INSTITUTE FOR FAMILIES IN SOCIETY

Guidelines and Methods for De-Identifying Protected Health Information

Providing Actionable Information—Protecting Individual Privacy

Population health improvement and community development require quality data to inform policy creation, planning, programming, evaluation, and other critical decision-making processes. Notably, “data can either be useful or perfectly anonymous but never both” (Ohm, 2009). Ultimately, a balance must be struck between the utility and anonymity of data. The Institute for Families in Society (IFS) at the University of South Carolina uses Expert Determination to de-identify protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA). Under the HIPAA Privacy Rule, de-identified data are not considered protected health information and can be shared and used for any purpose. Data de-identification by Expert Determination can provide valuable information about populations and communities, while safeguarding the privacy of individuals.

Health Insurance Portability and Accountability Act

The Health Insurance Portability and Accountability Act (HIPAA), federally enacted in 1996, establishes mandates for the continuity of health insurance coverage for workers and their families in the event of a job loss or change of employment; establishes national provider identifiers; and sets standards for health care information collection, processing, and exchange. To ensure the privacy of individuals, HIPAA’s Privacy Rule protects individually identifiable health information *(referred to in this guide as Protected Health Information)* that is recorded, stored, or exchanged in any form or medium (e.g., on paper, orally, or electronically).

Protected Health Information

Protected health information (PHI) is any individually identifiable health information, including demographic information, that is created, used, disclosed, or received by a health care provider, health plan, or health care clearinghouse. PHI includes data about past, present, and future health conditions, health care provision, and health care payments.

Examples of PHI include name, address, telephone number, birth date, and Social Security number (HHS, 2015).

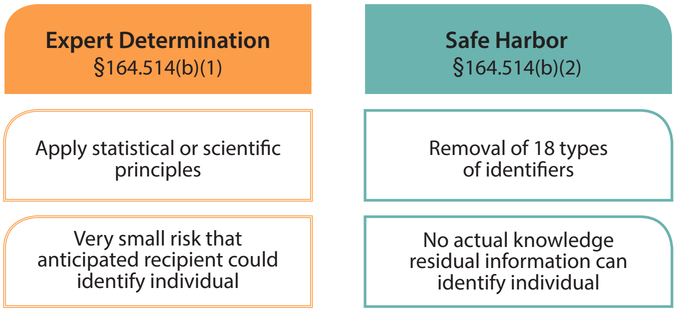
Data De-Identification Under HIPAA

HIPAA provides two mechanisms for the de-identification of protected health information: **Expert Determination and Safe Harbor** (Figure 1).

Data that have been de-identified by these methods are not considered PHI and can be shared and used for any purpose (HHS, 2015).

FIGURE 1.

HIPAA PRIVACY RULE DE-IDENTIFICATION METHODS



Expert Determination
§164.514(b)(1)

Apply statistical or scientific principles

Very small risk that anticipated recipient could identify individual

Safe Harbor
§164.514(b)(2)

Removal of 18 types of identifiers

No actual knowledge residual information can identify individual

Institute for Families in Society. (2022, May). Institute for Families in Society: Guidelines and methods for de-identifying protected health information [Infographic]. In *SC HealthViz*. University of South Carolina, prepared under contract to the SC Department of Health and Human Services. https://www.schealthviz.sc.edu/Data/Sites/1/media/images/USC_IFS_PHIDataDeIdentification_18.pdf

Expert Determination

The Expert Determination method, undertaken by persons with appropriate knowledge and experience, utilizes statistical and scientific principles to assess the risk that PHI could be used alone or in combination with other available data to identify specific individuals.

