Agenda IV: Personally Identifiable Information (PII) and Privacy

Merry Holliday-Hanson, Research Scientist Supervisor, HCAI

For Today

- HIPAA and HCAI
- Current HCAI record-level data release practices
- Managing risk of re-identification
- Current HCAI aggregate data release practices
    - CalHHS Data De-identification Guidelines (DDG)

HIPAA and HCAI

- HCAI is not a HIPAA “covered entity”
    - HPD is not HIPAA protected
- Rely on HPD Statute
- Use HIPAA as a guide, where applicable
- Data release regulations
    - Rely on HPD statute
    - Borrow some HIPAA definitions
        - e.g., “Limited Dataset”, “Research Identifiable”

Types of HPD Data

- Non-Confidential Data – De-identified

- Limited Data – includes indirect patient identifiers, e.g., dates specific to individuals with dd/mm/yyyy detail, 5-digit Zip Code.

- Research Identifiable Data – includes direct patient identifiers, e.g., name, street address, other identifiers unique to individuals (SSN); requires CPHS review

Current HCAI Record-Level Data Release

- Limited Data (Standardized datasets)
    - Detailed application, review and approval process
    - Hospitals (submitters) and Local Health Departments
    - Data Use Agreement
        - IT Security requirements
        - Return/destruction of data
- Researcher Data
    - Detailed application, enhanced review and approval process
    - Committee for the Protection of Human Subject (CPHS) approval
    - Minimum data necessary - justify need for PII/PHI
    - Data Use Agreement
        - IT Security requirements
        - Return/destruction of data upon completion of project

Managing Reidentification Risk – Enclave and Direct Transmission

Enclave – Research Identifiable & Limited

- IT security requirements conform to state/federal
- **Controlled environment, use of system governed in DUA**
  - No data imported or exported without approval
  - Only de-identified products exported, must follow CalHHS DDG
  - User access – timeframe and data permissions
- DUA limits the use of the data
- Minimum data necessary – justify need for PII/PHI
- Violation of DUA subject to legal action and penalties
- **Research Identifiable – risk for reidentifying *other*, non-HPD is reduced**
- **DRC approval for Research Identifiable only**

Direct Transmission – Research Identifiable & Standard Limited

- IT security requirements conform to state/federal
- **IT security requirements for recipients' systems specified in regulations (and DUA)**
- DUA limits the use of the data
- Minimum data necessary - justify need for PII/PHI
- **DUA requires use of CalHHS DDG**
- **DUA requires documented destruction of data**
- Violation of DUA subject to legal action and penalties
- **Research Identifiable - risk is that it may be used to reidentify *other*, non-HPD data**
- **DRC approval for <u>any</u> Direct Transmission**

Current HCAI Aggregate Data Release

Non-Confidential Data

- Simple application for custom analysis to obtain aggregate data
- Follows CalHHS Data De-identification Guidelines

Applicable for HPD

- Enclave: All exported data must be evaluated for de-identification using DDG
- Direct Transmission: Requires use of DDG in DUA

Data De-Identification Guidelines

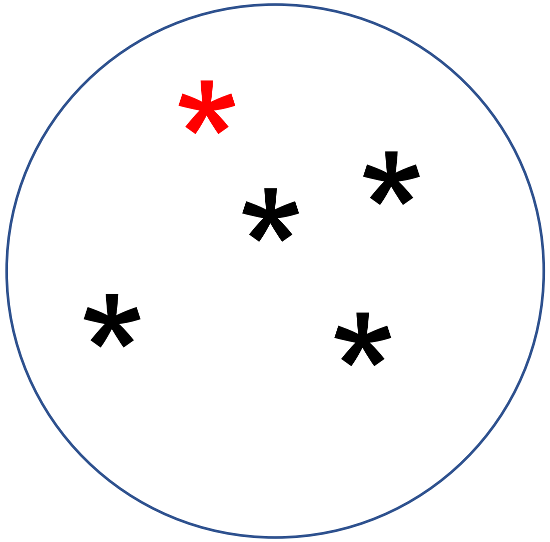
California Health and Human Services Data De-identification Guidelines (DDG)

