DATA DE-IDENTIFICATION GUIDELINES (DDG)

California Department of Health Care Services

November 13, 2025

Version 3.0

Based on California Health and Human Services Agency DDG Edition 2.0

CALIFORNIA DEPARTMENT OF
HEALTH CARE SERVICES

California Department of Health Care Services Revision History

Version	Date	Author	CalHHS Template	Brief Description of Changes
2.0	11/22/2016	L. Scott, MD, MPH	CalHHS DDG Template Edition applied, 1.0	Edits made to the CalHHS DDG 1.0 to reflect previously agreed to departmental processes as documented in the Public Aggregate Reporting - DHCS Business Reports (PAR-DBR) Guidelines V1.6.
2.1	01/08/2021	A. Valdez	CalHHS DDG Template Edition applied, 1.0	Edits made to the DHCS DDG V2.0 to reflect the title change of Chief Medical Information Officer to Chief Data Officer, along with miscellaneous formatting changes.
2.2	12/06/2022	D. Aggarwal, PhD	CalHHS DDG Template Edition applied, 1.0	Edits made to the DHCS DDG V2.1 to reflect updates approved by the CalHHS Peer Review Team and updated Appendix B to the current version listed on the Center for Data Insights and Innovation's website.
3.0	11/13/2025	D. Aggarwal, PhD	CalHHS DDG Template Edition applied, 2.0	Edits made to the DHCS DDG V2.2 to reflect updates approved by the CalHHS Peer Review Team and based on the CalHHS DDG Edition 2.0.

CalHHS DDG Template Edition History Summary

See Appendix J for the DDG Template's full development and revision history.

Edition	Date	Primary Author	Brief Description of Changes
1.0	9/23/16	L. Scott, MD, MPH	The First Edition was approved by CalHHS for implementation.
2.0	10/24/25	D. Aggarwal, PhD	The Second Edition was approved by CalHHS for implementation.

TABLE OF CONTENTS

1) PURPOSE

The State of California and the California Health and Human Services Agency (CalHHS) diligently protect the privacy of Californians while also ensuring transparency in government and making data available for research that benefits Californians and advances research related to health and human services. Recent changes in California law around sensitive information underscore the state’s commitment to securing the private medical and personally identifiable information of Californians, so they can continue to be free from unnecessary intrusions into their private lives. CalHHS’ update to the Data De-Identification Guidelines furthers California’s commitment to protecting Californians’ privacy while balancing the needs of the research and policy communities. CalHHS remains steadfast in its commitment to protecting the privacy of Californians while continuing to encourage the free flow of information to enrich the health and well-being of all Californians.

The California Department of Health Care Services (DHCS) Data De-Identification Guidelines (DDG) describes a procedure to be used by DHCS to assess data for public release and protect the privacy of individuals represented in the data in accordance with applicable state and federal laws. As part of the document, specific actions that may be taken for each step in the procedure are described. These steps are intended to assist departments in assuring that data is de-identified for purposes of public release that meet the requirements of the California Information Practices Act[1] (IPA) and the Health Insurance Portability and Accountability Act[2] (HIPAA) to prevent the disclosure of personal information.

Additionally, the DHCS DDG aligns with the CalHHS DDG which supports CalHHS governance goals to reduce inconsistency of practices across departments, align standards used across departments, facilitate the release of useful data to the public, promote transparency of state government, and support other CalHHS initiatives, such as the CalHHS Open Data Portal.

[1] Civ. Code § 1798 et seq.

[2] HIPAA Privacy Rule is located at 45 CFR Part 160 and Subparts A and E of Part 164.

2) BACKGROUND

CalHHS implemented an agency-wide governance structure in October 2014. The governance structure acts both in a decision-making and advisory capacity to Agency leadership and its departments and offices. Implementation of the governance framework supports information technology (IT) initiatives that are more tightly aligned with meeting business objectives, enhanced project prioritization and improved strategic IT investment decisions. The Executive Sponsor is the Undersecretary of CalHHS. The Interdepartmental Advisory Council (IDAC) consists of representatives of senior leadership from departments and offices in the Agency. There are seven subcommittees that report to IDAC, which include the Risk Management and Data Subcommittees.

- A Data De-Identification Workgroup was convened by the Data Subcommittee with representation from all departments and offices in CalHHS to develop the first Agency DDG. That first CalHHS DDG was approved for release in 2016.
- The Risk Management Subcommittee's DDG Workgroup, now known as the DDG Peer Review Team (PRT), developed the DDG Implementation Procedure (DDG-IP) and continues to support adoption and implementation by CalHHS departments.
    - The PRT membership of scientific researchers, privacy, security, or legal staff from CalHHS departments provide recommendations on topics related to the DDG-IP and DDG.

In addition, the PRT reviews and approves department DDGs and approves certain individuals to serve as a department's Statistical De-Identification Expert or Statistical De-Identification Supervisor Expert as indicated in the DDG-IP. CalHHS is engaged in improving transparency and public reporting. Data is Publishable State Data if it meets one of the following criteria: (1) data that are public by law such as via the Public Records Act[3] (PRA) or (2) the data are not prohibited from being released by any laws, regulations, policies, rules, rights, court order, or any other restriction. Data shall not be released if it is restricted under state or federal law. Data tables may fall into one of three categories:[4]

[3] Gov. Code 7920.000 et seq.

[4] CalHHS' Open Data Handbook, Version 1.1, October 2016, Data Levels Decision Tree, https://chhsdata.github.io/opendatahandbook/governance/

- Level One: Data tables that can be released to the public and published without restriction;
- Level Two: Data tables that have some level of restriction or sensitivity but currently can be made available to interested parties with a signed data use agreement; or
- Level Three: Level three data are restricted due to HIPAA, state, or federal law.

Data can change from Level 3 to Level 1 if appropriate de-identification processes are employed. The CalHHS DDG described in this document will support departments and offices in the evaluation of data to determine whether it has been adequately de-identified so that it can be considered Level 1.

Data sharing is very important for the DHCS. DHCS has adopted a commitment to hold ourselves and our providers, plans, and partners accountable for performance as part of our strategic plan. As part of this commitment, we have adopted a strategy to report publicly on our performance as a Department. DHCS has also made a strong public commitment to maintain a culture of privacy and security. All personally identifiable information is protected, and DHCS complies with federal law, specifically, the Privacy Rule and the Security Rule contained in the HIPAA and its regulations, 45 CFR Parts 160 and 164, and the Substance Abuse Confidentiality Regulations 42 CFR Part 2. DHCS is also committed to complying with California state privacy laws (e.g., Welfare and Institutions Code section 14100.2, the Information Practices Act, CA Civil Code section 1798, et seq.). In order to achieve both of these goals (public reporting and protection of personally identifiable information), procedures that appropriately and accurately de-identify data when publicly reporting are necessary.

The HIPAA Standard for De-identification of protected health information (PHI) states, "Health information that does not identify an 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 health information."[5] If the data are de-identified, and it is not reasonably likely that the data could be re-identified, the Privacy Rule no longer restricts the use or disclosure of the de-identified data.

[5] U.S. Department of Health & Human Services. "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

3) SCOPE

The DHCS DDG is based on the CalHHS DDG, which is focused on the assessment of aggregate or summary data for purposes of de-identification and public release. Aggregate data means collective data that relates to a group or category of services or individuals. The aggregate data may be shown in table form as counts, percentages, rates, averages, or other statistical groupings.

DHCS is sometimes asked to release record level data. Unlike aggregate or summary data, record level data refers to information that is specific to a person or entity. For example, a record for Jane Doe may include demographics and case information specific to Jane Doe. If record level data is to be publicly released, it must be assessed to ensure it is de-identified and does not include Personal Information (PI)[6] or Protected Health Information (PHI).[7] Although the DDG is focused on summarized data, it can be used to assist with review of individual or record level data. The record level data should be assessed both for uniqueness of the records and for the possibility that the data can be used in conjunction with other information available to the requester to identify individuals in the data and redacted based on the assessment. Record level data inherently has higher risk than summarized data, even after personal identifiers are removed. Therefore, record level data for public release should be assessed on a case-by-case basis.

DHCS collects, manages, and disseminates a wide range of data. The focus for the DDG is on data that includes personal characteristics of individuals who have a legal right to privacy. Personal characteristics include but are not limited to age, race, sex, and residence and other identifiers specified in the IPA and HIPAA and listed in Figure 1. These guidelines will focus on the assessment of personal characteristics that are included in various datasets or tables to assess risk for identification of the individuals to which they pertain.

[6] Personal Information is defined by California Civil Code section 1798.3 and Government Code section 11015.5.

[7] “PHI” is defined as information which relates to the individual’s past, present, or future physical or mental health or condition, the provision of health care to the individual, or the past, present, or future payment for the provision of health care to the individual, and that identifies the individual, or for which there is a reasonable basis to believe can be used to identify the individual. (45 CFR section 160.103)

Figure 1: Unique Identifiers

CA – Personal Information	HIPAA – Safe Harbor
Any information that identifies or describes an individual, including but not limited to:[8] » Name » Social security number » Physical description » Home address » Home telephone number » Education » Financial matters » Medical history » Employment history Electronically collected personal information:[9] » his or her name » social security number » physical description » home address » home telephone number » education » financial matters » medical or employment history » password » electronic mail address	» 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: ○ The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and ○ 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

[8] California Civil Code 1798.3 (a)

[9] California Government Code 11015.5 (d) (1)

CA – Personal Information	HIPAA – Safe Harbor
» information that reveals any network location or identity Excludes information relating to individuals who are users serving in a business capacity, including, but not limited to, business owners, officers, or principals of that business.	» Fax numbers » Email 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) addresses » Biometric identifiers, including finger and voice prints » Full-face photographs and any comparable images » Any other unique identifying number, characteristic, or code

Assessing the risk of an unauthorized disclosure that violates an individual’s right to privacy and/or confidentiality, as provided by statute, may be achieved by associating personal characteristics with a person’s identity or attributes. When these characteristics can successfully confirm an individual’s identity in a publicly released dataset, then release of this data results in disclosure of personal information.

Less obvious qualities in datasets and elements that may be used to identify individuals or groups can present uniqueness in data. Individual uniqueness in the released data and in the population is a quality that helps distinguish one person from another and is directly related to the re-identification of individuals in aggregate data. Disclosure risk becomes a concern when released data reveal characteristics that are unique in both the released data and in the underlying population. The risk of re-identifying an individual or group of individuals increases when unique or rare characteristics are “highly visible” or are readily accessible by the general public without any special or privileged knowledge. Unique or rare personal characteristics (e.g., height above 7 feet) or information that isolate individuals to small demographic subgroups (e.g., American Indian Tribal membership) increase the likelihood that someone can correctly attribute information in the released data to an individual or group of individuals.[10]

There are a variety of federal resources, such as the National Institute of Standards and Technology guide for “De-Identifying Government Datasets”,[11] that also address these risks.

3.1 Assessment of variables and their uniqueness

There are a number of variables that are unique to individuals that have been identified in various laws and are considered identifiers (PI/PHI). There are two primary laws that describe identifiers, shown in Figure 1, in California: the IPA and the federal HIPAA. Other variables that are commonly used to publish information to the public have been called quasi-identifiers because while they are not unique by themselves, they can become unique in the right combination. The variables shown in the Publication Scoring Criteria in Figure 6 can be considered quasi-identifiers and will be discussed further in Sections 4 and 16 (Appendix D).

3.2 Assessment of risk in the context of maximizing the usefulness of the information presented

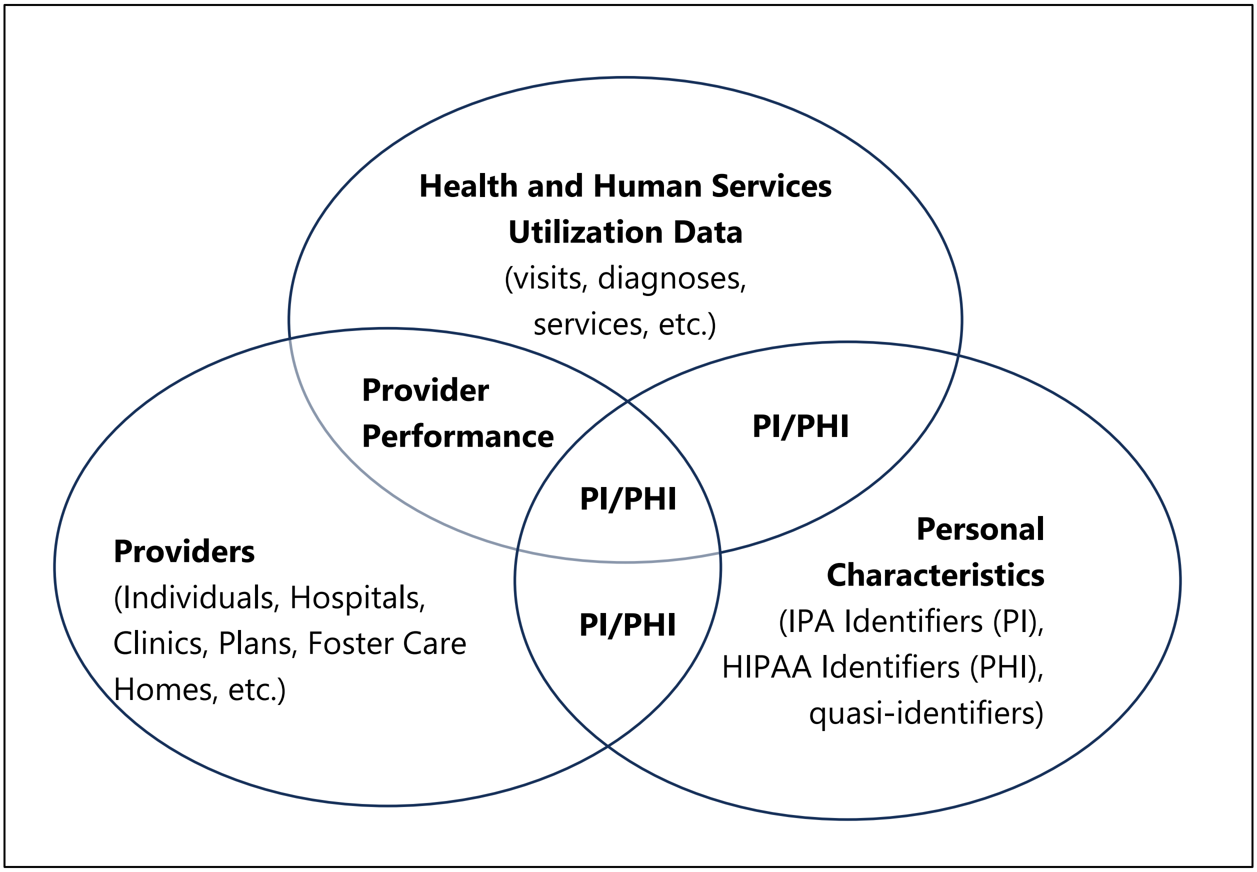
The removal of PI and PHI from datasets is often considered straight-forward, because as soon as data is aggregated or summarized the majority of the data fields defined as identifiers in the IPA and HIPAA are removed. However, various characteristics of individuals may remain that alone or in combination could contribute to identifying individuals. These characteristics have been described as quasi-identifiers. Figure 2 helps demonstrate the quasi-identifier concept. For instance, there is interest in reporting about providers, where providers may be individuals, clinics, group homes, or other entities. Each of these providers has a publicly available address and has publicly available characteristics. While patients may come to a provider from anywhere, they

[10] Introduction to Statistical Disclosure Control, Templ et al. 2014

[11] U.S. Department of Commerce, National Institute of Standards and Technology, “De-Identifying Government Datasets: Techniques and Governance”, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-188.pdf

typically will visit providers within a certain distance of their residence. Thus, by publicly publishing details on providers, data miners with malicious intent would have a targeted geography that lists locality information, types of services offered and received, and demographic information about patients. To expand on this example, data that states a provider saw two patients with heart disease does not indicate who had the heart disease nor does it reveal the identity of the two patients amongst the thousands of patients that provider sees. However, datasets that display a provider within a given region with two Black or African American female patients under age 10 with heart disease may release enough personal characteristics about the patients to successfully reveal their identity. These compounding patient details released about providers that give geography information (address), health condition (heart disease), and person-based characteristics (quasi-identifiers) of the patients put the dataset in the overlapping area of the diagram of Figure 2. This overlap, consequently, highlights potential risks associated with seemingly innocent summary data.

Figure 2: Relationship of Types of Reporting Variables


Health and Human Services Utilization Data
(visits, diagnoses, services, etc.)
Provider Performance
PI/PHI
PI/PHI
Providers
(Individuals, Hospitals, Clinics, Plans, Foster Care Homes, etc.)
PI/PHI
Personal Characteristics
(IPA Identifiers (PI), HIPAA Identifiers (PHI), quasi-identifiers)

4) STATISTICAL DE-IDENTIFICATION

The DDG describes the Data Assessment for Public Release Procedure shown in Figure 5, to be used by departments in the CalHHS to assess data for public release. Section 4 describes specific actions that may be taken for each step in the procedure with additional supporting information provided in Sections 5, 6 and 16. These steps are intended to assist departments in assuring that data is de-identified for purposes of public release that meet the requirements of the California IPA to prevent the disclosure of personal information.

The Data Assessment for Public Release Procedure includes the following steps:

Review the data to determine if it includes personal characteristics, directly or indirectly, that can be tied back to an individual;

1) If there is concern for personal characteristics, then assess the data for small numerators or denominators;
2) If there is concern for small numerators or denominators, assess potential risk of data release;
3) If there is potential risk identified, assess the need to apply statistical masking methods to de-identify the data;
4) Following statistical de-identification, the data release is reviewed by legal if indicated in departmental procedures; and,
5) After statistical de-identification, the data is reviewed and approved for release based on program and policy criteria pursuant to departmental procedures.

The steps above are represented in a step-wise process shown in Figure 5. Each step is described in further detail in Section 4.1 through 4.6.

Data summaries that originate from data which includes personal identifiers must be de-identified before release to the public. Additionally, data summaries about conditions experienced by individuals must be adequately de-identified to prevent re-identification of individuals represented by the summarized data. Various statistical methods are available to de-identify data statistically.

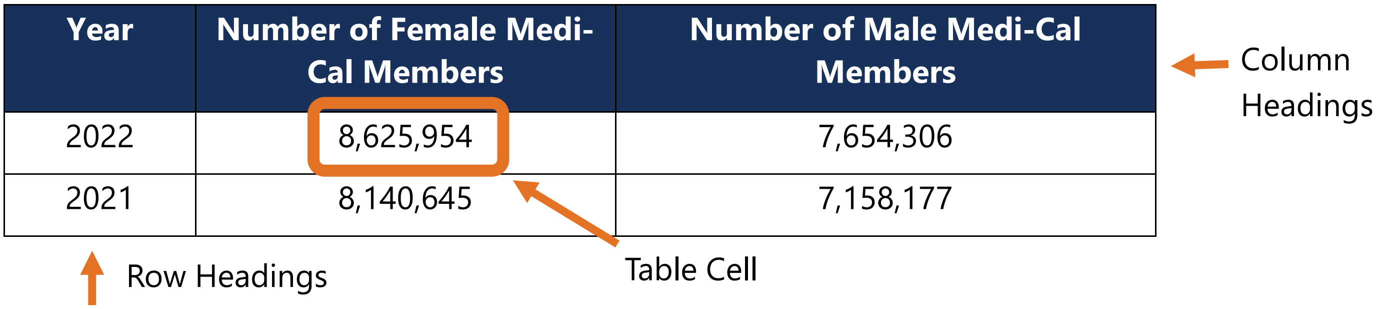
Summarized data may be reviewed in the context of the numerator and the denominator for the given presentation. The numerator represents the number of events being reported while the denominator represents the population from which the numerator is taken. For example, if it is reported that there are 50 cases of diabetes in

California then the numerator would be the number of cases (50) and the denominator would be the number of people in California that could have diabetes (more than 38 million people since diabetes can occur at any age or sex). While the numerator is relatively straight-forward to identify, the denominator can be difficult. Data summaries are frequently presented in tables in which numerators and denominators may be identified.

The numerator is typically the value in each table cell. However, the denominator can be difficult to identify given the various ways in which tables are prepared. Two examples of tables, Figure 3 and Figure 4, show the numerators and denominators in sample tables.

Figure 3. Illustration of Numerators and Denominators in a Table

Members Using Medi-Cal Long-Term Services and Support



Year	Number of Female Medi-Cal Members	Number of Male Medi-Cal Members
2022	8,625,954	7,654,306
2021	8,140,645	7,158,177

Column Headings
Row Headings
Table Cell

Numerator: Number of Female Medi-Cal Members in 2022 = 8,625,954

Denominator: Total Number of Medi-Cal Members in 2022 = 16,280,260

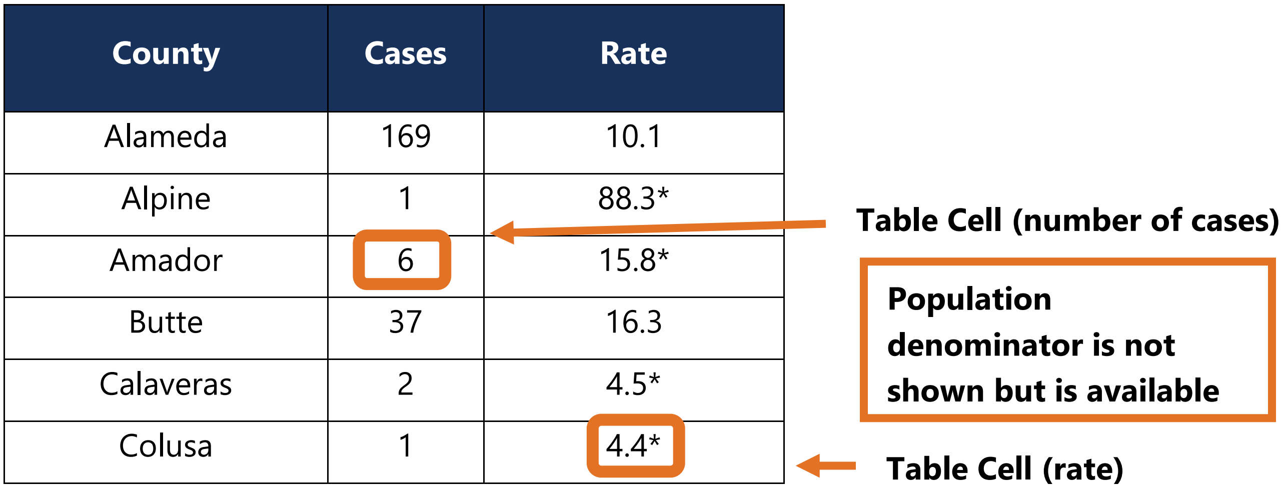
Figure 3 shows an example table with the numerator highlighted. The Cells in the table are the boxes with values in them, as opposed to the row and column headings. The row headings are 2021 and 2022. The column headings are Year, Number of Female Medi-Cal Members, and Number of Male Medi-Cal Members. In Figure 3, "8,625,954" is the value in a table cell and represents a numerator. The sum of the row for year 2022 (8,625,954 + 7,654,306 = 16,280,260) represents a denominator. In this context, the denominator may represent row totals, column totals or the total occurrences in the dataset released. Data in Figure 3 comes from the "Medi-Cal Long-Term Services and Supports Data" dashboard on the CalHHS Open Data Portal.[12]

[12] Data was extracted from the Medi-Cal Long-Term Services and Supports Data files for 2021 and 2022, CalHHS Open Data Portal: https://data.chhs.ca.gov/dataset/long-term-services-and-supports

Figure 4 shows another type of table that contains rates. In this case, the numerator is the number of Salmonellosis cases for a sample of California Local Health Jurisdictions in 2021. The table also includes the rate of Salmonellosis for these jurisdictions. In order to calculate the rate, the population size of each jurisdiction is required, but is not shown directly in this table. The population denominator is an important element for data de-identification.

