: 2025-03-07

# Introduction

In an effort to make this document general and comprehensive, material has been included that could be considered excessively detailed. This information is provided for those who want deeper background or more details.

# Resources

1. Federal law: the Health Insurance Portability and Accountability Act of 1996 (HIPAA)
2. “The HIPAA Privacy Rule establishes national standards to protect individuals' medical records and other individually identifiable health information (collectively defined as ‘protected health information’)” or PHI.

The HIPAA Privacy Rule defines PHI; defines how PHI can be used; and regulates research on human subjects. PHI includes MRNs, names, dates, locations, and others.

<https://www.hhs.gov/hipaa/for-professionals/privacy/index.html>

“The Privacy Rule is located at 45 CFR Part 160 and Subparts A and E of Part 164.”

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

(https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html)

1. Limited Data Set: “How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?”

<https://privacyruleandresearch.nih.gov/pr_08.asp>

“A limited data set is described as health information that excludes certain, listed direct identifiers (see below) but that may include city; state; ZIP Code; elements of date; and other numbers, characteristics, or codes not listed as direct identifiers.”

1. Limited Data Set FAQ

https://www.hhs.gov/hipaa/for-professionals/faq/limited-data-set/index.html

1. Limited Data Set (the HIPAA Journal)

https://www.hipaajournal.com/limited-data-set-under-hipaa/

1. Because the database is located in Illinois, we’re also required to comply with Illinois statutes regarding patient privacy, which define classes of “Special Protected Health Information” (SPHI). SPHI includes: mental health records; HIV/AIDS testing information; and genetic counseling information; but not genetic testing.

See Appendix A for sample SQL and lists of specific ICD-10 codes that should be excluded from data sets to comply with deidentification rules.

* Illinois Mental Health and Developmental Disabilities Confidentiality Act, 740 ILCS 110/1 et seq.
* <https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=2043>Illinois AIDS Confidentiality Act and related Regulations, 410 ILCS 305/1 et seq, and 77 IAC 697.10 et seq.

https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=1550

* Illinois Genetic Counselor Licensing Act, 225 ILCS 135/1 et seq.
* Illinois Genetic Information Privacy Act, 410 ILCS 513/1 et seq.
* Illinois Health Information Exchange and Technology Act 0f 2010

1. “Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule” (About 17,600 words)

https://privacyruleandresearch.nih.gov/pdf/HIPAA\_Privacy\_Rule\_Booklet.pdf

# Background

1. Text of Health Insurance Portability and Accountability Act of 1996:

<https://www.congress.gov/bill/104th-congress/house-bill/3103/text>

1. Summary of the HIPAA Privacy Rule

<https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html>

* “De-Identified Health Information. There are no restrictions on the use or disclosure of de-identified health information. De-identified health information neither identifies nor provides a reasonable basis to identify an individual. There are two ways to de-identify information; either: (1) a formal determination by a qualified statistician; or (2) the removal of specified identifiers of the individual and of the individual's relatives, household members, and employers is required, and is adequate only if the covered entity has no actual knowledge that the remaining information could be used to identify the individual.”
* **“Permitted Uses and Disclosures.** A covered entity is permitted, but not required, to use and disclose protected health information, without an individual's authorization, for the following purposes or situations: (1) To the Individual (unless required for access or accounting of disclosures); ... and (6) Limited Data Set for the purposes of research, public health or health care operations.”
* “**(6) Limited Data Set.** A limited data set is protected health information from which certain specified direct identifiers of individuals and their relatives, household members, and employers have been removed.43 A limited data set may be used and disclosed for research, health care operations, and public health purposes, provided the recipient enters into a data use agreement promising specified safeguards for the protected health information within the limited data set.”
* Note 43: Definition of Limited Data Set: “A limited data set is protected health information that excludes the following direct identifiers of the individual or of relatives, employers, or household members of the individual: (i) Names; (ii) Postal address information, other than town or city, State and zip code; (iii) Telephone numbers; (iv) Fax numbers; (v) Electronic mail addresses: (vi) Social security numbers; (vii) Medical record numbers; (viii) Health plan beneficiary numbers; (ix) Account numbers; (x) Certificate/license numbers; (xi) Vehicle identifiers and serial numbers, including license plate numbers; (xii) Device identifiers and serial numbers; (xiii) Web Universal Resource Locators (URLs); (xiv) Internet Protocol (IP) address numbers; (xv) Biometric identifiers, including finger and voice prints; (xvi) Full face photographic images and any comparable images.”