- Used for all data released to the public (custom analysis, reports, Legislature, PRAs, etc.)
- To be used by all CalHHS departments/offices
- Meet requirements of the California Information Practices Act (IPA) and HIPAA to prevent disclosure of personal information (table p. 8 in DDG)
- Aggregate data – collective data that relates to a group or category of services or individuals
  - Counts, percentages, rates, averages, etc.
- *Reduce the risk of re-identification*

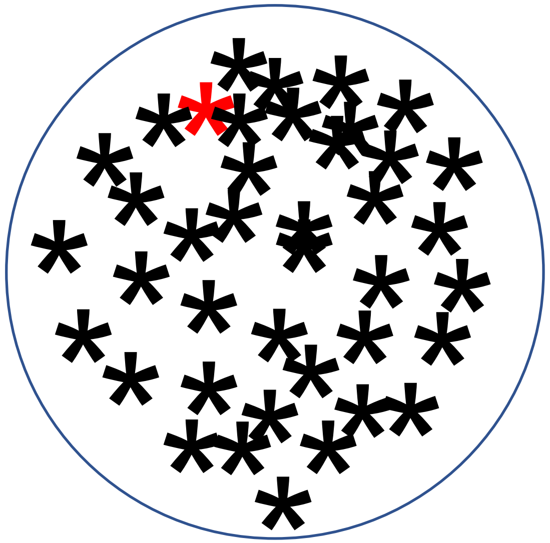
https://chhsdata.github.io/dataplaybook/documents/CHHS-DDG-V1.0-092316.pdf

DDG Background - Uniqueness

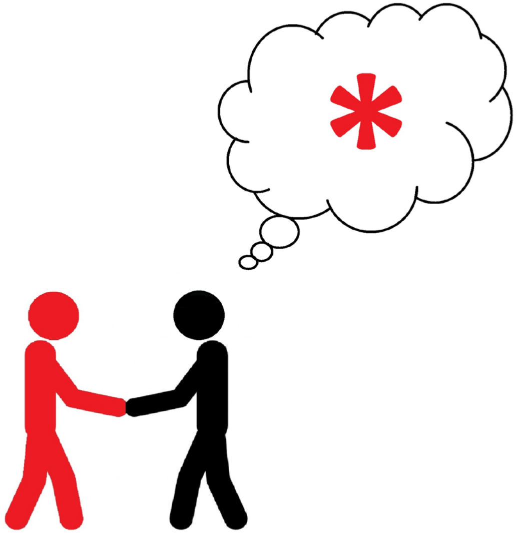
Aggregate Data File
n < 11



Aggregate Data File
n ≥ 11



DDG Background – Other Knowledge (data in the world)



EMSA
EMERGENCY MEDICAL SERVICES AUTHORITY
CALIFORNIA
DHCS
California Department of HealthCareServices
SAFETY TREATMENT RESPONSIBILITY
DEPARTMENT OF STATE HOSPITALS
California Department of Child Support Services
DOR
DEPARTMENT of REHABILITATION
Employment, Independence & Equality
CDPH
California Department of PublicHealth
CDSS
DDS
Department of Developmental Services
HCAI
Department of Health Care Access and Information
CENTER FOR DATA INSIGHTS AND INNOVATION
CALIFORNIA HEALTH & HUMAN SERVICES AGENCY
CSD
DEPARTMENT OF COMMUNITY SERVICES AND DEVELOPMENT
CALIFORNIA
DEPARTMENT OF Managed Health Care
CALIFORNIA DEPARTMENT OF AGING

Custom Data Request:

- Assault in a hospital by facility
  - Facility incident reports
  - Law enforcement/Court records