Figure 4: Illustration of Numerators and Denominators in a Table of Rates

Salmonellosis Cases by Selected County, 2021[13]



County	Cases	Rate
Alameda	169	10.1
Alpine	1	88.3*
Amador	6	15.8*
Butte	37	16.3
Calaveras	2	4.5*
Colusa	1	4.4*

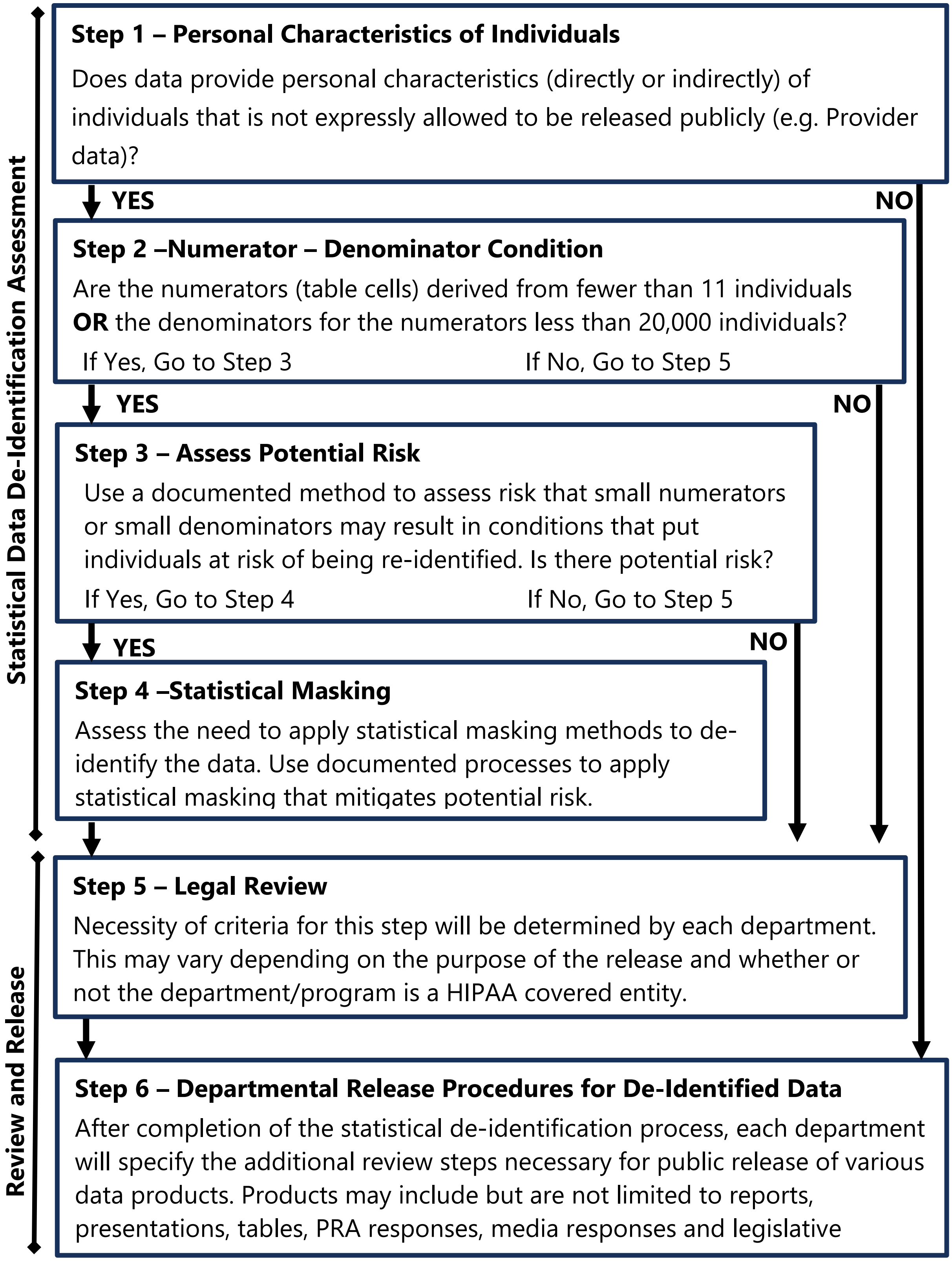
Table Cell (number of cases)
Population denominator is not shown but is available
Table Cell (rate)

Note: Rates are expressed as cases per 100,000 jurisdiction population per year.

*Potentially unreliable rate: relative standard error 23 percent or more.

[13] Data was sourced for the first six counties listed alphabetically within the “Salmonellosis, Cases and Rates by Health Jurisdiction” table in CDPH Infectious Diseases Branch (IDB) Yearly Summaries of Selected Communicable Diseases in California, 2013-2021, https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/YearlySummariesofSelectedCommDiseasesinCA2013-2021.pdf

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 (e.g. Provider data)?

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

4.1 Personal Characteristics of Individuals

As described in Section 3 and Figure 2, personal characteristics of individuals introduce the most significant risk with respect to identifying individuals in a dataset. The following are examples of personal characteristics.

- Identifiers as defined in the California IPA
- Identifiers as defined in HIPAA
- Demographics typically reported in census and other reporting
  - Race
  - Ethnicity
  - Language Spoken
  - Sex
  - Sexual Orientation and Gender Identity
  - Age
  - Socio-economic status as percent of poverty

Personal characteristics are those characteristics that are distinctive to a person and may be used to describe that person. Personal characteristics include a broader set of information than those data elements that may be specifically defined as identifiers (such as, driver license, address, birth date, etc.) Personal characteristics may also be inferred from characteristics related to provider or utilization data. For example, if presented with information about a provider that only sees women, it can be inferred that the clients are women even if that is not specifically stated in the data presentation.

4.2 Numerator – Denominator Condition

The Numerator – Denominator Condition represents a combination of both the Numerator Condition and Denominator Condition and for which both conditions must be met or else a more detailed assessment is required. This may be considered as an initial screening of a dataset.

$$\frac{Numerator - number\ of\ events\ with\ the\ characteristics\ of\ the\ given\ row\ and\ column}{Denominator - the\ population\ from\ which\ the\ events\ arise}$$

The Numerator Condition sets a lower limit for the cell size of cells displayed in a table. The DDG has set this limit as any value representing aggregated or summarized records which are derived from less than 11 individuals (clients). Of note, values of zero (0) are typically shown since a non-event cannot be identified.

The Denominator Condition sets a minimum value for the denominator. The DDG has identified the lower limit for the denominator to be a minimum value of 20,000.

Since this is a Numerator – Denominator Condition, both the minimum cell size for the numerator and denominator must be met. If these conditions are met, the table can move to Step 5 for consideration for release to the public. If either the numerator or denominator condition is not met, then the review of the data must proceed to Step 3.

4.3 Assess Potential Risk

This step requires the use of a documented method to assess the risk that small numerators or small denominators may result in conditions that put individuals at risk of being re-identified.

Assessment of potential risk for a given dataset must take into account a range of contributing considerations. This includes understanding particular characteristics of a given dataset that is being released. For example, if the potential values for a specific personal characteristic, such as race, results in many small numbers in dataset A but does not in dataset B, then the risk may be low for dataset B and high for data A if the groupings of the personal characteristics include the same categories. For this reason, each department or program may set different values for risk based on the underlying distribution of these variables in the datasets of interest.

There are many methods[14] used to assess potential risk. Many of the methods that are in use throughout the country are described in the various references provided in Section 11. While each department will document the method(s) chosen for use, all the CalHHS departments are directed to use the following description of the Publication Scoring Criteria as an example and as a method to assess potential risk.

4.3.1 Publication Scoring Criteria: Example of tool to assess potential risk

The Publication Scoring Criteria is used to identify the presence of small values that are considered sensitive in order to facilitate the assessment of potential risk. The

[14] Garfinkel, et al., “De-Identifying Government Datasets: Techniques and Governance”, National Institute of Standards and Technology Special Publication 800-188, https://doi.org/10.6028/NIST.SP.800-188

Publication Scoring Criteria combines a number of conditions that increase the risk of a given data table and allows the department to evaluate those risks in combination with each other. The variables included in the Publication Scoring Criteria are those variables routinely used to publish data but are not all inclusive. Explanations for the risk values assigned to variables can be found in Section 16, Appendix D. Section 16.2.14 addresses how to account for other variables that are not included in the Publication Scoring Criteria.

A variable is a symbol representing an unknown numerical or categorical value in an equation or table. A given variable may have different ranges assigned to it. Ranges assigned to the variable may be defined in many ways which may increase or decrease the risk of identification of an individual represented in the table. This is seen in the Publication Scoring Criteria in that ranges for variables which will produce smaller groupings have a higher score.

The Publication Scoring Criteria in Figure 6 quantifies with a score two identification risks: size of potential population and variable specificity. The Publication Scoring Criteria is used to assess the need to perform statistical masking as a result of a small numerator, small denominator, or both. The Publication Scoring Criteria takes into account both variables associated with numerators, such as Events, and with denominators, such as Geography or Insurance Coverage.

This method requires a score less than or equal to 12 for the data table to be released without additional masking of the data. Any score over 12 will require the use of statistical masking methods described in Section 4.4 or documentation regarding the specific characteristics of the dataset that mitigate the risk.

When identifying the score for each variable, use the highest scoring criteria. For example, if a table had age groups of 0 to 11 years, 12 to 14 years, and 15 to 18 years then the score for the "age range" variable would be +5 because the smallest age range is 12 to 14, which is an age range of three years.

If a variable has greater granularity than the score listed, use the highest score listed. For example, if the variable "Time" has a frequency of "weekly" then the score would be +5 which is the maximum score associated with the most granular level (monthly) of the variable in the Publication Scoring Criteria.

In addition to assessing the granularity of each variable, the interaction of the variables is also important. As discussed later in Section 4.4, decreasing the granularity or the number of variables are both techniques for increasing the values for the numerators.

The final criteria in Figure 6 are those for Variable Interactions. This provides for subtraction of points if the only variables presented are the events (numerator), time, and geography, and an addition of points for including more variables in a given presentation. With respect to the subtraction of points, the score is based on the minimum value for the Events variable. For example, if the smallest value for the Events is 5 or more, then the score would be -5. However, if the smallest value for the Events is 2, then the score would be 0. This is discussed in more detail in Section 16.2.

In assessing risk, scoring can be part of the justification to release or not release data but should not by itself be an absolute gateway to the release of data. The review must take into account additional considerations including those that are discussed in this document in addition to the scoring.

Figure 6: Publication Scoring Criteria Tables by Variable

Events (Numerator)

Characteristics	Score
1000+ events in a specified population	+2
100-999 events	+3
11-99 events	+5
<11 events	+7

Age Range

Characteristics	Score
>29-year age range	+1
11-29 year age range	+2
6-10 year age range	+3
3-5 year age range	+5
1-2 year age range	+7

Race or Race/Ethnicity

The following two tables can be used for data that complies with current OMB standards (which combines race/ethnicity into one variable) and data that complies with previous

1997-2024 OMB standards (which separated race and ethnicity into two variables) as the risk assessment is the same for both.

Race or Race/Ethnicity Combined

Characteristics	Score
White, Asian, Black or African American, Hispanic or Latino, Middle Eastern or North African	+2
White, Asian, Black or African American, Hispanic or Latino, Middle Eastern or North African, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Mixed	+3

Detailed Race or Race/Ethnicity Combined

Characteristics	Score
Detailed Race or Race/Ethnicity Combined with Population >4,000,000 e.g., Mexican	+1
Detailed Race or Race/Ethnicity Combined with Population 300,001 – 4,000,000 e.g., Chinese, Filipino, German, Asian Indian, Italian, Korean, Salvadoran, Guatemalan	+2
Detailed Race or Race/Ethnicity Combined with Population 100,001 – 300,000 e.g., Japanese, Armenian, Iranian, Aztec, Portuguese, Taiwanese, Hmong, Puerto Rican, Peruvian	+3
Detailed Race or Race/Ethnicity Combined with Population 20,001 – 100,000 e.g., Cambodian, Dutch, Pakistani, Egyptian, Thai, Maya, Afghan, Nigerian, Indonesian, Fijian, Native Hawaiian, Jamaican, Cuban, Colombian, Argentinean	+5
Detailed Race or Race/Ethnicity Combined with Population ≤20,000 e.g., Tongan, Chamorro, Bangladeshi, Sri Lankan, Brazilian, Mixtec, Kenyan, Zapotec, Malaysian, Belizean, Chumash, Sudanese, Pomo, Inca, Pipil	+7

Ethnicity

Use the following two tables to assess risk for the ethnicity variable, which will be present in data that follows pre-2024 OMB standards.

Ethnicity Only

Characteristics	Score
Hispanic or Latino - yes or no	+1

Detailed Ethnicity

Characteristics	Score
Detailed Ethnicity with Population >4,000,000 e.g., Mexican	+1
Detailed Ethnicity with Population 300,001 – 4,000,000 e.g., Salvadoran, Guatemalan, Central American, South American	+2
Detailed Ethnicity with Population 100,001 – 300,000 e.g., Puerto Rican, Spaniard, Peruvian, Nicaraguan, Honduran	+3
Detailed Ethnicity with Population 20,001 – 100,000 e.g., Cuban, Colombian, Argentinean, Dominican, Panamanian	+5
Detailed Ethnicity with Population ≤20,000 e.g., Bolivian, Uruguayan, Paraguayan	+7

Language Spoken

Characteristics	Score
English, Spanish, Other Language	+1
Detailed Language with Population 300,001 – 4,000,000 e.g., Chinese, Tagalog, Vietnamese, Korean	+2
Detailed Language with Population 100,001 - 300,000 e.g., Persian, Hindi, Arabic, Russian, Japanese, French	+3
Detailed Language with Population 20,001 - 100,000 e.g., German, Portuguese, Hmong, Hebrew, Bengali, Polish	+5
Detailed Language with Population ≤20,000 e.g., Haitian, Navajo	+7

Sex, Sexual Orientation, and Gender Identity

Variable	Characteristic	Score
Sex	Male or Female	+1
Sexual Orientation	Straight, Gay or Lesbian, Bisexual, Asexual	+2
Gender Identity	Man/Male, Woman/Female, Transgender or Non-Binary	+3
Gender Identity	Man/Male, Woman/Female, disaggregation of Transgender/Non-Binary category into more specific identities (e.g., Genderqueer, Two-Spirit, etc.)	+5

Intersex

Variable	Characteristic	Score
Intersex (asked as separate question)	Yes or No	+2
Intersex (combined with Sex question)	Male, Female, Intersex	+2

Immigration Status

Characteristic	Score
U.S. Citizen, Foreign Born (combines Naturalized Citizen and Noncitizen)	+1
U.S. Citizen, Naturalized Citizen, Noncitizen	+1
Detailed Immigration Status with Disaggregation of Noncitizen Statuses - Refer to High-Risk Populations (Section 5.6.2)	N/A

Insurance Coverage

Use the following table when reporting by insurance coverage, such as by health plan. See Appendix I for more details on scoring scenarios involving the overlap of Insurance Coverage, Expected Payer/Public Assistance and Means-Tested Programs, and Geography. Below are three key points that summarize all the scenarios:

1) If the data is ONLY related to Residence or Service Geography, then DO NOT USE Insurance Coverage or Means-Tested Tables.
2) Means-Tested Programs—Only add interaction if enrollment in the Public Assistance program is 10 million or fewer people. No interaction is needed for Medi-Cal as the current enrollment is approximately 14 million, which exceeds 10 million.
3) If the number of members enrolled in Insurance Coverage is less than the population of the geographic subdivision, then use the Insurance Table. If the number of members enrolled in Insurance Coverage is greater than or equal

to the population of the geographic subdivision, then use the Geography Table.

Characteristic	Score
Coverage with >2,000,000 members	-5
Coverage with 1,000,001 - 2,000,000 members	-3
Coverage with 560,001 - 1,000,000 members	-1
Coverage with 250,001 - 560,000 members	0
Coverage with 100,001 - 250,000 members	+1
Coverage with 50,001 - 100,000 members	+3
Coverage with 20,001 - 50,000 members	+4
Coverage with ≤ 20,000 members	+5

Expected Payer/ Public Assistance and Means-Tested Programs

Characteristic	Score
Enrollment > 10,000,000 people	+0
Enrollment 4,000,001 – 10,000,000	+1
Enrollment 300,001 – 4,000,000	+2
Enrollment 100,001 – 300,000	+3
Enrollment 20,001 – 100,000	+5
Enrollment ≤20,000	+7

Geography

If the level of reporting is best described by the geography of the individual/service, use one of the following two tables. Specifically, if the geography of the reporting is based on the residence of the individual, use the "Residence Geography" table. If the geography of the reporting is based on the location of service, use the "Service Geography" table.

Residence Geography

Characteristic	Score
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 250,001 - 560,000	0
Population 100,001 - 250,000	+1
Population 50,001 - 100,000	+3
Population 20,001 - 50,000	+4
Population 4,001 - 20,000	+5
Population ≤ 4,000	+7

Service Geography

Characteristic	Score
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 250,001 - 560,000	-1
Population of reporting region 20,001 - 250,000	0
Population of reporting region ≤20,000	+1
Address (Street and ZIP)	+3
Address in rural[15] area	+5
Address in frontier[16, 17] area	+7

Time – Reporting Period

Characteristic	Score
5 years aggregated	-5
2-4 years aggregated	-3
1 year (e.g., 2001)	0
Bi-Annual	+3
Quarterly	+4
Monthly	+5

[15] U.S. Census Bureau, 2020 Census "Urban and Rural" Classifications, https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html

[16] National Rural Health Association, "Definition of Frontier" Policy Paper, February 2016, Retrieved June 2025 from: https://www.ruralhealth.us/getmedia/132306e1-1643-4b40-818a-4d743317dc7e/NRHAFrontierDefPolicyPaperFeb2016.pdf

[17] U.S. Department of Agriculture Economic Research Service, "Frontier and Remote Area Codes", Retrieved June 2, 2025, from https://www.ers.usda.gov/data-products/frontier-and-remote-area-codes

Variable Interactions

Characteristics	Score
Only Events (minimum of 5), Time, and Population (Residence/Service Geography or Insurance Coverage)	-5
Only Events (minimum of 3), Time, and Population (Residence/Service Geo. or Insurance Coverage)	-3
Only Events (no minimum), Time, and Population (Residence/Service Geo. or Insurance Coverage)	0
Events, Time, and Population (Residence/Service Geo. or Insurance Coverage) + 1 variable	+1
Events, Time, and Population (Residence/Service Geo. or Insurance Coverage) + 2 variables	+2
Events, Time, and Population (Residence/Service Geo. or Insurance Coverage) + 3 variables	+4

4.4 Statistical Masking

Statistical masking provides an extensive set of tools that can be used to mitigate potential risk in a given data presentation. If Step 3 of the Data Assessment for Public Release Procedure (Figure 5) determined that the dataset has a risk that small numerators or small denominators may result in conditions that put individuals at risk of being re-identified, then the dataset must be assessed to determine the need for statistical masking of those small values and complementary values. In performing the statistical masking, the data producer must consider what level of analysis may be sacrificed in order to produce a table with lower risk.

4.4.1 Suppressing Small Counts

One common way of masking data is to suppress cells that contain small counts by following these steps:

1) Suppress cells (e.g., count of members and services provided to members) <11 (excluding 0) when total score ≥13.
2) After the cell suppression (<11 excluding 0) is completed, complementary cell suppression is also required so that the suppressed cells cannot be re-identified. See “Complementary Cell Suppression” section on Page 25.

3) Values of 0 should not be suppressed since a non-event cannot be identified.

4) Suppression is also required for financial data which can be associated with members or services provided to members <11 (excluding 0).

5) Suppression is required for all the associated statistical entries (e.g., Prevalence rates, percentages, Mean etc.) of the suppressed cells.

6) An additional complementary cell needs to be suppressed if (a) OR (b) is true:
   a. all of the values suppressed in a specific group (row or column) are each ≤ 3 (1, 2, 3 excluding 0)
   b. the sum of the values suppressed is less than 11

4.4.2 Other Masking Methods

Other masking methods may be applied when one of the following conditions is met:

a) Multiple variables. This most often occurs in a pivot table presentation or a query interface where a user may have occurrences of disease X, stratified by multiple variables, such as age, sex, race, and ethnicity.

b) Granular variables. The more granular the variable the smaller the potential numerator and denominator. This most commonly occurs with shortening the time period of reporting (weekly) or making the geography more specific (zip code or census tract). However, it can also occur when there are many categories for a variable. An example of this is aid codes in Medi-Cal where there are almost 200 aid codes.

c) Rare events. Examples include diseases such as hemophilia. Another example is mass trauma events such as a plane crash or multi-car accident.