1. HIPAA Final Rule (January 25, 2013): Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules

https://www.govinfo.gov/content/pkg/FR-2013-01-25/pdf/2013-01073.pdf

1. HHS: Additional guidance on research:

https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/research/index.html

1. 42 US Code 1320d-6 – Penalties for Wrongful disclosure of individually identifiable health information: “A person who knowingly and in violation of this part ....”

<https://www.govinfo.gov/app/details/USCODE-2010-title42/USCODE-2010-title42-chap7-subchapXI-partC-sec1320d-6>

# Overview

The HHS-HIPAA Guidance lists two methods of deidentification: “Expert Determination” and “Safe Harbor”. Expert Determination requires a statistical review of deidentification methods to show that the risk of identifying an individual is “is very small”. It doesn’t specify any particular threshold or definition of “very small”. Compliance is sufficiently difficult that the Expert Determination method is rarely used in practice.

The Safe Harbor method is much more commonly used and generally known.

# General Guidance

* Guidance on Satisfying the Safe Harbor Method: https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html#safeharborguidance
* Deidentifying data using the Safe Harbor method requires removing “The following identifiers of the individual or of relatives, employers, or household members of the individual”. (Exceptions for Limited Data Sets are noted for each item.)

1. Names
2. All geographic subdivisions smaller than a state (first three digits of ZIP code may be used if it includes more than 20,000 people, as derived from the most recent census)

(ZIP code is allowed in a Limited Data Set)

1. All elements of dates (except year) ... 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

“For example, if the patient’s year of birth is 1910 and the year of healthcare service is reported as 2010, then in the de-identified data set the year of birth should be reported as “on or before 1920.”

(See "Limited Data Set", below.)

1. Telephone numbers
2. Fax numbers
3. Device identifiers and serial numbers
4. Email addresses
5. Web URLs (specific to an individual)
6. Social Security numbers
7. IP addresses (associated with an individual)
8. Medical record numbers
9. Vehicle identifier
10. Health plan beneficiary numbers
11. Account numbers
12. Certificate/license numbers
13. Any other unique identifying number, characteristic, or code (Not required for a Limited Data Set)

# Limited Data Set

The intent for this project is for the contributing sites to share Limited Data Sets (LDS), which will contain information still classified as identifiable protected information. LDS must not be shared except as permitted by a Data Use Agreement.

“The list of identifiers in a limited data set can include the town, city, or state of the individual, their gender, and dates relating to the individual.” (See HIPAA Journal.)

For HeartShare data sets, full dates (month, day, and year) may be used for all events EXCEPT birth date, which will be limited to the year of birth only. Any birth dates indicating an age of 90 or older must be replaced by the year "1900".

# Further notes

* “***Identifying Code:*** A code corresponds to a value that is derived from a non-secure encoding mechanism. For instance, a code derived from a secure hash function without a secret key (e.g., ‘salt’) would be considered an identifying element.”
* “***Identifying Characteristic:*** A *characteristic* may be anything that distinguishes an individual and allows for identification. For example, a unique identifying characteristic could be the occupation of a patient, if it was listed in a record as “current President of State University.”
* **Special handling of age and birth date**:
  + Any **ages 90 or older should be recorded as 90** (see item C above).
  + Although the rules for a limited data set allow retaining complete dates, to help assure patient privacy, **please limit birth dates to the year of birth** only.
  + Further, if the date when patient data is extracted, or any of the structured data items containing dates, is more than 90 years after the patient's date of birth (indicating age over 90, including fractions), the **date of birth should be recorded as either 1900 or NULL**. (Both of these are already present in our data.)
  + Date of birth should be expressed as either an integer data type or a character string, so that it can be limited to year only. If it's necessary to use a date or date-time data type which requires month and day, use January 1.
* We have found instances in which the result of a procedure or a lab test contains PHI, such as a date (for example, “20220103”) or even a patient name (“Homer Simpson’s knee surgery completed”). In addition to being on the look-out for these cases, generally make sure that values that should be numeric really are numeric, are in the expected range, and don’t conform to known patterns.

String values should be safe as long as they conform to OMOP vocabularies.

# Appendix A: SPHI exclusions

We've created some example SQL files demonstrating some ways to remove data restricted under Illinois law. These will be posted in the HeartShare GitHub repository

* Example SQL to exclude selected diagnosis codes: See file "HeartShare Example Exclude Diagnoses.sql"
* Example SQL to exclude selected medications: See file "HeartShare Example Exclude Medications.sql"