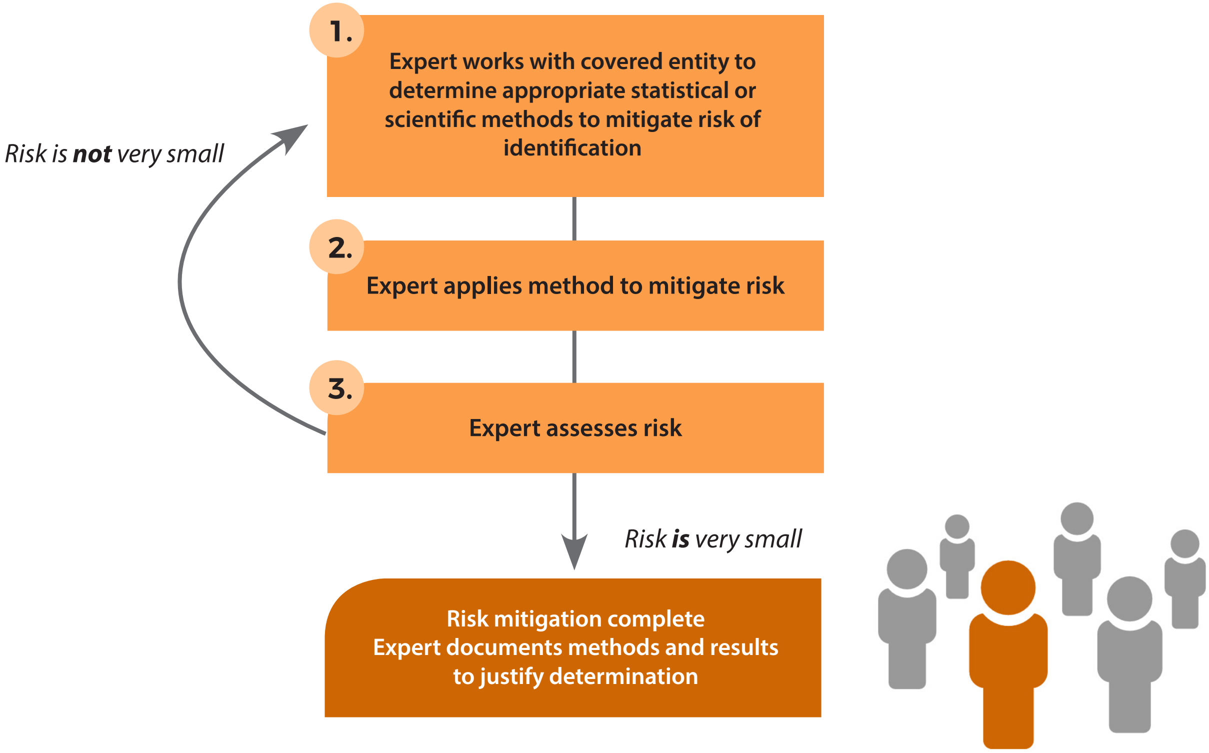
From Section 164.514(b)(1) of the Privacy Rule (HHS, 2015):

> "(b) Implementation specifications: requirements for de-identification of protected health information. A covered entity may determine that health information is not individually identifiable health information only if:
>
> > (1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
> >
> > > (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
> > >
> > > (ii) Documents the methods and results of the analysis that justify such determination."

At the core of Expert Determination is an assessment by qualified professionals that the "risk is very small" reported information, alone or in combination with other available data sources, could identify an individual who is the subject of the information. Expert Determination must be justified by means of documentation describing risk mitigation and risk assessment methods and results (Figure 2).

FIGURE 2.

EXPERT DE-IDENTIFICATION PROCESS
(Source: HHS, 2015)


1.
Expert works with covered entity to determine appropriate statistical or scientific methods to mitigate risk of identification
2.
Expert applies method to mitigate risk
3.
Expert assesses risk
Risk is **not** very small
Risk **is** very small
Risk mitigation complete
Expert documents methods and results to justify determination

Safe Harbor

Under HIPAA, PHI also may be de-identified using the Safe Harbor method. Safe Harbor requires the removal of all 18 types of identifiers specified by the Privacy Rule, Section 164.514(b)(2).

THOSE 18 TYPES OF IDENTIFIERS ARE:

Names

All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:

(1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and

(2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000

All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older

Telephone numbers

Vehicle identifiers and serial numbers, including license plate numbers

Fax numbers

Device identifiers and serial numbers

Email addresses

Web Universal Resource Locators (URLs)

Social Security numbers

Internet Protocol (IP) addresses

Medical record numbers

Biometric identifiers, including finger and voice prints

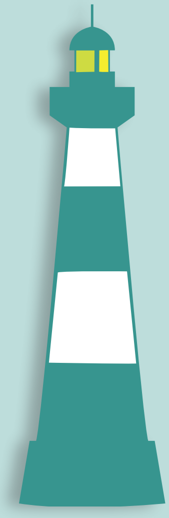
Health plan beneficiary numbers

Full-face photographs and any comparable images

Account numbers

Any other unique identifying number, characteristic, or code, except as permitted

Certificate/license numbers



Data from which these 18 identifiers have been removed are considered de-identified and therefore are not subject to the HIPAA Privacy Rule, unless actual knowledge exists that the remaining information could be used alone or in combination with other available data to identify an individual who is a subject of the information (HHS, 2015).