In each of these cases, statistical masking may be addressed in a number of ways. For this reason, it is important to keep in mind the purpose of the reporting so that the method chosen for masking can still maximize the usefulness of the data provided. Choices for each case are highlighted below.

a) Multiple variables. Options include separating the table into multiple tables that limit the number of variables included in each table; decreasing the granularity of the variables included in the table; or suppression of counts <11. For example, if there are six variables of interest for study, but a table that cross-tabulates all six variables produces a large number of small cells, the data producer could consider producing several tables with fewer variables so that the risk score is <13. This is especially effective if there are very few analytic questions requiring a cross-tabulation of all six variables.

b) Granular variables. A common approach to this situation would be to decrease the granularity of the variables (although suppressing counts <11 is also an option). This is especially useful for variables with a large number of categories that can be easily aggregated to fewer categories while still maintaining much of their utility. Geographic variables such as state or county can often be recoded into regional categories that still serve the analytic needs of the data user. It is also the only table restructuring option for tables with only two or three variables which have limited opportunities for variable reduction.

c) Rare events: In these cases, it is challenging to suppress the value so that it cannot be used with other public information to identify individuals. Additionally, with rare events, there is more significance in the variance of small numbers. The above-mentioned suppression rules minimize the risk of re-identification most times. However, an expert should treat each data on a case-by-case basis and add additional rules if there is a risk of re-identification in any data. Please see 4.4.3 for a couple of examples in which all the above rules are covered but note that if it is revealed that the cells are suppressed due to regular suppression (<11) and not for complementary suppression then all the suppressed cells can be re-identified.

4.4.3 Complementary Cell Suppression

Complementary cells are those that must be suppressed to prevent someone from calculating the suppressed cell based on row or column totals in combination with other data in that row or column. For example:

Example 1: 10-10-10

Count of Medi-Cal Members by Age

Age	Count
A1	10*
A2	14
A3	10*
A4	10*
A5	0
A6	0
A7	0
A8	30
Total	74

In the above example, if we suppress the counts for cells A1, A3 and A4, marked with an asterisk (*) and each with values of 10, and if we reveal that it is due to regular suppression of cells <11 then anyone can guess that each cell is 10. In this case, either we should not specify that the three cells are <11 or suppress a complementary cell A2 (with a value of 14) so that the three cells marked with an asterisk (*) could not be identified.

Example 2: 10-9

Count of Medi-Cal Members by Age

Age	Count
A1	10*
A2	14
A3	9*
A4	17
A5	0
A6	0
A7	0
A8	30
Total	80

In the above example, if we suppress the counts for A1 and A3, marked with an asterisk (*) and with values of 10 and 9, and if we reveal that it is due to regular suppression of cells <11 then there are only 2 possible combinations (A1=10, A3=9) or (A1=9, A3=10). In this case, either we should not specify that the two cells are <11 or suppress a complementary cell (A2, with a value of 14) so that the cells marked with an asterisk (*) could not be identified.

Example 3. When to suppress 0?

Counts and Percentages of Medi-Cal Members by County

County	Count	Percent
XXX	3	0.0
YYY	15	1.0
ZZZ	0	0.0

In this example, the percentage of 0.0 should not be suppressed for County ZZZ because it is based on a non-event. However, the percentage of 0.0 for County XXX needs to be suppressed because it is due to rounding of numbers. For example, if the denominator for the County XXX percentage is 7,500, a count of 4 would have a rounded percentage of 0.1. Therefore, it could be inferred that the count for County XXX is 1, 2, or 3 because a count of 3 is the highest value that would have a rounded percentage of 0.0 and counts of 0 are not suppressed. Consequently, summary statistics based on suppressed counts should not be reported even if the rounded value is 0 due to the potential for the information to be used for inference of suppressed values.

Example 4. When does indication of complementary suppression lead to data re-identification?

Count of Medi-Cal Members by Age

Age	Count
A1	14
A2	14
A3	1*
A4	11*
A5	0
A6	0
A7	0
A8	30
Total	70

In the above example, if we suppress the counts for A3 and A4 marked with an asterisk (*) and with values of 1 and 11, and if we reveal that A3 is due to regular suppression and A4 is due to complementary suppression then with the given total both the cells can be re-identified. In this case, we should not specify the nature of the suppressed cell so that the cells marked with an asterisk (*) could not be identified.

In these cases, it will be necessary to suppress small cells and perform complementary suppression to ensure that precise values of small cells cannot be calculated using the values of unsuppressed cells and marginal values. In the simplest case, this means ensuring that each column and row of a two-dimensional table has at least two suppressions. This ensures that the precise values of the suppressed cells cannot be calculated. Complementary suppressions are often selected using one of the methods listed below.

1) The ‘analytically least interesting’ level of a particular dimension. This is often ‘other’, or ‘I don’t know’.
2) The smallest cell available for complementary suppression. This is based on minimizing the ‘information loss’.

The cell most similar to the cell needing complementary suppression, such as adjacent age groups. This can produce complementary suppression that may be easier to interpret. It is important to clearly designate which cells have been suppressed due to complementary suppression. Use a symbol to indicate the cell has been suppressed. Identify any other cells (complementary cells) that can be used to calculate the small cell and use a symbol to indicate the cell has been suppressed. Please see below two ways to indicate cell suppression.

1) Suppression symbols for Machine Readable Version: The Open Data Portal requires submission of a machine-readable format. Therefore, the CalHHS Open Data Portal guidelines[18] provides instructions on the table structure.

   CalHHS Open Data Portal Small Cell Suppression Guidelines

- Guidelines: Use an annotation field (column) in each data table that corresponds to records that have suppressed cells.
- Small Cell Data Standard:
  - Value in cell is blank if blank due to “annotation”
  - “0” in cell if value is 0
  - Annotation field table:

Annotation Field	Definition
0 or blank	No annotation or blank
1	Cell suppressed for small numbers
2	Cell suppressed for complementary cell
3	No data is available
4	Statistically unstable value
5	Incomplete data

[18] CalHHS Open Data Handbook “Publication Guidelines for CalHHS Departments and Offices”, https://kb.data.chhs.ca.gov/odp/guidelines

    - Data Dictionary / Metadata indicates small cell method used (<11, etc.)

- Considerations:
    - Use metadata and documentation to inform users.
    - Consider highlighting and drawing attention to annotated fields.

2) Suppression Symbols for Non-Machine-Readable Version: Departments may want to present data in a non-machine-readable format for usability of the data. In this case, use the symbols "S," "*," or similar symbols for counts less than 11. Use symbols "C", "**", or similar symbols for complementary cells. When suppressing values, it is recommended to use the following footnote to indicate the suppression.

- Values or cells marked as "S", "*", or a similar symbol in the data would have the following footnote or note:
    - "S" represents Counts that are less than 11 which are not shown in accordance with the CalHHS DDG Edition 2.0.
    - Values or cells marked with "C", "**", or similar symbols indicate complementary cells would have the following footnote or note:
    - "C" represents counts for complementary data that are not shown in accordance with the CalHHS DDG Edition 2.0.

4.4.4 Masking a Three Category Variable (e.g. Intersex Status)

A special situation may occur when complementary cell suppression must be applied to a variable with three categories. This occurs when "Intersex" is an option for the Sex variable (Male, Female, Intersex) but can occur in any situation where a variable has two large categories and one small category (for example, if the options for Gender Identity included Man, Woman, and Nonbinary).

As an example:

Male	Female	Intersex	Total
545	545	10	1100

If the DDG risk assessment concludes that counts less than 11 must be suppressed, then complementary suppression must be applied to either the Male or Female categories so that the Intersex count of 10 cannot be back-calculated. The problem is that suppression of either category would deprive the public of vital information and hence be publicly unacceptable. Other masking methods (such as combining Intersex with either the Male or Female categories; or reassigning the Intersex individuals to both categories equally) will likely also be unacceptable as they would be viewed as erasure of the Intersex identity.

In these instances, the recommendation is to not show counts at all, but only percentages rounded to the nearest whole percent. Such a "rounding error" will effectively mask the smallest category without requiring complementary suppression.

Here are the raw percentages for the above example:

Male	Female	Intersex	Total
49.545%	49.545%	0.009%	100%

Here are the rounded percentages:

Male	Female	Intersex	Total
50%	50%	<1%	100%

Once the percentages are rounded, one cannot back-calculate to obtain the Intersex count, even if one knows the total is 1100. One would only know that the count is less than 11, which is the same knowledge as with standard suppression practices. It is recommended that a footnote is provided to explain why this was done, for example:

> "To protect individual privacy, counts are not provided for this table and percentages have been rounded to the nearest whole number. Due to rounding, totals for all categories may not add up to 100%."

A few other points regarding this method:

- Ensure that counts for the variable are not provided anywhere else in the document and are not available in the public domain. In the above example, one needs to make sure one cannot obtain a total count of 545 males and females elsewhere (for example, in a table that cross-tabulates sex by race).
- This method is only effective if the total number of individuals is 1100 or above. Below this threshold, one can back-calculate counts of less than 11 even when percentages are rounded.

- This method is only needed for a three-category variable. In the above example, if Intersex were presented as a separate "yes/no" variable and no cross-tabulation with Sex was provided, then complementary suppression would not be required, and this issue would not arise.

4.4.5 Balancing privacy and equity goals

Data de-identification, which ensures individuals are not made identifiable by public facing data, and the use of masking to preserve privacy and confidentiality, exist alongside the need to produce high-quality analyses focused on equity and disproportionality. Underrepresented or marginalized communities are more likely to have their data masked in data de-identification efforts due to smaller cell sizes, which can result in a lack of information that can be safely made publicly available about those populations. Section 4.4.2 outlines suggestions on how to preserve information while also preserving confidentiality. Intuitive ways to preserve information while also preserving confidentiality include presenting data as percentages, aggregating smaller groups into reportable-sized ones, or multi-year reporting that relates marginalized or underreported groups to an average. In addition, the DDG Peer Review Team remains committed to evaluating ways to continue to balance these goals through continued training and other innovative methodologies on an ongoing basis, including those listed in Appendix E.

4.5 Expert Determination Documentation and Legal Review

Expert Determination Review:

Review for Expert Determination will be performed by individuals who have been qualified as experts by the Office of Legal Services (OLS) and who meet the HIPAA Privacy Rule implementation specifications: "A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" [45 CFR Section 164.514(b)(1)].

Relevant expertise may be gained through various routes of education and experience. OLS will review the relevant professional experience and academic or other training of the expert, as well as the actual experience of the expert using health information De-identification methodologies.

The expert determination review, according to the regulation's requirements, will apply [i] "the generally accepted statistical and scientific principles and methods, in order to determine 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] document the methods and results of the analysis that justify such determination…" DHCS intends that these Guidelines provide the starting point for expert determination review; however, the facts of each case chosen for expert determination review must be analyzed on an individual, case-by-case basis by the expert. The documentation must include a general description of the principles, methods, and analyses used, as well as an explanation of the analysis that justifies the expert determination. If followed, the Guidelines may be referenced as part of the documentation used to support the expert determination.

The expert determination review will use the Expert Determination Template in Section 14.2 of Appendix B. The Expert Determination Template includes a confirmation 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. If methods that have been used to de-identify the data are not described in the Guidelines, then the Expert will need to provide additional documentation that explains the statistical and scientific principles and methods used, as well as the results of any additional analysis.

OLS Review:

This review will assess the data to be released for risk to the Department, and for potential implications on litigation, statutory or regulatory conditions on data release, and other legal considerations that may impact release. OLS review also includes reviewing the expert determination and expert determination documentation to ensure compliance with the HIPAA Privacy Rule.

4.6 Departmental Release Procedure for De-identified Data

The final step in preparing aggregate data for publishing is to submit the aggregate data analysis to the document review process. The document review process for documents that include data analysis includes the organizational entities listed below. The review by these entities is conducted in the order below, with the final review performed by the Office of Communications (OC). If any entity requests significant changes, then the document will be returned to Program Management Review and will undergo the process again. The documentation associated with the requested changes is to remain with the review documents.

- Program Management Review
- OC Review

Program Management Review:

The program management team is expected to review the table for both accuracy and to assess the publication in accordance with the DHCS Health Administrative Manual (HAM) Chapter 3 – Communications.

OC Review:

OC is to receive all publications, brochures, or pamphlets intended for public distribution to be printed or reproduced to review the material to determine if it requires Governor's Office approval (HAM Section 3-6040). OC will ensure that all reviews described in Step 5 have occurred prior to publication of the document(s).

For documents that will be posted to the public website, the OPA web unit reviews all content. This review will assess the data table for compliance with the Americans with Disabilities Act of 1990 (ADA). Prior to publication of data tables, OC will have the Data Analytics Division (DAD) provide a final quality assurance review of the web content to ensure that the items approved through the Expert Determination Review are those being released.

5) TYPES OF REPORTING

CalHHS programs develop a wide range of information based on different types of data. This is reflected in the various categories shown on the entry page for the CalHHS Open Data Portal, which include:

Diseases and Conditions

- Facilities and Services
- Health Care
- Workforce
- Environment
- Demographics
- Resources

Various types of reporting may or may not have a connection to personal characteristics that would create potential risk of identifying individuals.

5.1 Variables

The following two lists of variables are important to consider when preparing data for release.

1. Personal characteristics:
   - Age
   - Sex
   - Sexual Orientation and Gender Identity
   - Intersex
   - Race
   - Ethnicity
   - Language Spoken
   - Immigration Status
   - Location of Residence
   - Education Status
   - Financial Status
2. Event characteristics:
   - Number of events
   - Location of event
   - Time period of event
   - Provider of event

As stated previously, variables that are personal characteristics may be used to determine a person's identity or attributes. When these characteristics are used to confirm the identity of an individual in a publicly released dataset, then a disclosure of an individual's information has occurred. Individual uniqueness in the released data and in the population is a quality that helps distinguish one person from another and is directly related to re-identification of individuals in aggregate data. Disclosure risk is a concern when released data reveal characteristics that are unique in both the released data and in the underlying population. The risk of re-identifying an individual or group of individuals increases when unique or rare characteristics are "highly visible", or otherwise available without any special or privileged knowledge. Unique or rare personal characteristics (e.g., height above seven feet) or information that isolate individuals to small demographic subgroups (e.g., American Indian Tribal membership) increase the likelihood that someone can correctly attribute information in the released data to an individual or group of individuals.

Variables that are event characteristics are often associated with publicly available information.

Therefore, increased risk occurs when personal characteristics are combined with enough granularity with event characteristics. One could argue that if no more than two personal characteristics are combined with event characteristics then the risk will be low independent of the granularity of the variables. This hypothesis will need to be tested using various population frequencies to quantify the uniqueness of the combination of variables both in the potential data to be released as well as in the underlying population.

5.2 Survey Data

Survey data, often collected for research purposes, are collected differently than administrative data, and these differences should be considered in decisions about security, confidentiality, and data release.

Administrative data sources (non-survey data) such as vital statistics (e.g., births and deaths), health care administrative data (e.g., Medi-Cal utilization; hospital discharges), reportable disease surveillance data (e.g., measles cases) contain data for all persons in the population with the specific characteristic or other data elements of interest. Most of the discussions in this document pertain to these types of data.

On the other hand, surveys (e.g., the California Health Interview Survey) are designed to take a sample of the population and collect data on characteristics of persons in the sample, with the intent of generalizing to gain knowledge suggestive of the whole population.

The sampling methodology developed for any given survey is generally developed to maximize the sample size with the available resources while making the sample as un-biased (representative) as possible. These sampling procedures that are a fundamental part of surveys generally change the key considerations for the protection of security and confidentiality and give an extra layer of masking which can decrease the risk of re-identification. Given the world of artificial intelligence, increased availability of data on social platforms, increased incidences of data hacking, and the passage of new laws on data sharing, there is a much higher risk to re-identify an individual when small numbers are released. Thus, the Publication Scoring Criteria should be used to assess the potential risk and the same Data De-Identification Guidelines should be applied to the survey data.

Also, it is in the context of surveys that issues of statistical reliability often arise—which are distinct from confidentiality issues, but often arise in related discussions. We recommend that the estimates be considered statistically unreliable and suppressed if the denominator is based on fewer than 30 sample cases.

A few national and state surveys have adopted various guidelines to de-identify their survey data and various thresholds for the reliability of sample sizes.[19] If any state or national survey has not adopted its own guidelines to de-identify data and/or threshold of statistical reliability and if the survey is not included in the Appendix H list, then the above-mentioned CalHHS Data De-Identification Guidelines and threshold of 30 sample cases should be observed by the CalHHS departments.

5.3 Budgets and Fiscal Estimates

Budget reporting may include both actual and projected amounts. Projected amounts, although developed with models that are based on the historical actuals, reflect activities that have not yet occurred and, therefore, do not require an assessment for de-identification. Actual amounts do need to be assessed for de-identification. When the budgets reflect caseloads, but do not include personal characteristics of the individuals in the caseloads or other variables in the scoring criteria, then the budgets

[19] Klein RJ, et al. Healthy People 2010 Criteria for Data Suppression. Centers for Disease Prevention and Control, Statistical Notes. https://www.cdc.gov/nchs/data/statnt/statnt24.pdf

are reflecting data in the Providers and Health and Service Utilization Data circles of the Figure 2 Venn Diagram and do not need further assessment.

However, the budget reporting needs to be assessed for de-identification if the data:

- Reflects anything that can be tied back to member counts, or services provided to members
- Includes personal characteristics, such as age, sex, race or ethnicity, or other variables in the scoring criteria.

5.4 Facilities, Service Locations and Providers

Many CalHHS programs oversee, license, accredit or certify various businesses, providers, facilities, and service locations. As such, the programs report on various metrics, including characteristics of the entity and the services provided by the entity.

1) Characteristics of the entity are typically public information, such as location, type of service provided, type of license and the license status.
2) Services provided by the entity will typically need to be assessed to see if the reporting includes personal characteristics about the individuals receiving the services. Several examples are shown below.
    a. Reporting number of cases of mental illness treated by each facility – if the facility is a general acute care facility, then the reporting of the number of cases does not tell you about the individuals receiving the services.
    b. Reporting number of cases of mental illness treated by each facility – if the facility is a children's hospital, then the reporting of the number of cases does tell you about the individuals receiving the services.
    c. Reporting number of psychotropic medications prescribed by a general psychiatrist does not tell you about the patients receiving the medications.
    d. Reporting number of psychotropic medications prescribed by a general psychiatrist to include the number of medications prescribed by the age group, sex or race/ethnicity of the patients receiving the medications does tell you about the patients receiving the medications.

In (a) and (c) above, assessment for de-identification is not necessary as there are no characteristics about the individuals receiving the services. However, in (b) and (d) above, the inclusion of personal characteristics which may be quasi-identifiers, especially when combined with the geographical information about the provider, does require an assessment for de-identification.

5.5 Mandated Reporting

CalHHS programs are required to provide public reporting based on federal and California statute and regulations, court orders, and stipulated judgments, as well as by various funders. Although reporting may be mandated, unless the law expressly requires reporting of personal characteristics, publicly reported data must still be de-identified to protect against the release of identifying or personal information which may violate federal or state law.

5.6 Special Scenarios

5.6.1 Data with More Specificity

Specific diagnosis and procedure codes have higher risk of identification because of their unique nature. Thus, these diagnosis and procedure codes should be scored based on “Other Variables” in Section 16.2.14.

Some examples of data with more specificity include:

- Specific ICD codes for surgical procedures: Bariatric surgery (Z98. 84) [Examples of specific bariatric procedures: Open Roux-en-Y (0D160ZB), Laparoscopic Roux-en-Y (0DB64Z3), Laparoscopic Sleeve Gastrectomy (0DB64Z3), Biliopancreatic Diversion with Duodenal Switch (BPD/DS) (0DB60Z3+0D160ZB+0D194ZB), Adjustable Gastric Lap Band (Z46. 51), Single Anastomosis Duodeno-ileal Bypass with Sleeve Gastrectomy(0D16079)][20, 21, 22]
- Specific diagnoses: Disability [Complete Paraplegia G82.21], birth defects [e.g., Cleft Palate (Q35. 9)], neurological disorders [e.g., Dyslexia and alexia (R48.0), Cerebral Palsy (G80. 9)], and End stage renal disease (N18.6).
- Specific procedure codes: Transcranial doppler ultrasound (CPT code 93886 or 93971), Sickle cell screening (Z 13.0 or D57. 1).

[20] Mayo Clinic. Bariatric Surgery: Overview, https://www.mayoclinic.org/tests-procedures/bariatric-surgery/about/pac-20394258

[21] American Society for Metabolic and Bariatric Surgery: Public education committee. Bariatric Surgery Procedures, https://asmbs.org/patients/bariatric-surgery-procedures

[22] https://www.cms.gov/medicare/coding-billing/icd-10-codes

5.6.2 High-Risk Populations which require suppression of counts <11 regardless of score

911 calls and court records are considered public information. Therefore, a large amount of information is publicly available about individuals who interact with emergency services or the justice system, including identifying information such as name, gender, date of birth, age, and, when applicable, the location of where the patient is currently housed or is committed to. Examples of populations that may interact with emergency services or the justice system at a higher rate than other populations include:

- People with substance use disorders (SUD), serious mental illnesses (SMI), psychological or emotional trauma, or developmental disabilities;
- People who are involved in the child welfare system, experiencing housing instability or homelessness, victims of crime and abuse, or justice-involved;
- People without U.S. citizenship.

Therefore, regardless of the underlying population, suppression of counts less than 11 is always required for these populations. Note that most of the demographic variables are assigned risk scores that are consistent with those of Table 2 in Section 16.2. However, the above-mentioned high-risk populations have been exempted from this requirement due to the higher potential visibility of these individuals in public information (e.g., from police reports or news reports) that inherently increases risk of re-identification that is a risk that is not otherwise accounted for with the scoring criteria.