DDG Methodology

1. Personal Characteristics of Individuals
2. Numerator – Denominator Condition
3. Address Potential Risks
4. Statistical Masking
5. Legal Review
6. Departmental Release Procedures for De-Identified Data

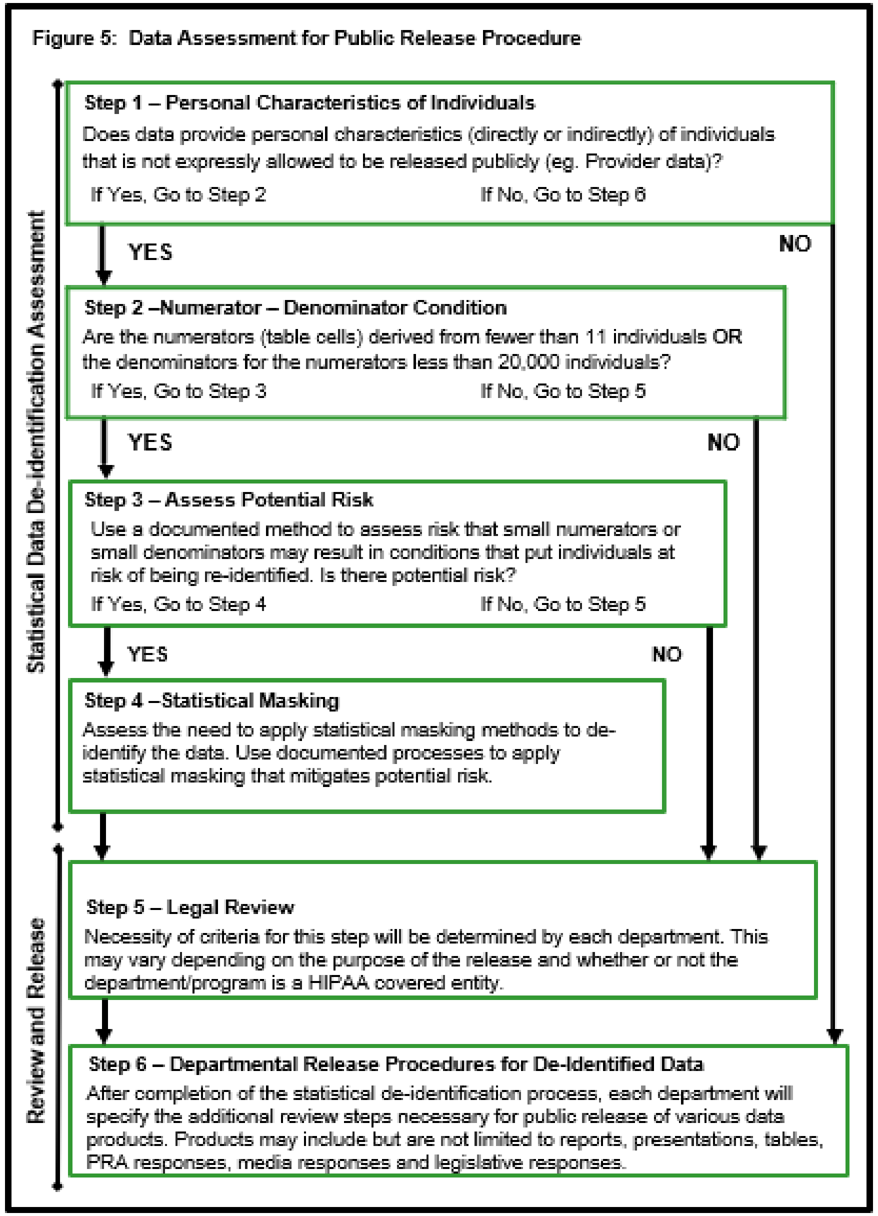


Figure 5: Data Assessment for Public Release Procedure

Step 1 – Personal Characteristics of Individuals
Does data provide personal characteristics (directly or indirectly) of individuals that is not expressly allowed to be released publicly (eg. Provider data)?
If Yes, Go to Step 2 If No, Go to Step 6

