Item #5: Ensuring Data Privacy, Confidentiality, and De-Identification

Chris Craig, Information Technology Manager, HCAI

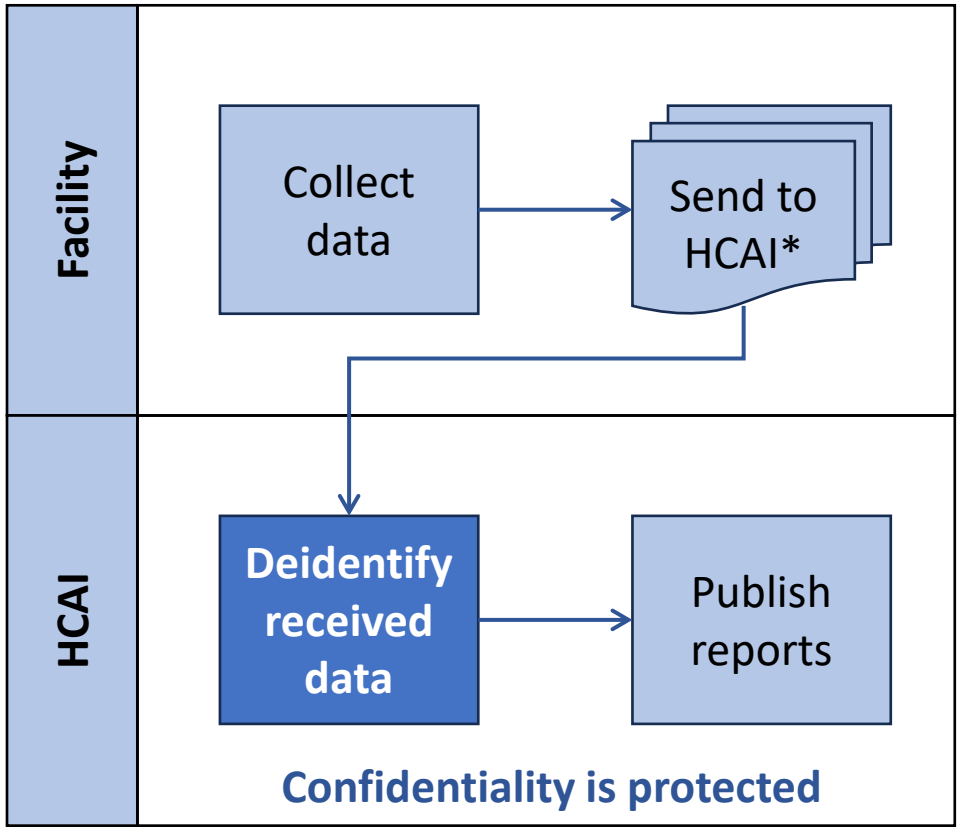
Ying Yang, Research Scientist Supervisor, HCAI