IFS Data De-Identification Guidelines

In developing actionable information products for population health and community stakeholders, IFS employs Expert Determination to de-identify PHI. Unless otherwise authorized under HIPAA rules, IFS typically aggregates (e.g., by age group, gender, race, and/or geographic unit) individual health information for reporting purposes. IFS's Aggregate Data De-Identification Decision Tree (Figure 3) guides the de-identification of PHI by Expert Determination in accordance with the HIPAA Privacy Rule.

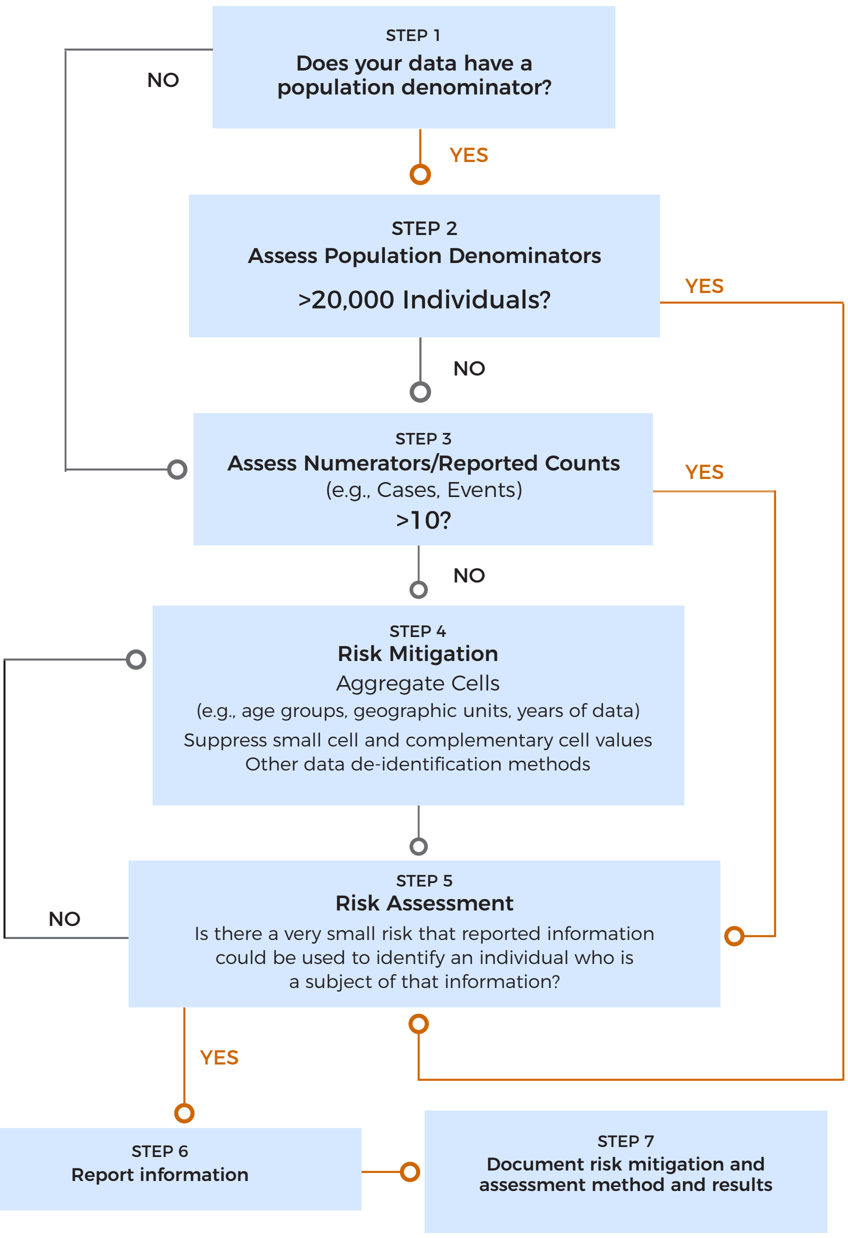
Consistent with the principles of Expert Determination, the Aggregate Data De-Identification Decision Tree emphasizes the assessment, mitigation, and re-assessment of risk that reported information could be used alone or in combination with other available data to identify an individual who is the subject of the information.