Step 2 –Numerator – Denominator Condition
Are the numerators (table cells) derived from fewer than 11 individuals OR the denominators for the numerators less than 20,000 individuals?
If Yes, Go to Step 3 If No, Go to Step 5

Step 3 – Assess Potential Risk
Use a documented method to assess risk that small numerators or small denominators may result in conditions that put individuals at risk of being re-identified. Is there potential risk?
If Yes, Go to Step 4 If No, Go to Step 5

Step 4 –Statistical Masking
Assess the need to apply statistical masking methods to de-identify the data. Use documented processes to apply statistical masking that mitigates potential risk.

Step 5 – Legal Review
Necessity of criteria for this step will be determined by each department. This may vary depending on the purpose of the release and whether or not the department/program is a HIPAA covered entity.

Step 6 – Departmental Release Procedures for De-Identified Data
After completion of the statistical de-identification process, each department will specify the additional review steps necessary for public release of various data products. Products may include but are not limited to reports, presentations, tables, PRA responses, media responses and legislative responses.

DDG Methodology

Step 1. Personal Characteristics of Individuals

Does data provide personal characteristics (directly or indirectly) of individuals that is not expressly allowed to be released publicly?

Step 2. Numerator – Denominator Condition

- Are the numerators (table cells) derived from fewer than 11 individuals