Some examples of the above include:

- SUD: medication assisted treatment services, prescription opioid misuse, other illicit drugs use
- SMI: schizophrenia, schizoaffective disorders, bipolar disorder, unspecified psychosis not due to a substance or known physiological condition, paraphilias, personality disorders, major depressive disorders, delusional disorders, stimulant related disorders, post-traumatic stress disorder
- Claims for emergency transport

Acknowledging the possibility of the use of reproductive health care data for law enforcement purposes, California law provides additional privacy protections for individuals receiving health care services related to pregnancy termination and for individuals receiving gender-affirming care.[23] Given the recognized risk associated with these services, suppression of counts less than 11 is always required regardless of the underlying population.

Some examples of the above include:

- Reproductive Health: surgical termination of pregnancy, medication/pharmaceutical termination of pregnancy
- Gender-Affirming Care: sex reassignment surgery, hormone replacement therapy

5.6.3 Re-Identification Risk with Increased Granularity

Sometimes, we may be asked to provide more granular data. However, increasing the number of demographic attributes or providing additional variables to achieve this[24, 25, 26] increases the risk that the unique combination of variables can effectively re-identify an individual.

If the expert determines that the provided information could be used, alone or in combination with other available information to identify or re-identify an individual, then the data should not be released. Instead, an alternative method that protects the privacy of individuals should be used to provide the requester with the data they need.

[23] AB 352 (Bauer-Kahan, Chapter 255, Statutes of 2023)

[24] Sweeney, "Uniqueness of Simple Demographics in the U.S. Population", LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.

[25] Golle, "Revisiting the Uniqueness of Simple Demographics in the US Population", Accessed from stanford.edu: https://crypto.stanford.edu/~pgolle/papers/census.pdf

[26] Rocher, et al. "Estimating the success of re-identifications in incomplete datasets using generative models", Nat Commun 10, 3069 (2019). https://doi.org/10.1038/s41467-019-10933-3

6) APPROVAL PROCESSES

Recognizing that some data analyses may be published as independent tables while other analyses will be part of larger reports, the final review of all data analyses must follow the department or office procedures for document review in addition to review procedures identified for the implementation of the DDG. The expectation is that the review of data for de-identification will fit into other routine review processes. Reviews outside the DDG portion may vary depending on whether data is being released for a PRA request, to the media, to the legislature, by the program as part of routine reporting, or for other reasons.

6.1 Statistical Review to Assess De-Identification (Steps 1, 2, 3 & 4)

The Chief Data Officer (CDO) will provide statistical review of data products before they are released to ensure the data has been de-identified with methods that are consistent with these guidelines. These individuals are considered experts for the purpose of performing expert determinations in compliance with the HIPAA Privacy Rule, and who meet the Rule's implementation specifications: "A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" [45 CFR Section 164.514(b)(1). This expert determination review, according to the regulation's requirements, will be performed by:

> "(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"[27]

When an expert determination review is requested, the Expert Determination Review must include a document that includes the expert's determination that "the risk is very

[27] 45 CFR section 164.514 (b)

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," attests that the requirements of 45 CFR section 164.514 (b)(1)(i) and (ii) have been met, and includes (or attaches) the documentation required by 45 CFR section 164.514(b)(1)(ii). This document must be signed by the expert.

These guidelines provide a starting point for expert determination review; however, the facts of each case chosen for expert determination review must be analyzed on an individual, case-by-case basis by the expert. If followed, the Guidelines may be referenced as part of the documentation used to support the expert determination. The documentation should also include a general description of the principles, methods, and analyses used, as well as an explanation of the analysis that justifies the expert determination.

The expert determination review may use the Expert Determination Template in Appendix B. The Expert Determination Template includes a confirmation 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."

If methods that have been used to de-identify the data are not described in the Guidelines, then the Expert will need to provide additional documentation that explains the statistical and scientific principles and methods used and the results of the additional analysis.

6.2 Legal Review (Step 5)

Step 5 in the Data Assessment for Public Release Process provides for a legal review within the department. This review will assess the data to be released for risk to the Department, and for potential implications on litigation, statutory or regulatory conditions on data release, and other legal considerations that may impact release. Legal Services will review the expert determination documentation to ensure compliance with the HIPAA Privacy Rule as applicable.

6.3 Departmental Release Procedures (Step 6)

Step 6 in the Data Assessment for Public Release Process provides for departmental release procedures for de-identified data. Products may include but are not limited to reports, presentation, tables, PRA responses, media responses and legislative responses.

OC is designated to receive all publications, brochures, or pamphlets intended for public distribution to be printed or reproduced, and review the material to determine if it requires Agency Approval or Governor's Office approval. OC has also been designated to review content to assess the data table for compliance with the Americans with Disabilities Act of 1990 (ADA).[28]

Quality Assurance Reviews: This is particularly pertinent to data products being added to the web sites to ensure that they have had appropriate reviews and De-identification steps. It also applies to reviews of updated reports. Many reports maintain the same variables and formats but have updated numbers/information on a periodic basis (monthly, quarterly, annually). For these reports, departments may consider a centralized review to ensure data products are consistent with previously reviewed reports and have not had changes that would change the previous assessment.

[28] 42 U.S.C 12101 et seq.

7) DDG GOVERNANCE

Governance for DDG will be provided by the Data Subcommittee with support from the Risk Management Subcommittee. The Subcommittees are part of the CalHHS governance structure as described in the CalHHS Information Strategic Plan.[29]

Governance for the CalHHS DDG will provide the following support for departments and offices.

- Maintain the CalHHS DDG, which will include updates and revisions to the document as well as annual reviews for currency.
- Coordinate integration of the CalHHS DDG into the Statewide Health Information Policy Manual (SHIPM), Section 2.5.0 De-Identification[30] and the CalHHS Open Data Handbook.
- Convene a Peer Review Team (PRT).
- Provide for escalation of issues that cannot be resolved by the PRT.

The CalHHS PRT will include no more than two representatives from each department or office. Membership of the PRT is expected to include individuals with the following background and experience.

- Knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable.
- Knowledge of and experience with legal principles associated with data de-identification in compliance with California IPA and HIPAA.

The PRT will have the following responsibilities:

- Provide review and consultation regarding a department's DDG to ensure it is consistent with the CalHHS DDG. This may be particularly useful if a department incorporates methods for de-identification in the department's DDG that have not already been documented in the CalHHS DDG.
- Provide for escalation and review of data de-identification questions or issues that a department is not comfortable resolving independently.

[29] California Health and Human Services Agency, Information Strategic Plan 2016.

[30] Statewide Health Information Policy Manual, https://www.cdii.ca.gov/compliance-and-policy/statewide-health-information-policy-manual-shipm/

- Develop training tools to be used by departments when developing and implementing department-specific DDGs based on the content of the CalHHS DDG.
- Review and approve each department's DDG procedure and any alternate forms of de-identification before implementation.
- Review and approve individuals designated by departments as Statistical De-Identification Experts and Statistical De-Identification Supervisor Experts. For departments without staff to meet the minimum qualifications of the Statistical De-Identification Supervisor Expert role, the PRT shall review and approve all individuals prior to a department entering into an agreement for the individual to complete the activities of the Statistical De-Identification Supervisor Expert.

The PRT will not review all disclosures or data released by each department but shall be available for consultation and second opinions.

8) DEVELOPMENT PROCESS

The DHCS developed the PAR-DBR Guidelines in 2014 through a collaborative departmental process. The development process included a literature review, case examples and broad discussion among DHCS programs, in addition to consultation with other CalHHS departments. The PAR-DBR V1.6 was finalized on August 26, 2014. In 2016, the PAR-DBR was updated to align with the CalHHS DDG and became identified as DHCS DDG 2.0. In 2021, DHCS updated the DHCS DDG to version 2.1 to reflect the title change of the Chief Medical Information Officer to Chief Data Officer. In 2022, DHCS added non-substantive updates to version 2.2, which were approved by the CalHHS Peer Review Team. Additionally, Appendix B was updated to align with the current version posted on the CDII's website.

This 2025 update, Version 3.0, introduces improvements made in the CalHHS DDG Edition 2.0. These include consistent scoring criteria based on California Population thresholds, updated scores for detailed race and ethnicity according to OMB guidelines and Assembly Bill 91, detailed language spoken, additional scores for residence and service geography tables considering small populations, and new scores for Insurance Coverage, Sexual Orientation and Gender Identity, Intersex, Immigration Status, and Expected Payer/Public Assistance and Means-Tested Programs. Additionally, some special scenarios, such as data with higher specificity and high-risk populations, were added. The update also covers new methodologies for survey data, symbols, and guidelines for data suppression, as well as more granular data. CalHHS DDG Edition 2.0. was approved on October 24, 2025, by CalHHS. DHCS DDG version 3.0 received approval from the CalHHS Peer Review Team on November 13, 2025.

9) ABBREVIATIONS AND ACRONYMS

Abbreviation	Acronyms
CDC	Centers for Disease Control and Prevention
CDPH	California Department of Public Health
CDSS	Department of Social Services
CalHHS	California Health and Human Services Agency
CMS	Centers for Medicare and Medicaid Services
CPHS	Committee for the Protection of Human Subjects
DDG	Data De-Identification Guidelines
DHCS	Department of Health Care Services
HCAI	Department of Health Care Access and Information
HIPAA	Health Insurance Portability and Accountability Act
IPA	Information Practices Act
MHSOAC	Mental Health Services Oversight and Accountability Commission
PHI	Protected Health Information
PI	Personal Information
PRA	Public Records Act
PRT	Peer Review Team

10) DEFINITIONS

Aggregate – formed or calculated by the combination of many separate units or items (Oxford Dictionary).

De-identified – generally defined under the HIPAA Privacy Rule (45 CFR section 164.514) as information (1) that does not identify the individual and (2) for which there is no reasonable basis to believe the individual can be identified from it.

Denominator – the portion of the overall population being referenced in a table or a figure representing the total population in terms of which statistical values are expressed (Oxford Dictionary).

Numerator – the number of specific cases as identified by the variable from a given population or the number above the line in a common fraction showing how many of the parts indicated by the denominator are taken (Oxford Dictionary).

Protected Health Information – information which relates to the individual's past, present, or future physical or mental health or condition, the provision of health care to the individual, or the past, present, or future payment for the provision of health care to the individual, and that identifies the individual, or for which there is a reasonable basis to believe can be used to identify the individual (HIPAA, 45 CFR section 160.103).

Personal Information – includes information that is maintained by an agency which identifies or describes an individual, including his or her name, social security number, physical description, home address, home telephone number, education, financial matters, email address and medical or employment history. It includes statements made by, or attributed to, the individual (California Civil Code section 1798.3).

Publishable State Data – Data is Publishable State Data if it meets one of the following criteria: (1) data that are public by law such as via the PRA or (2) the data are not prohibited from being released by any laws, regulations, policies, rules, rights, court order, or any other restriction. Data shall not be released if it is highly restricted due to the Health Insurance Portability and Accountability Act (HIPAA), state or federal law (such data are defined as Level 3 earlier in this handbook).[31]

Re-Identified – matching de-identified, or anonymized, personal information back to the individual.

[31] CalHHS Open Data Portal Guidelines, https://kb.data.chhs.ca.gov/odp/guidelines

11) REFERENCES

Abowd and Hawes, "21st Century Statistical Disclosure Limitation: Motivations and Challenges". Available through: United States Census Bureau Resource Library, Working Paper Number ced-wp-2023-002, https://www.census.gov/library/working-papers/2023/adrm/ced-wp-2023-002.html

Armstrong, MP, G Rusthon, and DL Zimmerman, 1999, Geographically Masking Health Data to Preserve Confidentiality. Statistics in Medicine, 18, 497-525.

Bambauer, Jane R., Tragedy of the Data Commons (March 18, 2011). Harvard Journal of Law and Technology, Vol. 25, 2011. Available at SSRN: http://ssrn.com/abstract=1789749 or http://dx.doi.org/10.2139/ssrn.1789749

Benitez K1, Malin B., Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010 Mar-Apr;17(2):169-77. doi: 10.1136/jamia.2009.000026. http://www.ncbi.nlm.nih.gov/pubmed/20190059

CalHHS Open Data Handbook. "Publication Guidelines for CalHHS Departments and Offices", https://kb.data.chhs.ca.gov/odp/guidelines

California Department of Public Health, Infectious Diseases Branch (IDB) Yearly Summaries of Selected Communicable Diseases in California, 2013-2021, https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/YearlySummariesofSelectedCommDiseasesinCA2013-2021.pdf

California Department of Social Services, CalFresh Dashboard, https://public.tableau.com/app/profile/california.department.of.social.services/viz/CFdashboard-PUBLIC/AnnualParticipation. 2025. Accessed March 25, 2025

California Department of Social Services. CalWORKs Take up. CalWORKs Interactive Summary. https://www.cdss.ca.gov/inforesources/calworks-summary/program-overview/take-up#C018. 2025. Accessed March 25, 2025

California Department of Social Services. Housing Support Program. CalWORKs Interactive Summary. https://www.cdss.ca.gov/inforesources/calworks-summary/program-services-utilization/housing-and-homelessness#C004. 2025. Accessed March 25, 2025

California Employment Development Department, California Open Data Portal datasets, https://data.ca.gov/organization/california-employment-development-department

California Health and Human Services Agency, Information Strategic Plan 2016.

California Health Interview Survey, 2021 population estimates for SOGI, AskCHIS,database, https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/access-chis-data

California Statewide Health Information Policy Manual (SHIPM), Section 2.5.0 De-Identification, Available through: Center for Data Insights and Innovation, https://www.cdii.ca.gov/compliance-and-policy/statewide-health-information-policy-manual-shipm/

Center for Data Insights and Innovation, Office of the Health Information Integrity, "2022 HIPAA Assessment Results", Health Information Entity Status Assessment: https://www.cdii.ca.gov/compliance-and-policy/health-information-entity-status-assessment/

Centers for Disease Control and Prevention, National Center for Health Statistics, https://www.cdc.gov/nchs/index.htm

Centers for Disease Control and Prevention, National Environmental Public Health Tracking Network, https://ephtracking.cdc.gov/

Centers for Disease Control and Prevention, WONDER database, https://wonder.cdc.gov/

Center for Migration Studies, "Estimates of Undocumented and Eligible-to-Naturalize Populations by State", http://data.cmsny.org/state.html

Centers for Medicare & Medicaid Services, Office of Information Products and Data Analytics. "Medicare Fee-For Service Provider Utilization & Payment Data Physician and Other Supplier Public Use File: A Methodological Overview." April 7, 2014.

Colorado Department of Public Health and Environment. "Guidelines for Working with Small Numbers." Retrieved April 2024 from http://www.cohid.dphe.state.co.us/smnumguidelines.html

Committee for the Protection of Human Subjects (CPHS), "About CPHS: Legal Authority and Scope" web content, https://www.cdii.ca.gov/committees-and-advisory-groups/committee-for-the-protection-of-human-subjects-cphs/.

Department of Finance, Demographic Research Unit, Various demographic reports and population estimates, https://dof.ca.gov/forecasting/demographics/

Department of Health Care Access and Information, Health Care Payments Data Snapshot, “Data Overview – Count of Individuals by Payer Type”, https://hcai.ca.gov/visualizations/healthcare-payments-data-hpd-snapshot/

Department of Health Care Access and Information (HCAI), Patient Origin & Market Share Reports

Department of Health Care Access and Information, Report to the Legislature “Program Report: Health Care Payments Data Program”, May 2024, https://hcai.ca.gov/wp-content/uploads/2024/04/HPD-Report-to-the-Legislature-March-2024-1.pdf

Department of Health Care Services, “Medi-Cal Long-Term Services and Supports Data” dashboard, CalHHS Open Data Portal, https://data.chhs.ca.gov/dataset/long-term-services-and-supports

Federal Committee on Statistical Methodology, Interagency Confidentiality and Data Access Group. “Checklist on Disclosure Potential of Proposed Data Releases.” Washington: Statistical Policy Office, Office of Management and Budget, July 1999.

Federal Committee on Statistical Methodology, “Statistical Policy Working Paper 22 – Report on Statistical Disclosure Limitation Methodology.” Washington: Statistical Policy Office, Office of Management and Budget, 1994.

Garfinkel S, Guttman B, Near J, Dajani AN, Singer P (2023) De-Identifying Government Datasets: Techniques and Governance. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) NIST SP 800-188. https://doi.org/10.6028/NIST.SP.800-188

Golle, Philippe. “Revisiting the uniqueness of simple demographics in the US population. In Proceedings of the 5th ACM Workshop on Privacy in the Electronic Society. ACM Press, New York, NY. 2006: 77-80.

Green, Ben, Gabe Cunningham, Ariel Ekblaw, Paul Kominers, Andrew Linzer, and Susan Crawford. 2017. Open Data Privacy (2017). Berkman Klein Center for Internet & Society Research Publication. Available at: http://nrs.harvard.edu/urn-3:HUL.InstRepos:30340010

Howe, H. L., A. J. Lake, and T. Shen. "Method to Assess Identifiability in Electronic Data Files." American Journal of Epidemiology 165.5 (2006): 597-601. Print.

Illinois Department of Public Health, Center for Health Statistics

Klein RJ, Proctor, Boudreault, and Turczyn. “Healthy People 2010 Criteria for Data Suppression”. Centers for Disease Prevention and Control, Statistical Notes. https://www.cdc.gov/nchs/data/statnt/statnt24.pdf

Maine Integrated Youth Health Survey, https://www.maine.gov/miyhs/

NAHDO-CDC Cooperative Agreement Project CDC Assessment Initiative. “Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk.” December 2004. Retrieved May 29, 2025, from https://www.nahdo.org/sites/default/files/publications/Data_Release_Guidelines.pdf

National Rural Health Association, “Definition of Frontier” Policy Brief, February 2016, Retrieved June 2025 from: https://www.ruralhealth.us/getmedia/132306e1-1643-4b40-818a-4d743317dc7e/NRHAFrontierDefPolicyPaperFeb2016.pdf

NCHS Staff Manual on Confidentiality. Hyattsville, MD: National Center for Health Statistics, Department of Health and Human Services, “NCHS Staff Manual on Confidentiality.” 2004. Retrieved from http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf.

NORC, “Case Study: The Disclosure Risk Implications of Small Cells Combined with Multiple Tables or External Data,” January 8, 2016.

NORC, “NORC Recommendations for California Department of Health Care Services (DHCS) Data De-Identification Guidelines (DDG),” January 8, 2016.

North American Association of Central Cancer Registries (NAACCR), “Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices,” October 2002.

Office of Civil Rights, U.S. Department of Health & Human Services. "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." November 26, 2012. Retrieved May 29, 2025, from https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

Office of Statewide Health Planning and Development (now HCAI), “Medical Service Study Areas – Census Detail, 2010”; Retrieved from the California State Geoportal on April 2, 2024

Ohio Department of Public Health. "Data Methodology for Public Health Practice." http://www.odh.ohio.gov/~/media/ODH/ASSETS/Files/data%20statistics/standards/methodological%20standards/disclimit.ashx.

Panel on Disclosure Review Boards of Federal Agencies: Characteristics, Defining Qualities and Generalizability, 2000, Proceedings of the Joint Statistical Meetings, Indianapolis, Indiana.

Parker JD, Talih M, Malec DJ, et al. National Center for Health Statistics. "Data Presentation Standards for Proportions". National Center for Health Statistics. Vital Health Stat 2(175). 2017. Retrieved from https://www.cdc.gov/nchs/data/series/sr_02/sr02_175.pdf

Privacy Technical Assistance Center, U.S. Department of Education. "Data De-Identification: An Overview of Basic Terms." May 2013. Retrieved June 2, 2025, from https://studentprivacy.ed.gov/sites/default/files/resource_document/file/data_deidentification_terms_0.pdfTerms

State of California, Department of Finance, Report P-1 (Race): State and County Population Projections by Race/Ethnicity, 2010-2060. Sacramento, California, January 2013. Retrieved from http://www.dhcs.ca.gov/services/MH/InfoNotices-Ltrs/Documents/InfoNotice-PrimaryLang-Enclosure1.pdf

Stoto, MA. Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems. RAND Health. September 19, 2002.

Sweeney, L. "Information Explosion, Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies," L Zayatz, P Doyle, J Theeuwes and J Lane (eds), Urban Institute, Washington, DC, 2001.

Sweeney, L. "K-anonymity: a model for protecting privacy." International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems. 2002; 10(5): 557-570.

Sweeney, L. Testimony before that National Center for Vital and Health Statistics Workgroup for Secondary Uses of Health information. August 23, 2007.

Templ et al., "Introduction to Statistical Disclosure Control", International Household Survey Network, Working Paper 007, 2014

U.S. Bureau of Labor Statistics, "Occupation Finder: Occupational Outlook Handbook", https://www.bls.gov/ooh/occupation-finder.htm

U.S. Census Bureau, Geography Program, “About Geographic Areas: Urban and Rural”, 2020 Census, Retrieved from: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html on April 2, 2024

U.S. Census Bureau, “About Program Income and Public Assistance,” (2025), Retrieved from: https://www.census.gov/topics/income-poverty/public-assistance/about.html on March 25, 2025

U.S. Census Bureau, American Community Survey, 5-Year 2017-2021

U.S. Census Bureau, American Community Survey 5-Year Estimates 2017-2021, “2021 Educational Attainment, California Estimates, Population 25 Years and Over”, Table S1501: https://data.census.gov/table/ACSST5Y2021.S1501?q=S1501&g=040XX00US06&y=2021

U.S. Census Bureau, "HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE," 2020. Decennial Census, DEC Redistricting Data (PL 94-171), Table P2, 2020. Accessed on October 6, 2023.

U.S. Census Bureau, "RACE," 2020. Decennial Census, DEC Redistricting Data (PL 94-171), Table P1, 2020. https://data.census.gov/table/DECENNIALPL2020.P1?g=040XX00US06,06$0500000. Accessed on October 6, 2023.

U.S. Census Bureau, " SELECTED CHARACTERISTICS OF THE NATIVE AND FOREIGN-BORN POPULATIONS," 2022. American Community Survey, ACS 5-Year Estimates Subject Tables, Table S0501, 2022, https://data.census.gov/table/ACSST5Y2022.S0501?q=S0501. Accessed on December 11, 2023.

U.S. Census Bureau, “Selected Characteristics of the Uninsured in the United States”, American Community Survey, ACS 5-Year Estimates Subject Tables, Table S2702, 2022, https://data.census.gov/table/ACSST5Y2022.S2702?g=040XX00US06. Accessed October 10, 2024.