FIGURE 3.

IFS AGGREGATE DATA DE-IDENTIFICATION DECISION TREE

WHY >20,000 AND >10?

The numerator/reported count and denominator thresholds indicated (numerator >10; denominator >20,000) are general guidelines based on research and best-evidence practice (Sharkova, Zerbe, and Stone, 2015). Depending on the level of detail inherent in specific data sets, these thresholds may be adjusted to protect individual privacy.


STEP 1
Does your data have a population denominator?
NO
YES
STEP 2
Assess Population Denominators
>20,000 Individuals?
YES
NO
STEP 3
Assess Numerators/Reported Counts
(e.g., Cases, Events)
>10?
YES
NO
STEP 4
Risk Mitigation
Aggregate Cells
(e.g., age groups, geographic units, years of data)
Suppress small cell and complementary cell values
Other data de-identification methods
STEP 5
Risk Assessment
Is there a very small risk that reported information could be used to identify an individual who is a subject of that information?
NO
YES
STEP 6
Report information
STEP 7
Document risk mitigation and assessment method and results

Data De-identification Methods

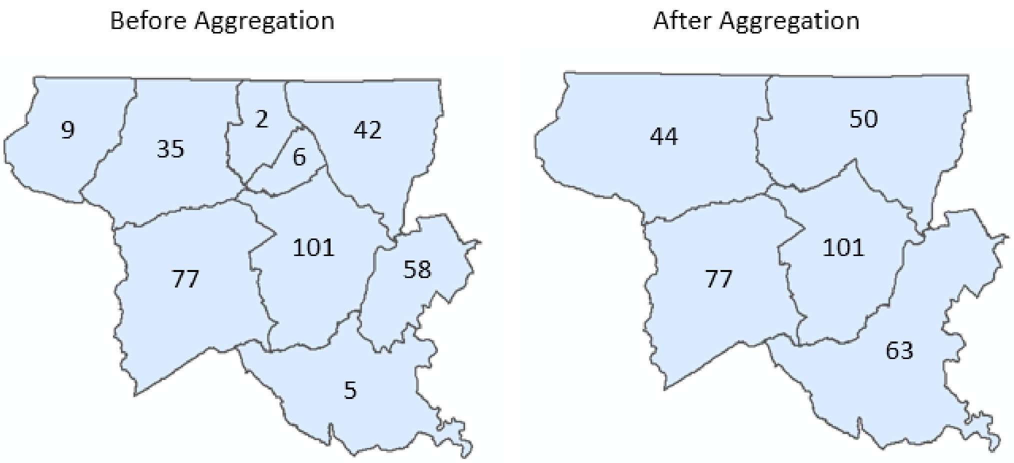
IFS uses a variety of statistical and geographic methods to mitigate the risk that any individual might be identified using reported data. The specific de-identification methods employed may vary depending on demographic, geographic, and temporal data detail, data sensitivity, and other data and project characteristics.

AGGREGATION

Population denominators and numerators/reported counts (e.g., cases, events) can be increased to minimum acceptable thresholds by aggregating tabular cells (e.g., age or racial/ethnic categories), geographic units, or time periods (e.g., months or years of data). Although the Aggregate Data De-Identification Decision Tree specifies a population denominator >20,000 and numerator >10 per cell, these thresholds are only guides, and may be adjusted depending on the detail and sensitivity of the data (Sharkova, Zerbe, and Stone, 2015). Aggregation of geographic units (Figure 4) may be assisted using the Geographic Aggregation Tool (GAT) or comparable software/geographic algorithms.

FIGURE 4.