Protecting Patient Confidentiality

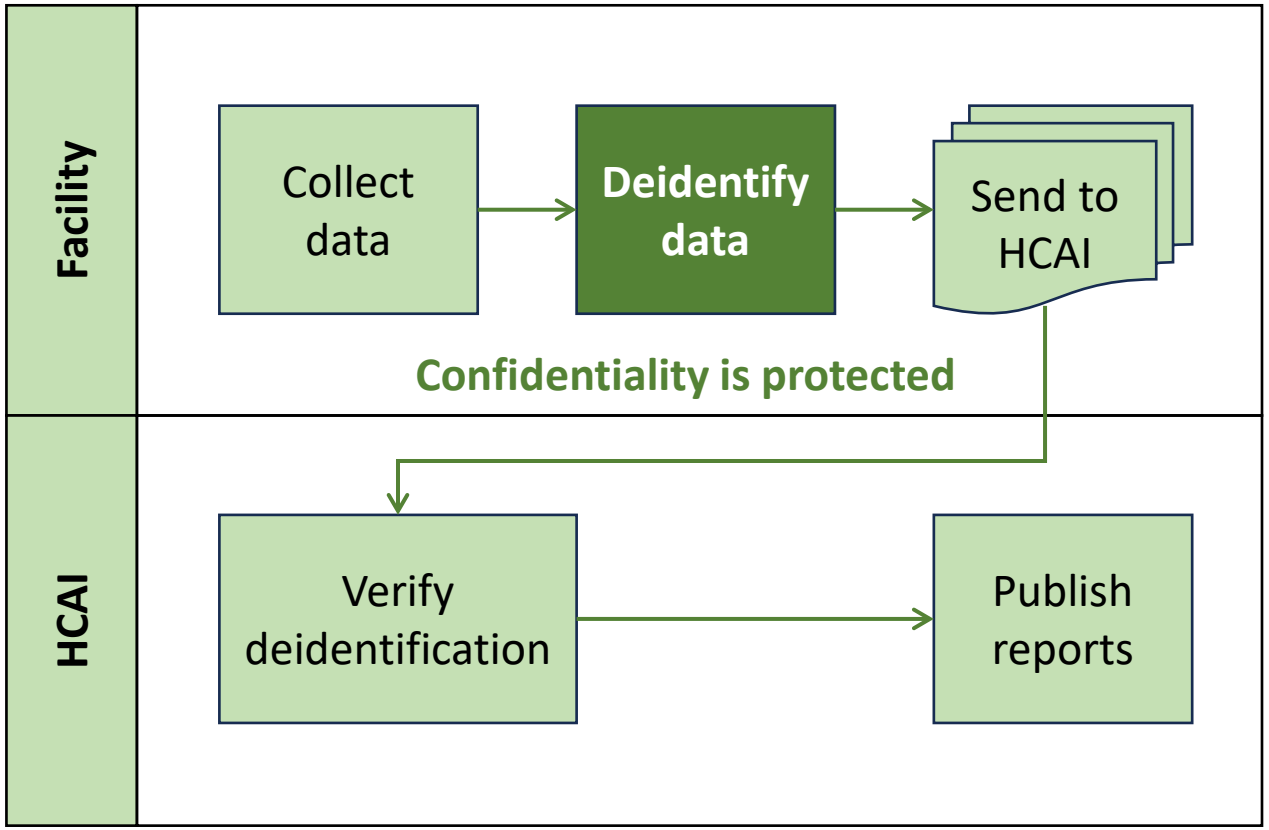
- Submitters of Hospital Equity Measures reports are subject to the federal Health Insurance Portability and Accountability Act (HIPAA), which provides strict, nationwide protections for individually identifiable health information
- Federal HIPAA regulations set limits on when, how, and to what extent individually identifiable health information may be shared
- Deidentification is the primary method to protect the privacy and confidentiality of patient data in Hospital Equity Measures reporting
    - Submitters are required to perform data deidentification prior to submitting reports to HCAI

Submission Workflow Comparison

Typical Data Reporting to HCAI



Hospital Equity Measures Reporting to HCAI



* Data connection is secure and encrypted, so confidentiality is protected during transmission

What Deidentification Means

- When data are properly deidentified, they no longer require protections under HIPAA
  - Deidentified data may be shared with the public and redistributed without restrictions
- Deidentification methods are used to create virtually all health and medical public reports, visualizations, research publications, and similar content intended for public use in the United States
- Following the long-standing federal model, the California Health and Human Services Agency (CalHHS) provides technical assistance through the publication of Data Deidentification Guidelines (DDG)
  - The current federal guidance has been in place since 2012[1]
  - The current CalHHS guidance has been in place since 2016[2]

1 https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html
2 https://chhsdata.github.io/dataplaybook/documents/CHHS-DDG-V1.0-092316.pdf

Data Deidentification

- HIPAA defines two methods of deidentification:
  - Expert determination method (45 CFR §164.514(b)(1))
  - Safe harbor method (45 CFR §164.514(b)(2))
- Hospital Equity Measures Program uses the expert determination approach
- HCAI must verify submitted reports have been properly deidentified (HSC §127374(e))

How Expert Determination Works

- The CalHHS DDG contains a list of risk factors that could be included in any set of individually identifiable health information
  - Characteristics, such as age, location, ethnicity, etc.
  - Criteria, such as interactions between multiple pieces of information
- Each risk factor is given a score
  - Lower scores indicate less risk of reidentification
  - Higher scores indicate greater risk of reidentification
- The analyst evaluates the data set and tallies up the DDG risk scores
  - A final score or 12 or less indicates the data are sufficiently deidentified
  - If the score is greater than 12, the data set must be adjusted to achieve proper deidentification