U.S. Census Bureau, "Selected Social Characteristics in the United States," 2021. American Community Survey, ACS 5-Year Estimates Data Profiles, Table DP02, 2021, https://data.census.gov/table/ACSDP5Y2021.DP02?g=040XX00US06. accessed on December 5, 2023.

U.S. Census Bureau, "SEX BY SINGLE-YEAR AGE," 2020. Decennial Census, DEC Demographic and Housing Characteristics, Table PCT12, 2020, https://data.census.gov/table/DECENNIALDHC2020.PCT12?g=040XX00US06. Accessed on October 6, 2023.

U.S. Core Data for Interoperability, Taxonomy, https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi

U.S. Department of Commerce, National Institute of Standards and Technology, "De-Identifying Government Datasets: Techniques and Governance", NIST Special Publication 800-188, Published 2023, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-188.pdf

U.S. Department of Education, Office of Federal Student Aid, "What federal benefits (means-tested benefits) might my family and I be eligible for?", https://studentaid.gov/help/means-tested-benefits. Accessed on March 2025.

U.S. Department of Homeland Security. (2022, March). Fiscal Year 2020 Refugees and Asylees Annual Flow Report. Dhs.gov. https://www.dhs.gov/sites/default/files/2022-03/22_0308_plcy_refugees_and_asylees_fy2020_1.pdf.

U.S. Department of Homeland Security. (2021, June). Population Estimates of Nonimmigrants Residing in the United States: Fiscal Years 2017-2019. dhs.gov. https://www.dhs.gov/sites/default/files/publications/immigration-statistics/Pop_Estimate/NI/ni_population_estimates_fiscal_years_2017_-_2019v2.pdf

U.S. Department of Homeland Security. (2022, September). September 2022 Population estimates of the Lawful Permanent Resident Population in the United States and the Subpopulation Eligible to Naturalize: 2022. dhs.gov. Retrieved June 2, 2025, from https://ohss.dhs.gov/sites/default/files/2023-12/2022_0920_plcy_lawful_permenent_resident_population_estimate_2022_0.pdf

U.S. Department of Veterans Affairs, "VetPop2020 State Estimates 2000 to 2020," updated on April 6, 2023, https://www.data.va.gov/dataset/VetPop2020-State-Estimates-2000-to-2020/fkjq-z6m8

U.S. Office of Management and Budget. "Initial Proposals for Updating OMB's Race and Ethnicity Statistical Standards", https://www.federalregister.gov/d/2023-01635.

U.S. Office of Management and Budget, “Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity”, https://www.federalregister.gov/d/97-28653

U.S. Office of Management and Budget, “Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity”, Federal Register: https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and

Wartman SA, Combs CD. Reimagining Medical Education in the Age of AI. AMA J Ethics. 2019 Feb 1;21(2):E146-152. doi: 10.1001/amajethics.2019.146. PMID: 30794124.

Washington State Department of Health. "Guidelines for Working with Small Numbers." N.p., 15 October 2012 (Revised May 2018). Retrieved on June 2. 2025, from https://doh.wa.gov/sites/default/files/legacy/Documents/1500/SmallNumbers.pdf

Note: Some references that appear above are no longer directly cited within the DDG but have been kept on this list due to their influence on the development of the original guidelines. Later DDG Editions replaced some older references with newer, more up-to-date resources as those became available.

12) APPROVALS

We have reviewed the document, “DHCS Department Data De-Identification Guidelines (DDG),” and hereby approve it as our official position for Department data release.

Original signed by ______________________ Date ______________________

Linette Scott, MD, MPH, Chief Data Officer and Deputy Director, Enterprise Data and Information Management, DHCS

Original signed by ______________________ Date ______________________

Judith Recchio, Chief Counsel and Deputy Director, Office of Legal Services, DHCS

Original signed by ______________________ Date ______________________

Anne Carvalho, Chief, Data Analytics Division, DHCS

APPENDIX

13) APPENDIX A: LEGAL FRAMEWORK

The overarching legal framework for the CalHHS Data De-Identification Guidelines (DDG) is the California Information Practices Act, California Civil Code 1798 et seq., which was established in 1977 and applies to all state government entities. The IPA includes requirements for the collection, maintenance, and dissemination of any information that identifies or describes an individual. The IPA and other California statutes limit the disclosure of personal information, consistent with the California Constitutional right to privacy. However, state agencies are generally permitted (and sometimes required under the California Public Records Act and other laws) to disclose data that have been de-identified. Summarized or aggregated data may still be identifiable; the DDG provides Guidelines for assessing whether data have been de-identified.

While most state agencies are covered by the IPA, some are also covered by or impacted by HIPAA. Unlike the IPA, which applies to all personal information, HIPAA only applies to certain health or health care-related information. HIPAA requirements apply in combination with IPA requirements. While the IPA does not include specific de-identification methods or criteria, the basic concept of statistical de-identification has no different meaning, and the basic standard of protection of identifiable data is no different for IPA covered PI than for HIPAA covered PHI.

“Personal Information” is defined by the California Civil Code section 1798.3(a) as “any information that is maintained by an agency that identifies or describes an individual, including, but not limited to:

- his or her name,
- social security number,
- physical description,
- home address,
- home telephone number,
- education,
- financial matters, and
- medical or employment history.

It includes statements made by, or attributed to, the individual.”

Under Section 1798.24 of the IPA, "An agency shall not disclose any personal information in a manner that would link the information disclosed to the individual to whom it pertains," unless it is disclosed as described in Section 1798.24.

Senate Bill 13 updated the IPA, effective January 1, 2006, to require Committee for the Protection of Human Subjects (CPHS) review and approval before personal information (linkable to any individual) that is held by any state agency or department can be released for research purposes. CPHS does not delegate reviews for compliance with the IPA to other institutional review boards. (https://www.cdii.ca.gov/committees-and-advisory-groups/committee-for-the-protection-of-human-subjects-cphs/)

13.1 California Laws Governing the Collection and Release of Confidential, Personal, or Sensitive Information

(Please note that this is not an exhaustive list.)

General State Collected Information and Data

- Civ. Code 1798.24, 1798.24a, 1798.24b (all personal information including health data)
- Gov. Code 11015.5 (electronically collected personal information)

General Medical Data

- Civ. Code 56.10 – 56.11
- Civ. Code 56.13
- Civ. Code 56.29
- Health & Saf. Code 128730
- Health & Saf. Code 128735
- Health & Saf. Code 128736
- Health & Saf. Code 128737
- Health & Saf. Code 128745
- Health & Saf. Code 128766

Birth Defects

- Health & Saf. Code 103850

Blood Lead Analysis

- Health & Saf. Code 124130

Cancer

- Health & Saf. Code 103875
- Health & Saf. Code 103885
- Health & Saf. Code 104315

Child Health Information

- Health & Saf. Code 130140.1

Child Health Screening

- Health & Saf. Code 124110
- Health & Saf. Code 124991

Cholinesterase Testing

- Health & Saf. Code 105206

Developmentally Disabled

- Health & Saf. Code 416.18
- Health & Saf. Code 416.8
- Welf. & Inst. Code 4514, 4514.3, 4514.5
- Welf. & Inst. Code 4517 (aggregation and publication of data)
- Welf. & Inst. Code 4659.22
- Welf. & Inst. Code 4744

Environmental Health Hazards

- Health & Saf. Code 59016

General Public Health Records

- Health & Saf. Code 100330
- Health & Saf. Code 121035

Genetic Information

- Health & Saf. Code 124975
- Health & Saf. Code 124980

- Health & Saf. Code 125105 (prenatal test)
- Civ. Code 56.17

HIV/AIDS

- Health & Saf. Code 120820
- Health & Saf. Code 120962
- Health & Saf. Code 120970
- Health & Saf. Code 120972
- Health & Saf. Code 120975
- Health & Saf. Code 120980
- Health & Saf. Code 121010
- Health & Saf. Code 121022
- Health & Saf. Code 121023
- Health & Saf. Code 121025
- Health & Saf. Code 121075
- Health & Saf. Code 121080
- Health & Saf. Code 121085
- Health & Saf. Code 121090
- Health & Saf. Code 121095
- Health & Saf. Code 121110
- Health & Saf. Code 121120
- Health & Saf. Code 121125
- Health & Saf. Code 121280
- Rev. & T. Code 19548.2

Immunizations

- Health & Saf. Code 120440

Independent Medical Review

- Health & Saf. Code 1374.33

Involuntary Mental Health (LPS covered records)

- Welf. & Inst. Code 4135
- Welf. & Inst. Code 5328 through 5328.9
- Welf. & Inst. Code 5329 (aggregation and publication of data)
- Welf. & Inst. Code 5540
- Welf. & Inst. Code 5610
- Educ. C. 56863

Medi-Cal Data

- Welf. & Inst. Code 14015.8
- Welf. & Inst. Code 14100.2
- Welf. & Inst. Code 14101.5

Neurological

- Health & Saf. Code 103871

Parkinson’s Disease Registry

- Health & Saf. Code 103865

Payment and Billing Info

- Health & Saf. Code 440.40 (applies only to GACHs)

Prenatal Tests

- Health & Saf. Code 120705
- Health & Saf. Code 125105

Public Assistance

- Welf. & Inst. Code 10850 (Confidential Information)

Public Social Services

- Welf. & Inst. Code 10850

Substance Abuse Treatment Data

- Health & Saf. Code 11845.5
- Health & Saf. Code 11812

Vital Records

- Health & Saf. Code 102425
- Health & Saf. Code 102426
- Health & Saf. Code 102430
- Health & Saf. Code 102455
- Health & Saf. Code 102460
- Health & Saf. Code 102465
- Health & Saf. Code 102475
- Health & Saf. Code 103025

13.2 Federal Laws Governing Public Data Release

(Please note this is not an exhaustive list.)

- HIPAA - Section 164.514 of the HIPAA Privacy Rule (45 CFR)
- 42 CFR Part 2
- Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99)
- Freedom of Information Act (FOIA) (5 U.S.C. § 552)

14) APPENDIX B: THE HIPAA DE-IDENTIFICATION STANDARD

The Center for Data Insights and Innovation (CDII) is authorized by state statute to coordinate and monitor HIPAA compliance by all California State entities within the executive branch of government covered or impacted by HIPAA. One difference between the California IPA and HIPAA is the documentation requirement in HIPAA for data de-identified using the Expert Determination method.

The HIPAA Standard[32] for de-identification of protected health information (PHI)[33] states "Health information that does not identify an 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 health information." If the data are de-identified, and it is not reasonably likely that the data could be re-identified, the Privacy Rule no longer restricts the use or disclosure of the de-identified data.

The following is quoted from the "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule", published November, 2012 by the U.S. Department of Health & Human Services, Office for Civil Rights: (https://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html)

14.1 The HIPAA De-Identification Standard

Section 164.514(a) of the HIPAA Privacy Rule (45 CFR) provides the standard for de-identification of protected health information. Under this standard, health information is not individually identifiable if it does not identify an individual and if the covered entity has no reasonable basis to believe it can be used to identify an individual.

§ 164.514 Other requirements relating to uses and disclosures of protected health information.

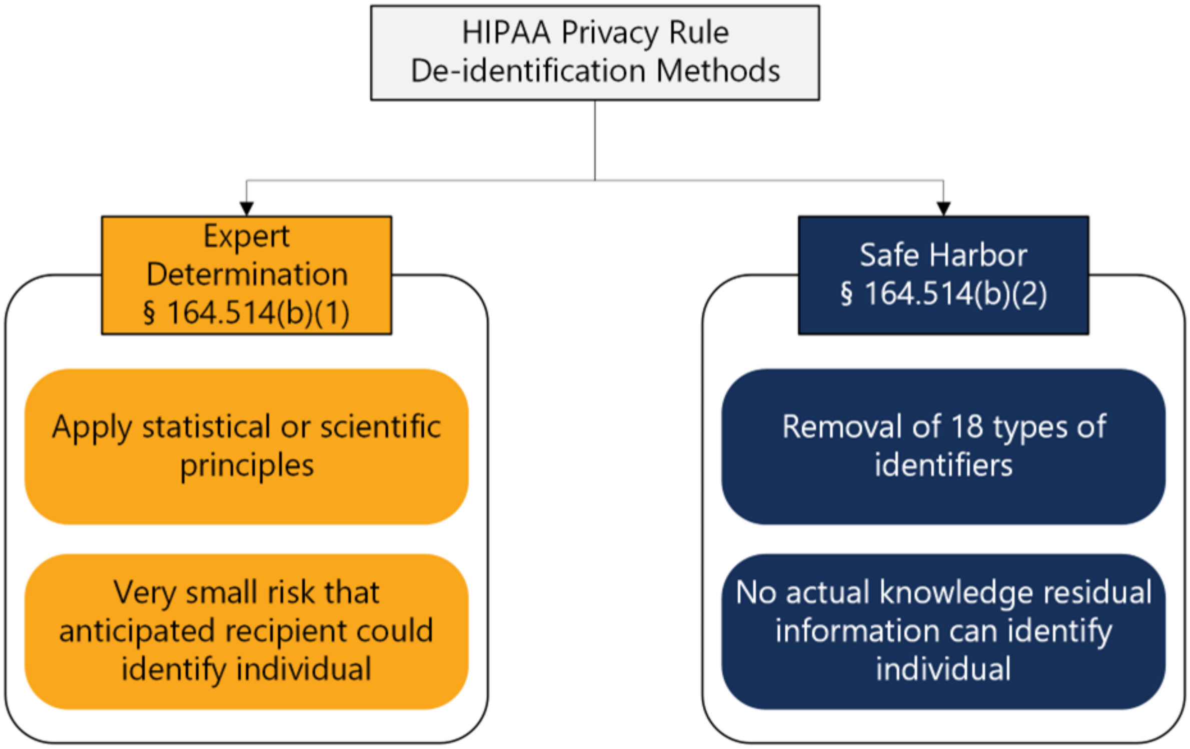
[32] The Standard is found in the HIPAA Privacy Rule, 45 CFR section 164.514(a).

[33] "PHI" is defined as information which relates to the individual's past, present, or future physical or mental health or condition, the provision of health care to the individual, or the past, present, or future payment for the provision of health care to the individual, and that identifies the individual, or for which there is a reasonable basis to believe can be used to identify the individual. (45 CFR section 160.103)

(a) Standard: de-identification of protected health information. Health information that does not identify an 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 health information.

Sections 164.514(b) and(c) of the Privacy Rule contain the implementation specifications that a covered entity must follow to meet the de-identification standard. As summarized in Figure 7, the Privacy Rule provides two methods by which health information can be designated as de-identified.

Figure 7. Two methods to achieve de-identification in accordance with the HIPAA Privacy Rule.


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

The first is the "Expert Determination" method:

> § 164.514(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; or

The second is the "Safe Harbor" method:

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names

(B) 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

(C) 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

(D) Telephone numbers

(E) Fax numbers

(F) Email addresses

(G) Social security numbers

(H) Medical record numbers

(I) Health plan beneficiary numbers

(J) Account numbers

(K) Certificate/license numbers

> (L) Vehicle identifiers and serial numbers, including license plate numbers
>
> (M) Device identifiers and serial numbers
>
> (N) Web Universal Resource Locators (URLs)
>
> (O) Internet Protocol (IP) addresses
>
> (P) Biometric identifiers, including finger and voice prints
>
> (Q) Full-face photographs and any comparable images
>
> (R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"]; and
>
> (ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Satisfying either method would demonstrate that a covered entity has met the standard in §164.514(a) above. De-identified health information created following these methods is no longer protected by the Privacy Rule because it does not fall within the definition of PHI. Of course, de-identification leads to information loss which may limit the usefulness of the resulting health information in certain circumstances. As described in the forthcoming sections, covered entities may wish to select de-identification strategies that minimize such loss.

Re-identification

The implementation specifications further provide direction with respect to re-identification, specifically the assignment of a unique code to the set of de-identified health information to permit re-identification by the covered entity.

> § 164.514 (c) Implementation specifications: re-identification. A covered entity may assign a code or other means of record identification to allow information de-identified under this section to be re-identified by the covered entity, provided that:
>
> (1) Derivation. The code or other means of record identification is not derived from or related to information about the individual and is not otherwise capable of being translated so as to identify the individual; and

(2) Security. 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.

If a covered entity or business associate successfully undertook an effort to identify the subject of de-identified information it maintained, the health information now related to a specific individual would again be protected by the Privacy Rule, as it would meet the definition of PHI. Disclosure of a code or other means of record identification designed to enable coded or otherwise de-identified information to be re-identified is also considered a disclosure of PHI.

14.2 Expert Determination Template

HIPAA covered entities in CalHHS must de-identify data in compliance with the HIPAA standard. Under the HIPAA standard, either Safe Harbor or Expert Determination must be used. If Expert Determination is used then the documentation of the review is essential. The following may serve as a template for this documentation with the reference to the CalHHS DDG to support the analysis documented.

Documentation of Expert Determination Template

Name of Report:

Reason for Data Release:

Identify why the data release does not meet Safe Harbor. For example:

The request does not meet the Safe Harbor standard because it includes counts by county (geographic area smaller than the state) or counts by month (which does not meet the criteria for dates). Therefore, the steps in the CalHHS DDG are being used to assess the tables.

Document how the conditions of each step are met or not met	**Result**
Step 1 – Presence of Personal Characteristics *Summary:*	
Step 2 – Numerator Denominator Condition *Summary:*	
Step 3 – Assess Potential Risk *Summary:*	
Step 4 – Statistical Masking *Summary:*	
Step 5 – Expert Review *Summary:* *"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"*	

15) APPENDIX C: JUNE 2022 HIPAA REASSESSMENT RESULTS

The Center for Data Insights and Innovation (CDII) is authorized by state statute to coordinate and monitor Health Insurance Portability and Accountability Act (HIPAA) compliance by all California State entities within the executive branch of government covered or impacted by HIPAA. To help ensure full compliance with HIPAA, CDII conducted a reassessment with all State Departments in February and March 2022.[34] Note the asterisk (*) means the entity was added to the list per the 2022 review.

Covered Entities and Business Associates - Subject to CDII oversight (including compliance reviews)

List Number	Organization
1	Aging, Department of (CDA)
2	Cal State East Bay*
3	Cal State Fullerton*
4	Cal State Long Beach*
5	Cal State Los Angeles*
6	Cal State Northridge*
7	Cal State San Bernardino*
8	California Maritime Academy*
9	Chico State*
10	Corrections and Rehabilitation (CDCR), Department of / California Correctional Health Care Services (CCHCS)
11	CSU San Marcos*
12	Developmental Services, Department of (DDS)

[34] https://www.cdii.ca.gov/wp-content/uploads/2023/01/2022-Entity-Assessment-Results_2022-06-02.pdf

List Number	Organization
13	Forestry and Fire Protection (Cal Fire), Department of
14	Fresno State*
15	General Services, Department of (DGS)
16	Health Care Services, Department of (DHCS)
17	Prison Industry Authority, California (Cal PIA)
18	Public Employees Retirement System (CalPERS), California
19	Public Health, Department of (CDPH)
20	Sacramento State*
21	San Diego State*
22	San Francisco State*
23	San Jose State*
24	Social Services, Department of (CDSS)
25	Sonoma State*
26	State Hospitals, Department of (DSH)
27	Systems Integration, Office of (OSI)
28	Technology, Department of (CDT)
29	Veterans Affairs, Department of (CalVET)

Impacted Entities – Impacted by State and/or Federal regulations and laws related to health information privacy (non-HIPAA).

Please reference California Health and Safety Code 1302023 (et seq.) CDII's revised statutory authority includes oversight of organizations subject to non-HIPAA health information privacy laws and regulations. CDII is developing a compliance program for non-HIPAA entities to address these other state and federal privacy laws not included in the current HIPAA compliance program. The organizations included in this list will be subject to that newly created program.

List Number	Organization
1	Alcoholic Beverage Control Appeals Board
2	Alcoholic Beverage Control, Department of
3	Business, Consumer Services and Housing Agency
4	Cal Poly Pomona
5	Cal Poly San Luis Obispo
6	California Energy Commission
7	California Institute for Regenerative Medicine
8	Cannabis Control Appeals Panel
9	Cemetery and Funeral Bureau
10	Central Valley Flood Protection Board
11	Child Support Services, Department of
12	Commission on Peace Officer Standards and Training
13	Community Services and Development, Department of
14	Conservation Corps, California
15	Conservation, Department of
16	Contractors' State License Board
17	CSU Bakersfield
18	CSU Channel Islands
19	CSU Dominquez Hills
20	CSU Monterey Bay
21	Delta Stewardship Council
22	Dental Board of California
23	Dental Hygiene Board of CA
24	Department of Cannabis Control

List Number	Organization
25	Department of Financial Protection and Innovation
26	Department of Health Care Access and Information (formally OSHPD)
27	Developmental Disabilities, State Council on
28	Education, Western Interstate Committee for Higher (WICHE)
29	Emergency Medical Services Authority
30	Employment Development Department
31	Employment Training Panel
32	Environmental Protection Agency
33	Equalization, Board of
34	Fair Employment and Housing Department
35	Fair Political Practices Commission
36	Fi$Cal (Financial Information System for California)
37	Fish and Wildlife, Department of
38	Franchise Tax Board
39	Gambling Control Commission
40	Governor's Office of Business and Economic Development
41	Health Benefit Exchange, Covered California - California
42	High-Speed Rail Authority
43	Horse Racing Board, California
44	Housing Finance Agency
45	Human Resources, CalHR - Department of
46	Humboldt Poly Technic
47	Industrial Relations, Department of
48	Inspector General, Office of

List Number	Organization
49	Insurance, Department of (Insurance Commissioner)
50	Justice, Department of (Attorney General)
51	Lottery, California State
52	Mandates, Commission on State
53	Medical Board of California
54	Motor Vehicles, Department of
55	Natural Resources Agency
56	New Motor Vehicle Board
57	Office of Law Enforcement Support
58	Office of Tax Appeals
59	Personnel Board, State
60	Physical Therapy Board of California
61	Planning and Research, Governor's Office of
62	Podiatric Medicine, Board of
63	Prison Industry Authority, California
64	Public Defender, State
65	Public Employment Relations Board
66	Public Utilities Commission, California
67	Real Estate, Department of
68	Registered Nursing, Board of
69	Resources Recycling and Recovery, Department of (CalRecycle)
70	Security and Investigative Services, Bureau of
71	Sierra Nevada Conservancy
71	Stanislaus State

List Number	Organization
73	State Coastal Conservancy
74	State Controller’s Office
75	State Lands Commission
76	State Water Resources Control Board
77	Tax and Fee Administration, Department of
78	Teacher Credentialing, Commission on
79	Teachers' Retirement Board
80	Toxic Substances Control, Department of
81	Transportation, Department of
82	Unemployment Insurance Appeals Board
83	Victim Compensation Board
84	Water Resources, Department of
85	Workforce Development Board

16) APPENDIX D: JUSTIFICATION OF THRESHOLDS IDENTIFIED

16.1 Establishing Minimum Numerator and Denominator

The DDG workgroup reviewed the published literature including information from other states and from the federal government. There was a great deal of variation in the numerical values chosen for the Numerator Condition. While the Centers for Disease Control and Prevention (CDC) WONDER database suppresses cells with numerators less than 10, the National Environmental Public Health Tracking Network suppresses cells that are greater than 0 but less than 6. Examples range from 3 to 40 with many being 10 to 15. The Centers for Medicare and Medicaid Services (CMS) uses a small cell policy of suppressing values derived from fewer than 11 individuals. As stated in a 2014 publication associated with a data release of Medicare Provider Data, "to protect the privacy of Medicare beneficiaries, any aggregated records which are derived from 10 or fewer beneficiaries are excluded from the Physician and Other Supplier PUF [public use file]."[35] Of note, CMS only uses a Numerator Condition.