Spatial Aggregation to Increase Number of Cases (e.g., Persons With a Chronic Condition) per Geographic Unit



SUPPRESSION OF SMALL CELL VALUES

Very small cell values (small Ns) also can be suppressed (Table 1). Often, it is necessary to suppress complementary cells as well to avoid the calculation of actual cell values by subtraction or other mathematical operations (Sharkova, Zerbe, and Stone, 2015; NCHS, 2004).

TABLE 1.
Example: Asthma Prevalence Rate per 1,000
The number of asthma cases and asthma prevalence per 1,000 persons are not reported (denoted by "--") for ZIP Code Tabulation Areas (ZCTAs) with ≤ 10 cases. Tabled data are hypothetical and for illustration purposes only.

Asthma Prevalence Rate per 1,000

ZCTA	ZCTA Name	Cases	Rate
29009	Bethune	--	--
29388	Woodruff	413	75.5
29532	Darlington	1,261	62.1
29541	Effingham	87	51.8
29546	Gresham	--	--
29625	Anderson	749	153.9
29923	Gifford	--	--

BLURRING

Blurring obscures data precision, thereby lessening the risk of individual identification (Privacy Technical Assistance Center, 2013). Examples of blurring include the conversion of discrete data values into data ranges (e.g., <20, 20-39, 40-79, 80 and above) and the transformation of continuous data into ordered categories (e.g., low, medium, high). Standard choropleth maps (employing color ramps to symbolize increasing/decreasing data values) typically use blurring in the establishment of mapped count (Figure 5) or rate data classes (Figure 6).

FIGURE 5.
Choropleth (Graduated Color) Map Showing Four Count Data Classes

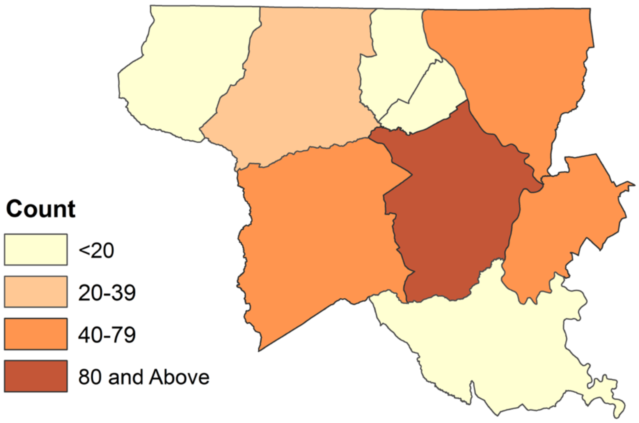
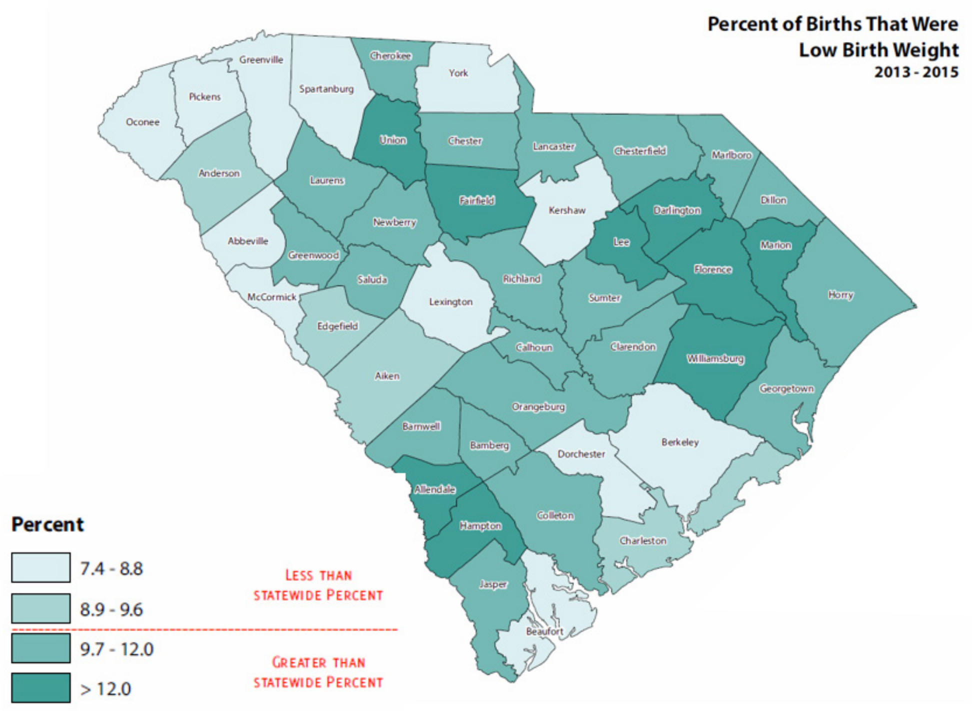