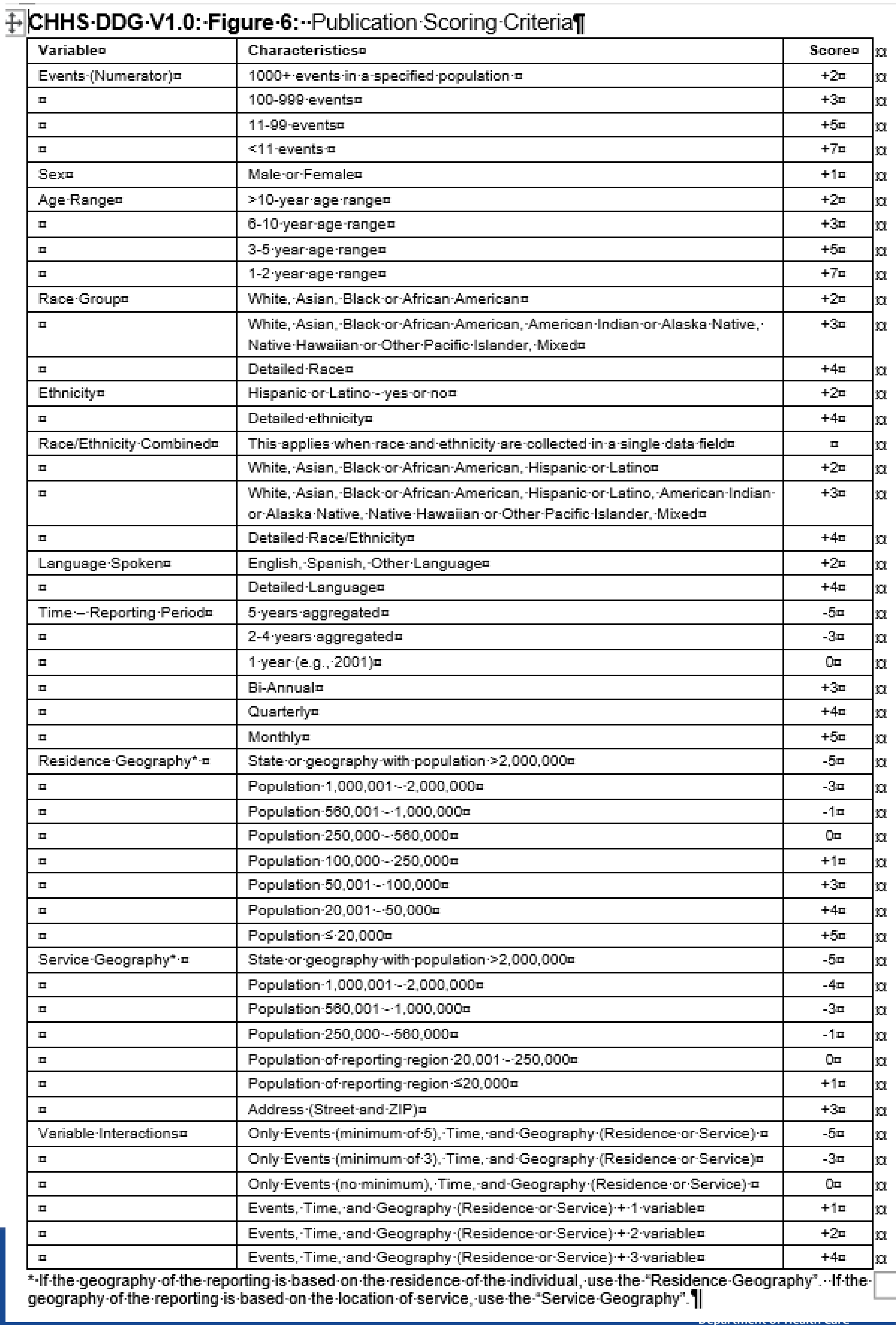
  OR

- The denominators for the numerators less than 20,000 individuals?

DDG Methodology

- **Step 3. Assess Potential Risk**
  - Use a documented method to assess risk that small numerators or small denominators may result in conditions that put individuals at risk of being re-identified.
  - **DDG Risk Scoring Criteria**
    - 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 masking
Score > 12 requires masking cells with values < 11



CHHS DDG V1.0: 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".

DDG Scoring Example

Amador County residents by age group (10 yr) with "Condition Y" by year, 2020 and 2021 (or 2020-2021)

Scoring	*2020/2021*	*2020-2021*
Events (<11)	+7	+7
Age (6-10 yrs)	+3	+3
Year (1) or (2-4)	0	-3
Geography (37,000)	+4	+4
Total	**+14**	**+11**

Score is >12, so proceed to Step 4

Score is ≤12, so no masking

DDG Methodology

Step 4. Statistical Masking

Assess the need to apply statistical masking methods to de-identify the data. Use documented processes to apply statistical masking that mitigates potential risk.

Step 5. Legal Review

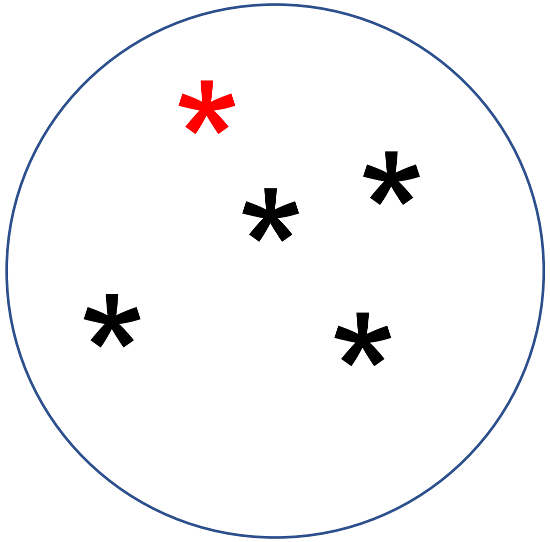
Necessity of criteria for this step will be determined by each department.

Step 6. Departmental Release Procedures for De-Identified Data

After completion of the statistical de-identification process, each department will specify the additional review steps necessary for public release of various data products.

Risk of reidentification and uniqueness

Aggregate Data File
$n < 11$



Aggregate Data File
$n \geq 11$