Just as there is no consistent value for the Numerator Condition, neither is there a consistent value for the Denominator Condition. Some examples include:

- National Center for Health Statistics (public micro-data) – 250,000
- National Environmental Health Tracking Network – 100,000
- Maine Integrated Youth Health Survey – 5,000

In establishing a minimum denominator to protect confidentiality, the DDG workgroup began by looking at the risk associated with providing geography associated with record level data. As noted in the "Guidance Regarding Methods for De-Identification of Protected HIPAA Privacy Rule", published November 2012 by the U.S. Department of Health & Human Services, Office for Civil Rights there is varying risk based on the level of zip code and how the zip code is combined with other variables. It has been estimated that the combination of a patient's Date of Birth, Sex, and 5-Digit ZIP Code is

[35] "Medicare Fee-For Service Provider Utilization & Payment Data Physician and Other Supplier Public Use File: A Methodological Overview," Prepared by: The Centers for Medicare and Medicaid Services, Office of Information Products and Data Analytics, April 7, 2014.

unique for over 50% of residents in the United States.[36, 37] This means that over half of U.S. residents could be uniquely described just with these three data elements. In contrast, it has been estimated that the combination of Year of Birth, Sex, and 3-Digit ZIP Code is unique for approximately 0.04% of residents in the United States.[38] For this reason, the HIPAA Safe Harbor rule specifies that the 3-Digit ZIP Code can be provided at the record level if the 3-Digit ZIP Code has a minimum of 20,000 people. By aggregating data for a given 3-Digit ZIP Code, the potential for identifying a unique individual is less than 0.04%. By combining with the Numerator Condition, the risk becomes less than 0.04% because there will be a minimum of 11 individuals with a particular age and sex for the 3-Digit ZIP Code. Additionally, most tables will provide additional levels of aggregation further reducing risk. This reduction of risk is discussed further with respect to the Publication Scoring Criteria.

A minimum denominator of 20,000 was chosen as part of the numerator-denominator condition to leverage the risk assessment cited above.

The Numerator-Denominator Condition serves as an initial screening to assess potential risk for a dataset. If this condition is met, additional analysis is not necessary. If the condition is not met, then the analysis proceeds to Step 3.

16.2 Assessing Potential Risk – Publication Scoring Criteria

The Publication Scoring Criteria is provided as an example of a method that meets the requirements of Step 3 in the Data Assessment for Public Release Procedure. It is a tool to assess and quantify potential risk for re-identification of de-identified data based on three identification risks: size of potential population, variable specificity, and the impact of re-identification. The Publication Scoring Criteria is used to assess the need to suppress small cells less than 11 as a result of a small numerator (less than 11), small denominator (less than 20,001), or both small numerator and small denominator. That is why the Publication Scoring Criteria takes into account both numerator (e.g., Events) and denominator (e.g., Geography or Insurance Coverage variables).

[36] See P. Golle. Revisiting the uniqueness of simple demographics in the US population. In Proceedings of the 5th ACM Workshop on Privacy in the Electronic Society. ACM Press, New York, NY. 2006: 77-80.

[37] See L. Sweeney. K-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems. 2002; 10(5): 557-570.

[38] See L. Sweeney. Testimony before that National Center for Vital and Health Statistics Workgroup for Secondary Uses of Health information. August 23, 2007.

The Publication Scoring Criteria is based on a framework that was used by the Illinois Department of Public Health, Illinois Center for Health Statistics when CalHHS Data De-Identification Guidelines Edition 1.0 was first prepared in 2016. Various other methods have been used to assess risk and the presence of sensitive or small cells. Public health has a long history of public provision of data and many methods have been used. Further discussion of other methods used to assess tables for sensitive or small cells is found in Section 4.4.

This section provides a more detailed review of the criteria that make up the Publication Scoring Criteria.

16.2.1 Events

Table 1: Events Scoring

Characteristics	Score
1000+ events in a specified population	+2
100-999 events	+3
11-99 events	+5
<11 events	+7

The Events score represents a score for the numerator. The Events category will be scored based on the smallest cell size in the table.

The lowest value for the Events variable (<11 events) which has the highest score (+7) was chosen to be consistent with the Numerator Condition. The Publication Scoring Criteria is used when the Numerator-Denominator Condition is not met. Therefore, when the Numerator Condition is not met with respect to the Events variable, a high score is given.

16.2.2 Generalized Scoring for Personal Characteristics

Scores for all personal characteristics (e.g., age, race/ethnicity, language spoken) have been determined based on thresholds derived from the statewide population sizes as below and the CalHHS DDG Edition 1.0[39] age range scores.

Table 2: Generalized Scoring Criteria Based on Statewide Population

Population Size	Score
Population >4,000,000	+1
Population 300,001 – 4,000,000	+2
Population 100,001 – 300,000	+3
Population 20,001 – 100,000	+5
Population ≤20,000	+7

16.2.3 Age Range

Table 3: Age Range Scoring

Characteristics	Score
>29-year age range	+1
11-29 year age range	+2
6-10 year age range	+3
3-5 year age range	+5
1-2 year age range	+7

The last four categories of Age Range scores in the above table are prepared based on the 2020 Census population counts and Table 2.

General fertility rates and sexually transmitted infection rates are often reported for the female population of childbearing age using the 30-year age range for 15-44 years old and the approximate population size for the female 30-year age range is four million.

[39] California Health and Human Services, Data De-Identification Guidelines (DDG), Edition 1.0, September 23, 2016, https://chhsa.dsh.ca.gov/wp-content/uploads/2021/04/CHHS-Data_Deidentification_Guidelines-V1.0-092316.pdf

Therefore, to incorporate these populations, an additional category with a score of +1 for age ranges greater than 29 years has been included which also reflects the lower identification risk for very large age ranges.

On the other side, age ranges receive a higher score for smaller ranges of years and smaller corresponding population sizes due to an increased risk for identification.

The smallest statewide populations for contiguous ranges of ages under 100 years, as reported for the 2020 Census in table below, were used to establish these thresholds. Note that the smallest population is always associated with the oldest age range and are given a score of seven (7) due to increased variable specificity. These are used to establish standard scoring criteria.

Table 4: Lowest Populations for Age Ranges

Age Range	Lowest Population Age Range	Lowest Population[40]
1-year	99 years	6,708
2-year	98-99 years	16,163
3-year	97-99 years	28,974
5-year	95-99 years	69,636
6-year	94-99 years	98,643
10-year	90-99 years	286,307
11-year	89-99 years	355,657
29-year	71-99 years	3,675,749
30-year	70-99 years	4,035,466

Be aware that other restrictions may apply when reporting age ranges, such as only reporting ages 90 and older as a single category for HIPAA Safe Harbor. Of note, the HIPAA Safe Harbor method specifically identifies the following as an identifier: "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

[40] U.S. Census Bureau, "SEX BY SINGLE-YEAR AGE," 2020. Decennial Census, DEC Demographic and Housing Characteristics, Table PCT12, 2020, accessed on October 6, 2023, https://data.census.gov/table/DECENNIALDHC2020.PCT12?g=040XX00US06.

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." Although dates are included in the Safe Harbor list, age (<90 years old) is not. The risk score to age ranges reflects the two components of the scoring criteria: size of the potential population and the variable specificity.

16.2.4 Race Group and Ethnicity

Table 5: Race or Race/Ethnicity Combined

Characteristics	Score
White, Asian, Black or African American, Hispanic or Latino, Middle Eastern or North African	+2
White, Asian, Black or African American, Hispanic or Latino, Middle Eastern or North African, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Mixed	+3

Table 6: Detailed Race or Race/Ethnicity Combined

Characteristics	Score
Detailed Race or Race/Ethnicity Combined with Population >4,000,000 *e.g., Mexican*	+1
Detailed Race or Race/Ethnicity Combined with Population 300,001 – 4,000,000 *e.g., Chinese, Filipino, German, Asian Indian, Italian, Korean, Salvadoran, Guatemalan*	+2
Detailed Race or Race/Ethnicity Combined with Population 100,001 – 300,000 *e.g., Japanese, Armenian, Iranian, Aztec, Portuguese, Taiwanese, Hmong, Puerto Rican, Peruvian*	+3
Detailed Race or Race/Ethnicity Combined with Population 20,001 – 100,000 *e.g., Cambodian, Dutch, Pakistani, Egyptian, Thai, Maya, Afghan, Nigerian, Indonesian, Fijian, Native Hawaiian, Jamaican, Cuban, Colombian, Argentinean*	+5
Detailed Race or Race/Ethnicity Combined with Population ≤20,000 *e.g., Tongan, Chamorro, Bangladeshi, Sri Lankan, Brazilian, Mixtec, Kenyan, Zapotec, Malaysian, Belizean, Chumash, Sudanese, Pomo, Inca, Pipil*	+7

Table 7: Ethnicity Scoring

Characteristics	Score
Hispanic or Latino - yes or no	+1

Table 8: Detailed Ethnicity Scoring

Characteristics	Score
Detailed Ethnicity with Population >4,000,000 *e.g., Mexican*	+1
Detailed Ethnicity with Population 300,001 – 4,000,000 *e.g., Salvadoran, Guatemalan, Central American, South American*	+2
Detailed Ethnicity with Population 100,001 – 300,000 *e.g., Puerto Rican, Spaniard, Peruvian, Nicaraguan, Honduran*	+3
Detailed Ethnicity with Population 20,001 – 100,000 *e.g., Cuban, Colombian, Argentinean, Dominican, Panamanian*	+5
Detailed Ethnicity with Population ≤20,000 *e.g., Bolivian, Uruguayan, Paraguayan*	+7

Race and Ethnicity are collected in several different ways on the different state and federal data collection tools. At the federal level from 1997 to 2024 Office of Management and Budget required federal agencies to use a minimum of five race categories:[41]

- White,
- Black or African American,
- American Indian or Alaska Native,
- Asian, and

[41] U.S. Office of Management and Budget. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, https://www.federalregister.gov/d/97-28653.

- Native Hawaiian or Other Pacific Islander.

The 1997 OMB guidance required a separate question on Ethnicity which asked whether individuals are Hispanic or Latino. The 2020 Census asked if individuals were of Hispanic, Latino, or Spanish origin, and additional specific detail for ethnicity was requested. The US Census Bureau often reports this information as Hispanic or Latino origin and Hispanic origin interchangeably.

Table 9: California Population by Race Group based on the 2020 Census

Race Group	Population[42]
White alone	16,296,122
Black or African American alone	2,237,044
American Indian and Alaska Native alone	631,016
Asian alone	6,085,947
Native Hawaiian and Other Pacific Islander alone	157,263
Some Other Race alone	8,370,596
Two or more races	5,760,235

Table 10: California Population by Ethnicity or Race/Ethnicity based on the 2020 Census

Ethnicity or Race/Ethnicity	Population[43]
Hispanic or Latino, any race	15,579,652
Not Hispanic or Latino, any race	23,958,571
White alone, not Hispanic or Latino	13,714,587
Black or African American alone, not Hispanic or Latino	2,119,286

[42] U.S. Census Bureau, "RACE," 2020. Decennial Census, DEC Redistricting Data (PL 94-171), Table P1, 2020, accessed on October 6, 2023, https://data.census.gov/table/DECENNIALPL2020.P1?g=040XX00US06,06$0500000.

[43] U.S. Census Bureau, "HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE," 2020. Decennial Census, DEC Redistricting Data (PL 94-171), Table P2, 2020, accessed on October 6, 2023, https://data.census.gov/table/DECENNIALPL2020.P2?g=040XX00US06,06$0500000.

Ethnicity or Race/Ethnicity	Population[43]
American Indian and Alaska Native alone, not Hispanic or Latino	156,085
Asian alone, not Hispanic or Latino	5,978,795
Native Hawaiian and Other Pacific Islander alone, not Hispanic or Latino	138,167
Some Other Race alone, not Hispanic or Latino	223,929
Two or more races, not Hispanic or Latino	1,627,722

In 2024, the Office of Management and Budget approved an update[44] to “Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity”. The three major changes are:

- Addition of “Middle Eastern and North African” category. Previously these individuals were classified as White.
- Combination of race and ethnicity into a single question.
- Allowing individuals to select multiple racial groups that they identify with, rather than allowing only one selection.

The scores as described for population size (Table 2) were used to assess risk for race group, ethnicity, and race/ethnicity, based on comparable California populations as reported for the 2020 Census. Note that these risk scores are higher than the equivalent population numbers for location, as demographic traits such as age and race/ethnicity are publicly identifiable in a way that residence is not.

- Race group reported with the categories White, Asian, Black or African American, Middle Eastern and North African, and Hispanic or Latino have been assigned a +2 score. The smallest population is Middle Eastern and North African, estimated to be above 700,000 for California. The +2 score is consistent with the score in Table 2 for a statewide population of 300,001 to 4 million.
- Race group and race/ethnicity reporting that include the above categories plus American Indian and Pacific Islander are assessed a +3 score. The smallest (exclusionary) categories are Pacific Islander (157,263) and non-Hispanic/non-

[44] U.S. Office of Management and Budget. Initial Proposals For Updating OMB's Race and Ethnicity Statistical Standards, https://www.federalregister.gov/d/2023-01635.

Latino Pacific Islander (138,167). The +3 score is consistent with the score in Table 2 for a statewide population of 100,001 to 300,000.

- Scores for race, ethnicity and race/ethnicity are based on populations associated with these categories as reported for the 2020 Census, except for Middle Eastern and North African which was not recorded as a separate category but was allowed as a write-in response. If reported categories differ from the 2020 Census classification methodology, scores may need to be adjusted accordingly. For more information on these populations and categories see Appendix G.

While preparing the scores of race/ethnicity groups, it was not possible to consider the population size of counties since the distribution of each race/ethnicity group is different for each county in California. Please see the distribution of race/ethnicity groups within each county population in Appendix G. Thus, only statewide distributions of race/ethnicity groups within the California population are considered to score the race/ethnicity groups.

Examples

Three scenarios are presented below to help demonstrate how to use the race group and ethnicity scoring criteria for data that conforms to 1997-2024 OMB standards.

First Scenario – Complete Cross-Tabulation between Race Group and Ethnicity

Consider this table:

Table 11: First Scenario Example

Race Group	Hispanic	Non-Hispanic	Any Hispanic response
Black	50	250	300
White	200	1,000	1,200
Asian	5	95	100
Any race group	255	1,345	1,600

With this cross-tabulation, you would add both the Race score and Ethnicity score independently to the overall total for your scoring metric (i.e., greatest risk for re-identification). Note that you can replace “Ethnicity” with “Sex” and the principle still applies—you have a cross-tabulated table of Race and Sex.

Second Scenario – Race and Ethnicity merged into exclusive categories

Usually, the algorithm is that Ethnicity trumps Race when categorizing. This results in a Hispanic category, with the other categories effectively becoming "Non-Hispanic Race." Accordingly, the above table would become:

Table 12: Second Scenario Example

Mutually exclusive race/ethnicity	Number of Events
Non-Hispanic Black	250
Non-Hispanic White	1,000
Non-Hispanic Asian	95
Hispanic	255
Total	*1,600*

The second scenario is when you would use the combined Race/Ethnicity score in the guidelines for your scoring metric.

Third Scenario – No Interaction between Race and Ethnicity

Without an interaction between Race and Ethnicity, this could be reported as follows:

Table 13: Third Scenario Example

Race or Ethnicity	Number of Events
Black, any Hispanic response	300
White, any Hispanic response	1,200
Asian, any Hispanic response	100
Hispanic, any race group	255

Note that as displayed above, you cannot add up the categories to get a total population. For assigning a score, this is the same as reporting in two separate tables that are each scored independently:

Table 14: Third Scenario Example Continued

Race Group	Number of Events
Black	300
White	1,200
Asian	100
Total	*1,600*

Table 15: Third Scenario Example Continued

Ethnicity	Number of Events
Hispanic	255
Non-Hispanic	1,345
Total	*1,600*

Also, you would need to run the scoring metric separately for your Race-only and Ethnicity-only datasets. Like the First Scenario, you can replace Ethnicity with Sex and it still makes sense—you now have two tables, one displaying Race and the other Sex, with no interaction between the two—which lessens the Small Cell Size problem.

Risk Assessment for Detailed Race and Ethnicity Groups

The scores for detailed race, ethnicity, and race/ethnicity categories are harmonized with the scores for the minimum OMB categories and created based on Table 2.

Be aware that when reporting hierarchical data with multiple levels of the hierarchy, such as broad race/ethnicity alongside detailed race/ethnicity groups, that any complementary suppression algorithm will need to account for this dependent relationship between the values.

Risk Assessment for Multi-Racial and Multi-Ethnic Populations

Both OMB SPD 15 and California Government Code Section 8310.9 recommend providing the ability for individuals to “select all that apply” in order to capture multi-racial and multi-ethnic identities accurately. Data display would then aggregate all responses where a specific choice is selected.

- For example, the previous “exclusionary” approach would only count an individual as “Black” if “Black” was the only race selected. If another race group was also selected, that individual would be labeled as “Multi-race.”
- In the new "inclusionary" approach, all responses where “Black” was selected would be included in the “Black” category, regardless of whether another racial or ethnic choice was selected as well.

Statistically, this “inclusionary” approach would increase both numerator and denominator values for specific racial/ethnic groups and subgroups. Because groups would no longer be exclusionary of one another, it would also prevent the back-calculation of suppressed cells by subtracting non-suppressed cells from the total. Both these factors result in decreased re-identification risk. Therefore, no additional risk score is needed for an “inclusive” approach when displaying data on specific race/ethnicity groups and subgroups.

When identifying the score for a variable, use the highest scoring criteria. For example, if a table had age groups of 0 to 11 years (+2), 12 to 14 years (+5), and 15 to 18 years (+5) then the score for the Age Range variable would be +5 because the smallest age range is 12 to 14, which is an age range of three years. Similarly, if a table had race groups of Chinese (+2), Japanese (+3), Cambodian (+5), and Malaysian (+7) then the score for the Detailed Race Group variable would be +7 because it is the highest score for the reported groups.

16.2.5 Language Spoken

Table 16: Language Spoken Scoring

Characteristics	Score
English, Spanish, Other Language	+1
Detailed Language with Population 300,001 - 4,000,000 *e.g., Chinese, Tagalog, Vietnamese, Korean*	+2
Detailed Language with Population 100,001 - 300,000 *e.g., Persian, Hindi, Arabic, Russian, Japanese, French*	+3
Detailed Language with Population 20,001 - 100,000 *e.g., German, Portuguese, Hmong, Hebrew, Bengali, Polish*	+5
Detailed Language with Population ≤20,000 *e.g., Haitian, Navajo*	+7

Language spoken is captured in a variety of data systems to support individuals in receiving services in the language they speak. Language Spoken is not collected as part of the decennial Census, so the following estimates of language spoken at home are used to assess the population sizes of various languages. These estimates of language spoken are taken from the 2017-2021 5-year American Community Survey (ACS). These Statewide population estimates of language spoken along with Table 2 are used to determine the groupings for the scoring above.

Table 17: California Population Estimates by Language Spoken

Language Spoken at Home for the Population 5 years and older	California Estimate	Margin of Error
Total	37,105,018	±669
Speak only English	20,833,290	±49,117
Spanish	10,514,821	±34,689
Chinese (incl. Mandarin, Cantonese)	1,259,668	±13,770
Tagalog (incl. Filipino)	780,024	±10,075
Vietnamese	556,398	±8,380
Korean	358,018	±7,365
Persian (incl. Farsi, Dari)	211,089	±6,309
Hindi	203,238	±6,535
Arabic	198,914	±7,807
Armenian	195,413	±5,581
Russian	170,508	±4,817
Punjabi	142,450	±6,035
Japanese	136,009	±4,890
French (incl. Cajun)	126,338	±4,041
Ilocano, Samoan, Hawaiian, or other Austronesian languages (aggregated group)	120,223	±4,014
German	93,471	±2,737
Portuguese	91,042	±3,852
Thai, Lao, or other Tai-Kadai languages (aggregated group)	75,646	±3,518
Other Indo-European languages (aggregated group)	75,233	±3,511
Hmong	74,317	±3,670
Telugu	67,956	±3,101
Khmer	67,756	±3,552
Other languages of Asia (aggregated group)	65,969	±3,512

Language Spoken at Home for the Population 5 years and older	California Estimate	Margin of Error
Amharic, Somali, or other Afro-Asiatic languages (aggregated group)	63,318	±3,433
Tamil	60,594	±2,885
Nepali, Marathi, or other Indic languages (aggregated group)	58,552	±2,832
Urdu	54,569	±2,938
Italian	53,954	±2,128
Gujarati	51,662	±2,772
Hebrew	44,540	±2,542
Ukrainian or other Slavic languages (aggregated group)	41,606	±2,630
Malayalam, Kannada, or other Dravidian languages (aggregated group)	37,709	±2,672
Yoruba, Twi, Igbo, or other languages of Western Africa (aggregated group)	34,305	±2,830
Bengali	30,223	±2,043
Yiddish, Pennsylvania Dutch or other West Germanic languages (aggregated group)	26,359	±1,870
Polish	21,304	±1,639
Serbo-Croatian	20,022	±1,470
Greek	19,783	±1,576
Swahili or other languages of Central, Eastern, and Southern Africa (aggregated group)	16,547	±1,558
Haitian	7,878	±1,071
Other Native languages of North America (aggregated group)	5,841	±713
Navajo	1,043	±305
Other and unspecified languages	37,418	±2,389

Based on the above numbers, the majority of individuals speak English or Spanish. Therefore, if the table includes “English”, “Spanish”, and “Other Language” as the categories for “Language Spoken”, then the score is +1 which is comparable to reporting Hispanic or Latino Ethnicity as a “Yes or No”.