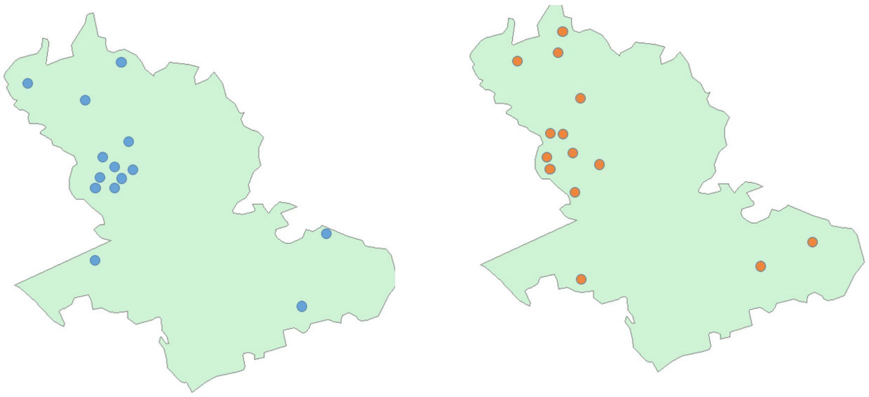
FIGURE 6.
Choropleth (Graduated Color) Map Showing Four Low Birth Weight Rate Classes
Additional blurring of data is achieved by combining three years of data (2013-2015).



PERTURBATION

To protect individual privacy, data uncertainty or error can be introduced intentionally by "swapping" certain cell values, for example, or by randomly misclassifying some data elements (Privacy Technical Assistance Center, 2013). On maps, point data representing address locations can be geographically altered to protect individual privacy, while preserving overall spatial patterns (Figure 7), a perturbation technique called dithering or random offsetting (Sharkova, Zerbe, and Stone, 2015).

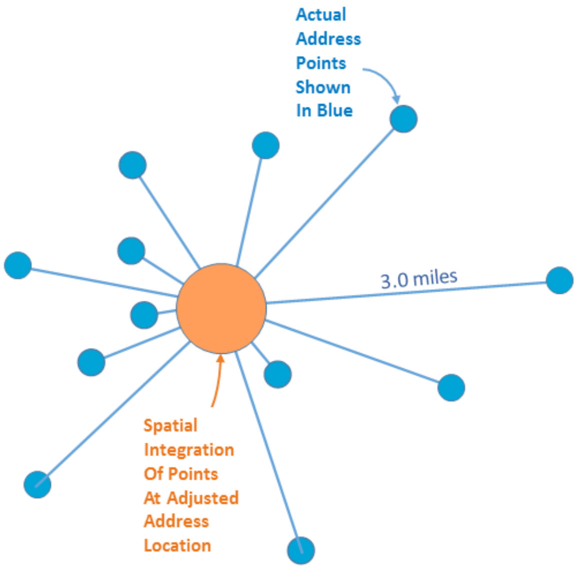
FIGURE 7.
Example of Random Offsetting
The perturbation of address data (shown in orange) obscures actual address locations (shown in blue), while preserving the overall spatial pattern.



SPATIAL INTEGRATION

Point locations also can be spatially aggregated to protect individual location privacy. In this process, called spatial integration, nearby point locations within a specified search radius (3 miles in Figure 8, for example) are "snapped together," or integrated, at a single, spatially adjusted location (shown in orange). Typically, the size of the integration point is graded (small to large) to reflect the total number of original points aggregated at that location.

FIGURE 8.
Example showing the spatial integration and location adjustment of nearby points (within a 3-mile search area).