Data De-Identification Guidelines (DDG) for Hospital Equity Measures

- Reports shall include the numerator, denominator, and rate of each measure broken down by each stratification category, to the extent the data is available and consistent with the California Health and Human Services Agency's "Data De-Identification Guidelines (DDG)," dated September 23, 2016.
- If the stratification category is not reportable per DDG then indicate "suppressed". If data are not available, leave category blank.

Minimum Numerator and Denominator

The Publication Scoring Criteria method is used to assess the need to suppress small cells as a result of a small numerator, small denominator, or both small numerator and small denominator:

- Numerator is less than 11 (events)
- Denominator is less than 20,001

Publication Scoring Criteria

HEM variables specified in the Publication Scoring Criteria:

- Events (Numerator)
- Sex
- Age range
- Race/Ethnicity
- Language spoken
- Time – reporting period
- Service Geography
- Variable combination

Figure 6: Publication Scoring Criteria

Variable	Characteristics	Score
Events (Numerator)	1000+ events in a specified population	+2
	100-999 events	+3
	11-99 events	+5
	<11 events	+7
Sex	Male or Female	+1
Age Range	>10-year age range	+2
	6-10 year age range	+3
	3-5 year age range	+5
	1-2 year age range	+7
Race Group	White, Asian, Black or African American	+2
	White, Asian, Black or African American, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Mixed	+3
	Detailed Race	+4
Ethnicity	Hispanic or Latino - yes or no	+2
	Detailed ethnicity	+4
Race/Ethnicity Combined	This applies when race and ethnicity are collected in a single data field	
	White, Asian, Black or African American, Hispanic or Latino	+2
	White, Asian, Black or African American, Hispanic or Latino, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Mixed	+3
	Detailed Race/Ethnicity	+4
Language Spoken	English, Spanish, Other Language	+2
	Detailed Language	+4
Time – Reporting Period	5 years aggregated	-5
	2-4 years aggregated	-3
	1 year (e.g., 2001)	0
	Bi-Annual	+3
	Quarterly	+4
	Monthly	+5
Residence Geography*	State or geography with population >2,000,000	-5
	Population 1,000,001 - 2,000,000	-3
	Population 560,001 - 1,000,000	-1
	Population 250,000 - 560,000	0
	Population 100,000 - 250,000	+1
	Population 50,001 - 100,000	+3
	Population 20,001 - 50,000	+4
	Population ≤ 20,000	+5
Service Geography*	State or geography with population >2,000,000	-5
	Population 1,000,001 - 2,000,000	-4
	Population 560,001 - 1,000,000	-3
	Population 250,000 - 560,000	-1
	Population of reporting region 20,001 - 250,000	0
	Population of reporting region ≤20,000	+1
	Address (Street and ZIP)	+3
Variable Interactions	Only Events (minimum of 5), Time, and Geography (Residence or Service)	-5
	Only Events (minimum of 3), Time, and Geography (Residence or Service)	-3
	Only Events (no minimum), Time, and Geography (Residence or Service)	0
	Events, Time, and Geography (Residence or Service) + 1 variable	+1
	Events, Time, and Geography (Residence or Service) + 2 variable	+2
	Events, Time, and Geography (Residence or Service) + 3 variable	+4

* If the geography of the reporting is based on the residence of the individual, use the "Residence Geography". If the geography of the reporting is based on the location of service, use the "Service Geography".

Assess Potential Risk

- Quantifies based on two identification risks:
  - Size of potential population
  - Variable specificity
- Takes into account variables associated with numerators (events) and with denominators (e.g., geography)
- Score ≤ 12 data can be released without suppression (masking)
- Score > 12 requires suppression of cells with values < 11

Common Questions for DDG

- Common Questions:
  - Risk score calculations: The rate ratio should be calculated after applying the DDG.
  - Service geography for hospital reporting: Use the facility address for scoring
  - Service geography for hospital system reporting: Use estimated population for scoring