As noted for Race and Ethnicity demographics, language spoken demographics may vary significantly based on geography as well as based on particular conditions. So although the scoring criteria presents a guideline for assessing risk, the population frequencies for the specific geography and/or condition should also be taken into account.

If more specificity for Language Spoken is being requested with respect to reporting on the other languages in the table above, the request will need to be reviewed on a case-by-case basis. The additional review is necessary given the variability of language spoken by different populations or geographies and the consideration for potential increased risk of identification.

16.2.6 Sexual Orientation and Gender Identity

There are no census estimates for Sexual Orientation and Gender Identity (SOGI). Instead, we have based our risk scores on the California Health Interview Survey (CHIS) population estimates, as CHIS is the largest California survey and has publicly reproducible results via AskCHIS and harmonized with the risk scores assigned to the age and race/ethnicity category population thresholds.

AskCHIS 2021 population estimates for SOGI are as follows (Gender Identity population estimate is from pooled years 2019-2021 as per CHIS recommendations):

- Gay/lesbian: 1.2 mil
- Bisexual: 1.5 mil
- Asexual: 400K
- Transgender or non-conforming: 279K

Based on the risk scores for age and race/ethnicity, Sexual Orientation data aggregated to the above CHIS categories would be assigned a +2 score as the smallest group (Asexual) falls within the "300,001 – 4 million" risk category. Gender Identity data aggregated to the above CHIS categories would be assigned a +3 score as "Transgender or non-confirming" falls within the "100,001 – 300,000" risk category.

There are no estimates for more granular categories such as genderqueer and two-spirit. A +5 risk score was assigned based on the “Other Variable” category for 5-9 groups in the DDG.

The option for a variable limited to "Male or Female" has been kept for datasets that are limited to these options, or in cases where other categories aren't displayed due to confidentiality concerns. The risk score remains at +1 as both populations are greater than 4 million.

Table 18: Sex, Sexual Orientation, and Gender Identity Scoring

Variable	Characteristics	Score
Sex	Male or Female	+1
Sexual Orientation	Straight, Gay or Lesbian, Bisexual, Asexual	+2
Gender Identity	Man/Male, Woman/Female, Transgender or Non-Binary	+3
Gender Identity	Man/Male, Woman/Female, disaggregation of Transgender/ Non-Binary category into more specific identities (e.g. Genderqueer, Two-Spirit, etc.)	+5

Distinction Between Sexual Orientation and Men Having Sex with Men (MSM) Activity

While the CDC definition of sexual orientation includes an implicit behavior in "sexual attraction", this guidance defines Sexual Orientation and Gender Identity as identifiable factors (akin to race, ethnicity, and age) to distinguish them from sexual behaviors such as "Men having Sex with Men" (MSM). This is because the former may be a publicly identifiable trait and thus a re-identification risk, while the latter is, generally, not an identification risk. For example, some individuals may engage in MSM activity but not identify as gay, while others may identify as gay but practice abstinence.

We acknowledge that the boundary between the two is not clear cut and programs will need to assess on a case-by-case basis as to whether MSM should be considered a risk factor for redisclosure. An example of this is an incident cited in NIST SP 800-188 where geographic location from app data was used to infer MSM activity for an individual:

> *According to commercially available records of app signal data obtained by The Pillar, a mobile device correlated to Burrill emitted app data signals from the location-based hookup app Grindr on a near-daily basis during parts of 2018,*

> *2019, and 2020 — at both his USCCB office and his USCCB-owned residence, as well as during USCCB meetings and events in other cities.*

This incident demonstrates the possibility that if an individual is known to use an app or frequent certain locations, one may infer MSM activity for the individual. This makes it potentially usable as a quasi-identifier to re-identify an individual. As mentioned by Abowd and Hawes in 21st Century Statistical Disclosure Limitation: Motivations and Challenges:

> *The release of any statistic derived from a confidential source always carries some incremental risk of disclosure of identifiable information.*

16.2.7 Intersex

AB 1163 (Chapter 832, Statutes of 2023) added the requirement to collect intersex status in Government Code 8310.8 when collecting "ancestry or ethnic origins of Californians." There are no California estimates of the intersex population, but the national estimate is 1.7% of the total population. That would be a population of ~595K in California, which would be a +2 risk score as per the risk score table. The question can be asked as part of the Sex question or as its own question.

Table 19: Intersex Scoring

Variable	Characteristic	Score
Intersex (asked as separate question)	Yes or No	+2
Intersex (combined with Sex question)	Male, Female, Intersex	+2

16.2.8 Immigration Status

Table 20: Immigration Status Scoring

Characteristics	Score
U.S. Citizen, Foreign Born (combines Naturalized Citizen and Noncitizen)	+1
U.S. Citizen, Naturalized Citizen, Noncitizen	+1
Detailed Immigration Status with Disaggregation of Noncitizen Statuses – Refer to High-Risk Populations (Section 5.6.2)	N/A

In this version of DDG, immigration status is added as a new variable because immigration status in the summarized health care data may increase the re-identification risk of individuals due to increased granularity of the aggregated data. This additional information has the potential to make it easier to narrow down and identify individuals, especially if the dataset is combined with other publicly available sources. Immigration status often intersects with other demographic factors such as age, gender, ethnicity, and location. When combined, these factors can create a more distinct profile for certain individuals.

The U.S. Census Bureau[45] collects citizenship data via the American Community Survey (ACS) and categorizes the population as either "U.S. citizen" or "foreign born". The "foreign born" population is then categorized further as "natural citizen" or "noncitizen". The noncitizen status is disaggregated into the following statuses:

- Lawful Permanent Resident
- Nonimmigrant
    - Temporary Workers
    - Students
    - Exchange Visitors
    - Refugees and Asylees
- Undocumented Immigrant

Based on reporting by the U.S. Census Bureau and the U.S. Department of Homeland Security, population estimates for different characteristics of immigration status are as

[45] U.S. Census Bureau, " SELECTED CHARACTERISTICS OF THE NATIVE AND FOREIGN-BORN POPULATIONS," 2022. American Community Survey, ACS 5-Year Estimates Subject Tables, Table S0501, 2022, accessed on December 11, 2023, https://data.census.gov/table/ACSST5Y2022.S0501?q=S0501

below.[46, 47] However, the population estimates for undocumented immigrants are based on population statistics reported by the Center for Migration Studies.[48]

Table 21: Population Estimates Related to Immigration Status

Immigration Status	Population Estimate
U.S. Citizen	~28.8 million
Foreign Born	~10.5 million
Naturalized Citizens	~5.7 million
Noncitizens	~4.7 million
Lawful Permanent Residents	~2.2 million
Nonimmigrants	~560,000
Temporary Workers	~300,000
Students	~210,000
Exchange Visitors	~50,000
Refugees/Asylees	<10,000
Undocumented Immigrants	2,251,756[49]

Immigration Status scores are prepared based on the U.S. Census Bureau[50] population data collected on U.S. citizen and foreign-born populations (Table 21 rows for U.S.

[46] U.S. Department of Homeland Security, Office of Homeland Security Statistics, "Population Estimates of Nonimmigrants Residing in the United States: Fiscal Years 2017-2019" May 2021, Retrieved June 2, 2025, from https://ohss.dhs.gov/sites/default/files/2023-12/ni_population_estimates_fiscal_years_2017_-_2019v2.pdf

[47] U.S. Department of Homeland Security, Office of Homeland Security Statistics, "Refugees and Asylees: 2020" Annual Flow Report for Fiscal Year 2020. March 2022, Retrieved June 2, 2025, from https://ohss.dhs.gov/sites/default/files/2023-12/2022_0308_plcy_refugee_and_asylee_fy2020v2.pdf

[48] 2019 estimate from Estimates of Undocumented and Eligible-to-Naturalize Populations by State, Center for Migration Studies, http://data.cmsny.org/state.html.

[49] 2019 estimate from Estimates of Undocumented and Eligible-to-Naturalize Populations by State, Center for Migration Studies, http://data.cmsny.org/state.html.

[50] U.S. Census Bureau, " SELECTED CHARACTERISTICS OF THE NATIVE AND FOREIGN-BORN POPULATIONS," 2022. American Community Survey, ACS 5-Year Estimates Subject Tables, Table S0501, 2022, accessed on December 11, 2023, https://data.census.gov/table/ACSST5Y2022.S0501?q=S0501

Citizen, Foreign Born, Naturalized Citizens and Noncitizens). Disaggregated immigration status population sizes are prepared based on the Department of Homeland Security.

Population Estimates for lawful permanent residents[51] (Table 21 row for Lawful Permanent Resident) and nonimmigrants[52] (Table 21 rows for Nonimmigrants, Temporary Workers, Students, and Exchange Visitors), including Refugees/Asylees.[53] The scores are then harmonized with the risk scores assigned to the latest population size thresholds, Table 2. Immigration statuses for disaggregated statuses, and smaller corresponding population sizes are added to the high-risk populations due to the sensitive nature of these populations and an increased risk for identification.

16.2.9 Insurance Coverage

Insurance Coverage is a factor that may be a proxy for other identifiers. For example, employment-based payers may provide information about a person's job status. Other types of payers, such as Medicaid, may provide information about a person's income. There is demographic data that indicates populations by income level in the public domain as published by the U.S. Census Bureau[54] as well as public information related to various jobs and employment status through labor agencies.[55, 56] Given the opportunities to use information about the payer in combination with other public

[51] U.S. Department of Homeland Security. (2022, September). September 2022 Population estimates of the Lawful Permanent Resident Population in the United States and the Subpopulation Eligible to Naturalize: 2022. dhs.gov https://www.dhs.gov/sites/default/files/2022-10/2022_0920_plcy_lawful_permenent_resident_population_estimate_2022_0.pdf.

[52] U.S. Department of Homeland Security, Office of Homeland Security Statistics, "Population Estimates of Nonimmigrants Residing in the United States: Fiscal Years 2017-2019" May 2021, Retrieved June 2, 2025, from https://ohss.dhs.gov/sites/default/files/2023-12/ni_population_estimates_fiscal_years_2017_-_2019v2.pdf

[53] U.S. Department of Homeland Security, Office of Homeland Security Statistics, "Refugees and Asylees: 2020" Annual Flow Report for Fiscal Year 2020. March 2022, Retrieved June 2, 2025, from https://ohss.dhs.gov/sites/default/files/2023-12/2022_0308_plcy_refugee_and_asylee_fy2020v2.pdf

[54] U.S. Census Bureau, Census Datasets on income topics, (census.gov)

[55] U.S. Bureau of Labor Statistics, Occupation Finder: Occupational Outlook Handbook (bls.gov)

[56] California Employment Development Department, California Open Data Portal datasets, https://data.ca.gov/organization/california-employment-development-department

information, this variable is given a risk score. The risk is scored as below given the member size of the Insurance Coverage in the various categories.[57][58]

Table 22: Insurance Coverage

Characteristic	Score
Coverage with >2,000,000 members	-5
Coverage with 1,000,001 - 2,000,000 members	-3
Coverage with 560,001 - 1,000,000 members	-1
Coverage with 250,001 - 560,000 members	0
Coverage with 100,001 - 250,000 members	+1
Coverage with 50,001 - 100,000 members	+3
Coverage with 20,001 - 50,000 members	+4
Coverage with ≤ 20,000 members	+5

See Appendix I for more details on scoring scenarios involving the overlap of Insurance Coverage, Expected Payer/Public Assistance and Means-Tested Programs, and Geography. Below are three key points that summarize all the scenarios:

1) If the data is ONLY related to Residence or Service Geography, then DO NOT USE Insurance Coverage or Means-Tested Tables.
2) Means-Tested Programs—Only add interaction if enrollment in the Public Assistance program is 10 million or fewer people. No interaction is needed for Medi-Cal as the current enrollment is approximately 14 million, which exceeds 10 million.
3) If the number of members enrolled in Insurance Coverage is less than the population of the geographic subdivision, then use the Insurance Table. If the number of members enrolled in Insurance Coverage is greater than or equal to the population of the geographic subdivision, then use the Geography Table.

[57] Department of Health Care Access and Information, Health Care Payments Data Snapshot "Data Overview – Count of Individuals by Payer Type", https://hcai.ca.gov/visualizations/healthcare-payments-data-hpd-snapshot/

[58] Department of Health Care Access and Information, Report to the Legislature "Program Report: Health Care Payments Data Program", May 2024, https://hcai.ca.gov/wp-content/uploads/2024/04/HPD-Report-to-the-Legislature-March-2024-1.pdf

16.2.10 Expected Payer/ Public Assistance and Means-Tested Programs

Expected Payer is a factor that may be a proxy for other identifiers. For example, employment-based payers may provide information about a person's job status. Other types of payers, such as Medicaid, may provide information about a person's income. There is demographic data that indicates populations by income level in the public domain as published by the U.S. Census Bureau,[59] as well as public information related to various jobs and employment status through labor agencies.[60, 61] Given the opportunities to use information about the payer in combination with other public information, this variable is given a risk score.

It is important to be aware of the potential risks associated with this data and to ensure its security and protection. For instance, eligibility for benefits in the Medi-Cal program may be determined based on income, property, and assets. Similarly, self-pay data, often associated with the uninsured, can also indicate low income, highlighting the need for data security.

Table 23: Expected Payer/ Public Assistance and Means-Tested Programs Scoring

Size of program enrollment	Score
Enrollment > 10,000,000 people	+0
Enrollment > 4,000,001 – 10,000,000	+1
Enrollment 300,001 – 4,000,000	+2
Enrollment 100,001 – 300,000	+3
Enrollment 20,001 – 100,000	+5
Enrollment ≤20,000	+7

It is important to note that Medi-Cal is a Means-Tested Program, and overall enrollment is managed at the statewide level. In 2024, approximately 14 million Californians were enrolled in the Medi-Cal program, therefore, the Medi-Cal category has been assigned a +0 DDG score. Also, more than six million California residents had Medicare coverage, therefore, the Medicare category has been assigned a +1 DDG score. There are

[59] U.S. Census Bureau, Census Datasets on income topics, (census.gov)

[60] U.S. Bureau of Labor Statistics, Occupation Finder: Occupational Outlook Handbook (bls.gov)

[61] California Employment Development Department, California Open Data Portal datasets, https://data.ca.gov/organization/california-employment-development-department

approximately 5.5 million patients captured in the HCAI Health Care Payments Database who have private insurance and thus assigned as +1 score. The uninsured population, who we presume would self-pay seems to be less than four million (2,752,067 based on the 2022 five-year American Community Survey estimate for California), has been assigned a +2 score.

In addition to Medi-Cal, CalHHS administers and collects data on a variety of public assistance programs[62] or means-tested programs.[63] In these programs, a granting entity uses information on an individual's income and resources to determine eligibility for the program. Knowing that an individual is on a means-tested program or a public assistance program can reveal sensitive information about their income, employment status, or other personal characteristic related to eligibility. For example, recipients of SNAP are typically at 130% of the Federal Poverty Level. To qualify for TANF, assistance units must be at 100% of the Federal Poverty Level. Major federal programs include Medicaid (Medi-Cal in California), the Earned Income Tax Credit, the Supplemental Nutrition Assistance Program (CalFresh), Supplemental Security Income (SSI), and Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). Additionally, there are several state-funded public assistance programs, including state anti-poverty tax credits, housing and homelessness programs, and programs such as the Cash Assistance Program for Immigrants (CAPI) and the California Food Assistance Program (CFAP).

The size of these programs ranges widely; as noted in the previous section, 14 million Californians are enrolled in Medi-Cal. Roughly 5.5 million Californians participated in CalFresh in 2024.[64] CalWORKs served a little over 900,000 Californians in 2023.[65] In the

[62] U.S. Census Bureau, "About Program Income and Public Assistance," (2025) https://www.census.gov/topics/income-poverty/public-assistance/about.html, Accessed March 25, 2025

[63] https://studentaid.gov/help/means-tested-benefits

[64] California Department of Social Services. CalFresh Dashboard. https://public.tableau.com/app/profile/california.department.of.social.services/viz/CFdashboard-PUBLIC/AnnualParticipation. 2025. Accessed March 25, 2025.

[65] California Department of Social Services. CalWORKs Take up. CalWORKs Interactive Summary. https://www.cdss.ca.gov/inforesources/calworks-summary/program-overview/take-up#C018. 2025. Accessed March 25, 2025.

Housing Support Program, fewer than 13,000 applications were approved.[66] Risk scores should be assigned to data related to (or breaking out) participation by program enrollment. To prevent the inadvertent disclosure of an individual's participation in a means-tested or public assistance program, the following risk scores can be assigned to data that includes information on program participation, based on the size of the program (note, these numbers are derived from Table 2) and Means-Tested Programs, and Geography.

16.2.11 Geography

If the level of reporting is best described by the geography of the individual/service, use one of the following two tables. Specifically, 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". Also see Appendix I for details on scoring scenarios involving the overlap of Insurance Coverage, Expected Payer/Public Assistance and Means-Tested Programs, and Geography.

Table 24: Residence Geography Scoring

Characteristics	Score
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 250,001 - 560,000	0
Population 100,001 - 250,000	+1
Population 50,001 - 100,000	+3
Population 20,001 - 50,000	+4
Population 4,001- 20,000	+5
Population ≤ 4,000	+7

[66] California Department of Social Services. Housing Support Program. CalWORKs Interactive Summary. https://www.cdss.ca.gov/inforesources/calworks-summary/program-services-utilization/housing-and-homelessness#C004. 2025. Accessed March 25, 2025

Table 25: Service Geography Scoring

Characteristics	Score
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 250,001 - 560,000	-1
Population of reporting region 20,001 - 250,000	0
Population of reporting region ≤20,000	+1
Address (Street and ZIP)	+3
Address in rural [67] area	+5
Address in frontier [68, 69] area	+7

The Geography score, while it may or may not represent the denominator of the table, does provide a reference to the base population about which the reporting is occurring. This will often be reflected in the title of the table if a statewide table. Otherwise, the geography may be represented in the rows or columns. There are two different scoring sets based on whether the geography reporting is based on the residence of the individual to which the information applies or to the service location.

The scores are higher for geography related to residence address because so much information is publicly available about individuals and their address of residence. For large populations greater than 560,000, which is equivalent to the size of a state, there is a negative score because the size of the denominator masks the individual. The number 560,000 was chosen as a cut-off because this is the size of the smallest state (Wyoming).

[67] U.S. Census Bureau, 2020 Census "Urban and Rural" Classifications, https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html

[68] National Rural Health Association, "Definition of Frontier", Retrieved June 2, 2025, from https://www.ruralhealth.us/getmedia/132306e1-1643-4b40-818a-4d743317dc7e/NRHAFrontierDefPolicyPaperFeb2016.pdf

[69] U.S. Department of Agriculture Economic Research Service, "Frontier and Remote Area Codes", Retrieved June 2, 2025, from https://www.ers.usda.gov/data-products/frontier-and-remote-area-codes

We chose to use the cut-off at the smallest state's population because state level reporting is not listed as one of the 18 identifiers in the HIPAA Safe Harbor method.

The scores for the service geography are lower because clients can generally come from diverse locations for services. Although people often seek services or have health conditions close to their homes, they may also travel extensive distances. Reviewers do need to make sure that there are no constraints associated with services that would mean the service geography and resident geography are the same. For example, if a program publishes service utilization by county and the county services can only be used by county residents, then the service utilization by county is also the county of residence. Scoring should be based on the criteria that result in the highest score and, thus, the highest risk.

There are smaller areas within counties where the population is significantly lower than the overall county population. One example of this is census tracts, which are small, relatively stable statistical subdivisions of a county or an equivalent statistical entity. These tracts can be updated by local participants before each decennial census through the Census Bureau's Participant Statistical Areas Program (PSAP). Census tracts[70] typically cover contiguous areas, but their size can vary widely based on population density. The boundaries of these tracts are designed to remain unchanged over time, allowing for consistent statistical comparisons from one census to the next. Generally, census tracts have populations ranging from 1,200 to 8,000 people, with an optimum size of population around 4,000. To account for the higher re-identification risk associated with smaller populations, a score of +7 is assigned when the population size is 4,000 or fewer in the residential area.

Service Geography includes a level of detail that is identified as "Address (Street and ZIP)." This deals with reporting by provider (hospital, clinic, provider office, etc.) Provider addresses are public information and are public at the street address level. A given provider will tend to have a standard catchment area or the geographic boundaries from which most patients come from. This information is published by the Department of Health Care Access and Information (previously the Office of Statewide Health

[70] United States Census Bureau, Geography Program Glossary, https://www.census.gov/programs-surveys/geography/about/glossary.html#:~:text=Census%20Tract,-Census%20Tracts%20are&text=The%20primary%20purpose%20of%20census,optimum%20size%20of%204%2C000%20people.

Planning and Development - OSHPD)[71] for hospitals. While this addresses where most patients or clients come from, patients or clients may also come from outside the catchment area. For that reason, this does not score as high as the more detailed geography under Residence Geography.