Actual Address Points Shown In Blue
3.0 miles
Spatial Integration Of Points At Adjusted Address Location

Does Mapping Data at the County Level Ensure Individual Privacy Protection?

Health data (including demographic data) often are mapped at the county level. Commonly, it is assumed that the population of a county is sufficiently large to protect the identity of individuals for whom data are reported. County populations, however, vary substantially across the nation (Table 2). In Kansas, for instance, the smallest county has only 1,224 residents, while the largest has over 560,000. Population size in California counties is even more variable, ranging from only 1,131 to more than 10 million. Although 10% of all U.S. counties have more than 200,000 residents, more than 40% of counties have fewer than 20,000 residents.

TABLE 2.
County Population Variability in the U.S. and Four Sample States
(Source: 2015 ACS 5-Year Estimates)

State	US Census Region	# of Counties	Minimum County Population	Maximum County Population	Mean County Population	Standard Deviation
California	West	58	1,131	10,038,388	662,439	1,442,050
Kansas	Midwest	105	1,224	566,814	27,552	76,464
New Jersey	Northeast	21	65,120	926,330	424,020	252,530
South Carolina	South	46	9,838	474,903	103,860	112,986
United States		**3,142**	**85**	**10,038,388**	**100,737**	**322,983**

Population variability aside, mapping data at the county level does not ensure the confidentiality of PHI as required by the HIPAA Privacy Rule. In fact, Safe Harbor requires the removal of identifiers for all geographic subdivisions smaller than a state except for 3-digit ZIP Codes with a minimum of 20,000 people. Under Safe Harbor, county of residence is PHI and cannot be reported. When mapping data at the county level, it is necessary to use Expert Determination to assess and mitigate the risk that mapped data could be used to identify an individual who is the subject of the reported information. Notably, the Expert Determination method also can be used to de-identify data for mapping at smaller geographic scales (e.g., ZIP Code Tabulation Area, census tract, and census block group).

Summary

HIPAA provides two mechanisms to de-identify PHI: Expert Determination and Safe Harbor. Data that have been de-identified in accordance with HIPAA are no longer considered PHI and can be used for any purpose. Based on knowledge of and experience with statistical and scientific principles and methods, IFS employs Expert Determination to de-identify PHI. Data de-identification by Expert Determination can provide valuable information to strengthen policies and programs aimed at improved population health at state, county, and local levels, while safeguarding the privacy of individuals.

References

National Center for Health Statistics. (2004). *NCHS Staff manual on confidentiality.* Retrieved from https://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review, 57,* 1701-1777.

Privacy Technical Assistance Center. (2013). *Data de-identification: An overview of basic terms.* Retrieved from U.S. Department of Education website: http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf

Sharkova, I., Zerbe, J., & Stone, H. (2015). *Washington State Department of Social Services: Geospatial data confidentiality guidelines.* Retrieved from https://www.dshs.wa.gov/sites/default/files/SESA/rda/documents/DSHS%20Geospatial%20Data%20Confidentiality%20Guidelines%20-%2004May2015.pdf

U.S. Census Bureau. *American Factfinder. 2015 ACS 5-Year Estimates.* Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

U.S. Department of Health & Human Services. (2015). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* Retrieved from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Related Resources

Barth-Jones, D. C. *De-identification of confidential health data: Principles, methods, and policy; balancing privacy protection with scientific accuracy: Challenges for de-identification practice.* Retrieved from http://privacyoffice.med.miami.edu/documents/De-Identification_of_Confidential_Health_Data.pdf

Scott, L. (2014). *California Department of Health Care Services (DHCS): Public aggregate reporting—DHCS business reports guidelines.* Retrieved from http://www.dhcs.ca.gov/dataandstats/data/DocumentsOLD/IMD/PublicReportingGuidelines.pdf

U.S. Department of Health & Human Services (HHS). (2003). *OCR privacy brief: Summary of the HIPAA Privacy Rule.* Retrieved from https://www.hhs.gov/sites/default/files/privacysummary.pdf

Wiggins, L. (Ed). (2002). *Using geographic information systems technology in the collection, analysis, and presentation of cancer registry data: A handbook of basic practices.* Springfield, IL: North American Association of Central Cancer Registries.