However, addresses associated with rural and frontier areas have a higher re-identification risk due to lower population density per square mile and isolation of communities in these areas. Thus, higher scores are assigned for Providers' addresses, except where the Provider is defined as Hospital, in rural and frontier areas.[72, 73]

16.2.12 Time – Reporting Period

Table 26: Time – Reporting Period Scoring

Characteristics	Score
5 years aggregated	-5
2-4 years aggregated	-3
1 year (e.g., 2001)	0
Bi-Annual	+3
Quarterly	+4
Monthly	+5

Many reports are published based on the calendar year. However, the combination of years of data is an excellent way to provide increased aggregation in a way that allows for more specificity elsewhere, such as county identifiers. Inversely, the smaller the time period in the data, the closer the time period comes to approximating a date. Thus, monthly reported data has a high score of +5.

Of note, the HIPAA Safe Harbor method list includes "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

[71] Department of Health Care Access and Information, "Facility Market Share and Patient Origin" Reports, Retrieved June 2, 2025, from https://hcai.ca.gov/visualizations/facility-market-share-and-patient-origin/

[72] Medical Service Study Areas 2010, Retrieved on April 2, 2024, from: https://gis.data.ca.gov/maps/fe411f2d74494b89a74ab181b22fc8a1/about

[73] United States Census Bureau, "About Geographic Areas: Urban and Rural", Retrieved on April 2, 2024, from: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html.

year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older." This is a potential identifier when in combination with other information. This potential as an identifier influences the higher scores in the Publication Scoring Criteria as the time period for aggregation gets smaller.

The "0" value for this variable is set at one year as this is the criteria for Safe Harbor under the HIPAA de-identification standard.

16.2.13 Variable Interactions

Table 27: Variable Interactions Scoring

Characteristics	Score
Only Events (minimum of 5), Time, and Population (Residence/Service Geography or Insurance Coverage	-5
Only Events (minimum of 3), Time, and Population (Residence/Service Geo. or Insurance Coverage)	-3
Only Events (no minimum), Time, and Population (Residence/Service Geo. or Insurance Coverage)	0
Events, Time, and Population (Residence/Service Geo. or Insurance Coverage) + 1 variable	+1
Events, Time, and Population (Residence/Service Geo. or Insurance Coverage) + 2 variables	+2
Events, Time, and Population (Residence/Service Geo. or Insurance Coverage) + 3 variables	+4

These criteria specifically address the interaction of the variables in a given data presentation and require the analyst to identify dependent as opposed to independent variables. These criteria are used with respect to dependent variables. This is demonstrated in the two tables below.

Illustration A: Dependent Variables

In this example the Event (counts of Disease A) is shown for Males who are also 0-17 years old or Males who are also 18-25 years old. In this case Sex and Age are dependent because the stratification for each variable is stacked. This commonly occurs in pivot tables.

Table 28: Illustration A: Dependent Variables Example

Counts of disease A by year	Males and 0-17 years old	Males and 18-25 years old	Females and 0-17 years old	Females and 18-25 years old
Year 1	6	10	5	8
Year 2	8	14	3	20

Illustration B: Independent Variables

In this example the Event (counts of Disease A) is for Males or Females which is shown side by side to a table with ages 0-17 years old or 18-25 years old. In this case Sex and Age are independent because the stratification for each variable is not stacked. Although the two variables Sex and Age are shown in the same table, they are presented independently of each other. While you can compile the data in Example B from Example A, the reverse is not true.

Table 29: Illustration B: Independent Variables Example

Counts of disease A by year	Males	Females	0-17 years old	18-25 years old
Year 1	16	13	11	18
Year 2	22	23	11	34

These criteria are structured to have less impact if personal characteristics outside of time and geography are excluded and more impact if multiple personal characteristics are included. This provides for a subtraction of points if the only variables presented are the events (numerator), time and geography and an addition of points for including more variables in a given presentation. With respect to the subtraction of points, the score is based on the minimum value for the Events variable. For example, if the smallest value for the Events is 5 or more, then the score would be -5. However, if the smallest value for the Events is 2, then the score would be 0.

The minimum value for Events of 3 (Only Events (minimum of 3), Time, and Geography (Residence or Service)) is used as a threshold to address concern for pre-existing knowledge by users about individuals. For example, if an entity knows who one person is with disease A and the count for Events is “1” or “2”, then the entity could identify the person they know of or the person they know of plus information about the other person. The use of a minimum of 3 does not protect against two entities colluding to determine a third person.[74] For this reason, the threshold of 5 for Events is also given. The threshold of 5 is frequently used in public health reporting regarding various events.

In contrast, if additional demographic variables are added, then the risk increases significantly. For example, for Events, Time and Geography (Residence or Service) with three additional variables, a table would show how many individuals are female by age group by race for a given time period and geography. This allows for a more detailed comparison to census data and assessment of the number of individuals with a particular set of characteristics.[75] For this reason, additional points are added because of the inclusion of multiple dependent variables.

16.2.14 Other Variables

Variables not specified in the Publication Scoring Criteria can be released only after an additional, case-by-case review by the department's Statistical De-Identification Expert or Statistical De-Identification Supervisor Expert. We suggest that this review considers population size-based scores if population size of the characteristics of a variable is present, or otherwise scores based on number of groups or categories.

1. Other Variable Scores Based on Population Size

Table 30: Generalized Scoring Criteria based on Statewide Population

Population Size	Score
Population >4,000,000	+1
Population 300,001 – 4,000,000	+2
Population 100,001 – 300,000	+3
Population 20,001 – 100,000	+5
Population ≤20,000	+7

[74] NORC, "NORC Recommendations for California Department of Health Care Services (DHCS) Data De-Identification Guidelines (DDG)," January 8, 2016.

[75] NORC, "Case Study: The Disclosure Risk Implications of Small Cells Combined with Multiple Tables or External Data," January 8, 2016.

The above table shows scoring criteria for population size in general and can be used to score veteran status and educational attainment. Examples follow below of how these variables can be considered by using data at the Census American Community Survey.[76]

Veteran Status

For the California statewide population estimate of 1,467,025 listed below for the adult civilian veteran population, a score of +2 would be assigned from Table 2 based on Statewide Population (Population 300,001 – 4,000,000).

Table 31: Veteran Status recorded from the 2021 American Community Survey

Veteran Status for Population 18 years and over	California Estimate	Margin of Error
Civilian veterans	1,467,026	±9,913

Note that even though there is an implicit age associated with veteran status, this information is incorporated into the population estimate used for the score. Therefore, an additional score modifier for the age should not be applied.

Other aspects of the subpopulation should be considered when assessing risk. For example, more than 90 percent of the veteran population is male.[77] Therefore, reporting of veterans by sex would have more risk than the scores for sex based on the statewide population due to the small number of female veterans.

[76] U.S. Census Bureau, "Selected Social Characteristics in the United States," 2021. American Community Survey, ACS 5-Year Estimates Data Profiles, Table DP02, 2021, accessed on December 5, 2023, https://data.census.gov/table/ACSDP5Y2021.DP02?g=040XX00US06.

[77] U.S. Department of Veterans Affairs, “VetPop2020 State Estimates 2000 to 2020,” updated on April 6, 2023, Retrieved December 2024 from https://www.data.va.gov/dataset/VetPop2020-State-Estimates-2000-to-2020/fkjq-z6m8.

Educational Attainment

Similarly, the population estimates[78] in the table below can be used to score Education Attainment Status by using Table 2.

Table 32: Educational Attainment Status recorded from the 2021 American Community Survey

Educational Attainment for Population 25 years and over	California Estimate	Margin of Error
Less than 9th grade, no high school diploma	2,342,364	±15,809
9th to 12th grade, no high school diploma	1,893,671	±12,037
High school graduate (includes equivalency)	5,477,154	±28,244
Some college, no degree	5,496,578	±16,961
Associate degree	2,135,865	±13,333
Bachelor degree	5,855,383	±22,797
Graduate or professional degree	3,596,055	±25,828
High school graduate or higher	22,561,035	±20,560
Bachelor's degree or higher	9,451,438	±40,058

For example, the smallest California population with a recorded educational attainment level is 2,135,865 for an Associate degree. Thus, a score of +2 (Population 300,001 – 4,000,000) from the Table 2 would be assigned. Whereas the score would be +1 (Population >4,000,000) if educational attainment is combined from the table into two categories as follows:

- No college (sum of rows 1-3): 9,713,189 (2,342,364 + 1,893,671 + 5,477,154)
- At least some college (sum of rows 4-7): 17,083,881 (5,496,578 + 2,135,865 + 5,855,383 + 3,596,055)

[78] U.S. Census Bureau, 2017-2021 American Community Survey 5-Year Estimates, "2021 Educational Attainment, California Estimates, Population 25 Years and Over", Table S1501: https://data.census.gov/table/ACSST5Y2021.S1501?q=S1501&g=040XX00US06&y=2021

2. Other Variable Scores Based on Number of Groups or Categories

Table 33: Other Variable Scoring Based on Number of Groups or Categories

Number of Groups or Categories	Score
<5 groups or categories	+3
5-9 groups	+5
10+ groups	+7

If population size is not available for a variable, then score the variable based on number of groups or categories as well as the characteristics of the variables.

For example, legal class groupings associated with a patient's commitment to the state hospital system are an example of groups or categories. At the highest level, patients can be categorized into two groups: forensic commitment or civil commitments. At a more granular level, patients can be categorized into specific legal classes such defendants found incompetent to stand trial, parolees diagnosed with mental health disorders, individuals found to be not guilty by reason of insanity, mentally ill prisoners transferred from prison to a state hospital for mental health care, individuals committed to the state hospital system as sexually violent predators, or patients civilly committed to a state hospital as described in the Lanterman-Petris-Short Act.

Table 34: Example Scoring for Legal Class Groupings

Groups or Categories	Score
2 Groups Forensic commitments Civil commitments	+3
6 Groups Incompetent to stand trial Offenders with a mental health disorder Not guilty by reason of insanity Mentally ill prisoners Sexually violent predators Lanterman-Petris-Short Act commitments	+5

Some additional examples can be found in "Data with more Specificity" Section 5.6.1. Along with number of groups or categories, consider the specificity of the groups or categories, that is, whether the variable represents an aggregation (e.g., Diagnosis Related Groups) or a specific item (e.g., ICD-10 Code).

Also consider the availability of the variable to the public when also associated with other information, particularly the availability of variables that are personal characteristics.

How publicly identifiable the trait in question is should also be considered. For example, unreconstructed cleft palate is a physically identifiable trait. Dyslexia is a condition that, while not physically identifiable, may be something an individual mentions they have and should also be considered a publicly identifiable trait.

17) APPENDIX E. MORE ABOUT STATISTICAL MASKING METHODS

There are a number of alternative methods that are available for statistical masking. Use of any alternative method for de-identification must be reviewed and approved by the PRT.

Methods discussed in the Federal Committee of Statistical Methodology's "43333333333Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology"[79] include the following for tables of counts or frequencies and for magnitude data. Additional information is available through the committee's Data Protection Toolkit.[80]

- Tables of Counts or Frequencies
    - Sampling as a Statistical Disclosure Limitation Method
    - Defining Sensitive Cells
        - Special Rules
        - The Threshold Rule
    - Protecting Sensitive Cells After Tabulation
        - Suppression
        - Random Rounding
        - Controlled Rounding
        - Controlled Tabular Adjustment
    - Protecting Sensitive Cells Before Tabulation
- Tables of Magnitude Data
    - Defining Sensitive Cells – Linear Sensitivity Rules

[79] Federal Committee on Statistical Methodology, "Statistical Policy Working Paper 22 – Report on Statistical Disclosure Limitation Methodology," Second Version, 2005, Office of Management and Budget, Accessed from: https://www.fcsm.gov/assets/files/docs/spwp22WithFrontNote.pdf

[80] Federal Committee on Statistical Methodology, Data Protection Toolkit: Report and Resources on Statistical Disclosure Methodology and Tiered Data Access (formerly "Statistical Policy Working Paper #22") rev. 2024-11-21, Available at: https://nces.ed.gov/fcsm/dpt/versions

- Protecting Sensitive Cells After Tabulation
- Protecting Sensitive Cells Before Tabulation

18) APPENDIX F: PUBLICLY AVAILABLE DATA

A critical step in reviewing data for public release is the consideration of what other data may be publicly available that could be used in combination with the newly released data to identify the individuals represented in the data. This section will highlight some specific datasets that are publicly available that may be used in combination with CalHHS data that would contribute to potential increased risk.

Common kinds of data with personal information include: real estate records, individual licensing databases (MD, RN, contractors, lawyers, etc.), marriage records, news (and other) media reports, commercially available databases (data brokers, marketing), court documents, etc.

18.1 Vital Records Data

Another common dataset for programs to be aware of are the publicly available electronic birth and death indices from Vital Records, as specified in Health and Safety Code section 102230(b).

The following are provided in the birth record indices:

- First, middle, and last name
- Sex
- Date of birth
- Place of birth

The following are provided in the death record indices:

- First, middle, and last name
- Sex
- Date of birth
- Place of birth
- Date of death
- Place of death
- Father’s last name

Other potential sources of publicly available data to consider are informational certified copies of birth and death certificates. In California, anyone can obtain an informational certified copy of birth and death certificates, which are clearly marked as un-authorized copies that cannot be used to verify identity. In reality, it is difficult to use these as a dataset for the following reasons:

- Certified copies of birth and death certificates must be obtained on an individual basis, and you must be able to identify the record. In other words, an individual cannot simply ask for a stack of certificates for purposes of creating a dataset.
- Certified copies are issued on specialized banknote paper, not in electronic format, which creates a problem of scale when trying to create a dataset.
- There is a $29 fee for each certified copy of a birth certificate and $24 fee for a certified copy of a death certificate, which also creates a problem of scale when trying to create a dataset.
- Certified copies are meant for individual use. A request for a large number of certificates may generate an investigation among vital records staff as to why so many certificates were requested at once.

18.2 CalHHS Open Data Portal

As additional data sets are added to the Open Data Portal, programs need to take that information into account when considering potential risk for any given dataset. The CalHHS Open Data Workgroup will be providing easier access to both lists of data currently on the portal as well as datasets planned for addition to the portal. While significant with over 100 datasets, this is not exhaustive because of the PRA, which allows for an extremely broad amount of information to be released in a sporadic way. So, some specificity can occur but not completely. CalHHS departments have a duty of due diligence in the de-identification process regarding consideration of published identifiable data, published de-identified data, and the soon to be published de-identified data. Additional information[81] that addresses the balance of transparency and privacy include the Berkman Klein Center for Internet & Society's "Open Data Privacy Playbook".[82]

Listed below are examples of individual records or documents that the Department of Rehabilitation have available to the public:

[81] Bambauer, "Tragedy of the Data Commons". Harvard Journal of Law and Technology, Vol. 25, 2011

[82] Green, et al. 2017. Open Data Privacy (2017). Berkman Klein Center for Internet & Society Research Publication, http://nrs.harvard.edu/urn-3:HUL.InstRepos:30340010.

- Fair Hearing Decisions include the appellant's initials and possibly other information, depending on the issue the appellant presents for hearing, such as sex, disability, employment, education, vocational rehabilitation services, etc.; and
- Monthly Operating Reports and information therefrom includes names of licensees and financial information regarding the operation of the licensees' operation of vending facilities in the Business Enterprises Program for the Blind. To be eligible for this program, the individuals must be legally blind.

18.3 Public Census and Demographic Information

The Demographic Research Unit (DRU) of the California Department of Finance is designated as the single official source of demographic data for state planning and budgeting.[83] The DRU produces the following products which serve as the basis for understanding the population characteristics and distributions that frequently make up the denominators in the review of datasets.

- Estimates – Official population estimates of the state, counties and cities produced by the Demographic Research Unit for state planning and budgeting.
- Projections – Forecasts of population, births and public school enrollment at the state and county level produced by the Demographic Research Unit.
- State Census Data Center – Demographic, social, economic, migration, and housing data from the decennial censuses, the American Community Survey, the Current Population Survey, and other special and periodic surveys.

18.4 Commonly Shared Information

With the growth of social media, people frequently share information through tools such as Facebook, LinkedIn, Instagram, TikTok, YouTube, X (formerly Twitter), dating apps, and AI platforms such as Chat GPT, Open AI, and Google Cloud Vertex AI. While it would be impossible to take into account all information that people make public about themselves, there is an expectation that a certain amount of information is likely to be in the public domain based on information individuals frequently provide about themselves. Examples of such information include wedding dates, birth dates, education (high school, college), and professional certifications.

[83] California Department of Finance Demographics Research Unit, https://dof.ca.gov/forecasting/demographics/

18.5 Geographic Information

Geographic information is particularly suited to being combined with other geographic information given the relatively standardized way data is coded (latitude, longitude, county, etc.). With the use of mapping tools, various information can be combined in a way that is called a "mash up." A "mashup", in web development, is a web page, or web application, that uses content from more than one source to create a single new service displayed in a single graphical interface. For example, you could combine the addresses and photographs of your library branches with a Google map to create a map mashup. The term implies easy, fast integration, frequently using open application programming interfaces (open API) and data sources to produce enriched results that were not necessarily the original reason for producing the raw source data."[84]

18.6 Artificial Intelligence

With the rapid advancement in the use of health care data for artificial intelligence, machine learning tools, and data-generative models, there may be an increased risk of reidentification. These technologies enable the processing of large volumes of complex, unstructured raw data, but issues of liability and accountability remain unclear and unaddressed. Continued risk assessment is an important aspect of the guidelines presented in the DDG.

[84] http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)

19) APPENDIX G: STATE AND COUNTY POPULATION PROJECTIONS

View the “CalHHS Data De-Identification Guidelines Reference Dataset” resource on the California Health & Human Services Agency’s Open Data Portal.

This DDG resource:

- Provides California population estimates related to the race and ethnicity composition at the county level based on data from the 2020 U.S. Census.
- Includes more detailed race and ethnicity breakdowns than provided in prior DDG versions.

20) APPENDIX H: SURVEY AND SUPPRESSION RULES

The table in this appendix provides information from a 2002 article[85] with additional updates regarding rules for data suppression, statistical unreliability, and data quality in the display of data for Healthy People 2010. The main reasons for suppressing data in Healthy People 2010 included:

- The number of events is too small to produce reliable estimates or may violate confidentiality requirements.
- The sample design does not yield representative estimates for a particular group.
- There is a high level of item nonresponse or a significant number of unknown entries.

Due to the varying criteria for data suppression adopted by the different data systems that monitor the health status of the U.S. population, Klein, et al, summarized these criteria and provided information on specific population groups for which data systems cannot reliably estimate health metrics. They outlined the criteria for the following surveys:

- Behavioral Risk Factor Surveillance System (BRFSS)
- Medical Expenditure Panel Survey (MEPS)
- Monitoring the Future (MTF)
- National Ambulatory Medical Care Survey (NAMCS) and National Hospital Ambulatory Medical Care Survey (NHAMCS)
- National Crime Victimization Survey (NCVS)
- National Health Interview Survey (NHIS)
- National Health and Nutrition Examination Survey (NHANES) & Continuing Survey of Food Intake by Individuals (CSFII)
- National Household Survey on Drug Abuse (NHSDA)

[85] Klein RJ, Proctor SE, Boudreault MA, Turczyn KM. Healthy People 2010 criteria for data suppression. Healthy People 2010 Stat Notes. 2002 Jul;(24):1-12. PMID: 12117004. https://www.cdc.gov/nchs/data/statnt/statnt24.pdf

- National Survey of Family Growth (NSFG)
- School Health Policies and Programs Study (SHPPS)
- Youth Risk Behavior Surveillance System (YRBSS)
- National Hospital Discharge Survey (NHDS)

We have added the California Health Interview Survey (CHIS) and Consumer Assessment of Healthcare Providers & Systems (CAHPS) survey to the list above. The updated information for each survey is organized by a brief background, rules for statistical unreliability, and data de-identification/protection rules. Additionally, we have further updated guidance with information from a 2017 publication[86] by the National Center for Health Statistics (NCHS).

[86] Parker, et al, "Data Presentation Standards for Proportions", National Center for Health Statistics. Vital Health Stat 2(175). 2017; Retrieved from: https://www.cdc.gov/nchs/data/series/sr_02/sr02_175.pdf

Table 35: Surveys and Suppression Rules

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
California Health Interview Survey (CHIS)	The California Health Interview Survey (CHIS) began in 2001 as a biennial population-based, omnibus health survey of Californians. CHIS is a mixed-mode (web and telephone) survey that uses an address-based sampling (ABS) frame, making it representative of the state's population. CHIS data can be analyzed at the county level for the state's 41 most populated counties. The remaining 17 counties are combined into three different groups. Overall, the CHIS sample is designed to provide population-based estimates for most California counties and all major ethnic groups, including several ethnic subgroups.	Survey data (including the California Health Interview Survey) values are suppressed if the Relative Standard Error (RSE) is greater than 20% and, if statewide or non-stratified county-wide data, deemed likely to be misleading based on individual review of the data. The RSE is calculated using the standard methods noted above–this approach differs slightly from some other users of the California Health Interview Survey data, where the divisor for the RSE is 100 minus the percent, if the estimate is > 50%. While 30% or 23% cut points are more standard, CHIS determined that a 20% cut point suppressed many potentially misleading values not suppressed with the standard values. CHIS did not use the "100 minus the percent" approach because they determined that it	The CHIS Data Access Center generally does not allow the release of output containing the following information, unless special approval has been received: • Frequencies that do not meet the cell suppression guidelines; • Estimates run on sub-stratum geographical areas, i.e., by aggregated zip codes or for counties that comprise part of a stratum; • Analyses that include the most highly sensitive or most highly identifiable variables with small cell size issues; or • If most tables in the output require full suppression. CHIS reviews all output for small (raw frequency) cell sizes and makes release decisions accordingly. CHIS does not generate frequency/counts or

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
	https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/chis-design-and-methods/chis-design	suppressed many values that were unlikely to be misleading. https://letsgethealthy.ca.gov/progress/understanding-this-data/	cross-tabulations at the sub-county level due to confidentiality and weighting concerns. Small cell values are defined as less than 3 (unweighted) and less than 500 (weighted). In cross-tabulations, complementary cell values are also suppressed to avoid back-calculation. https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/access-chis-data/chis-frequently-asked-questions-faqs#dac
Consumer Assessment of Healthcare Providers & Systems (CAHPS)	The CAHPS Survey is used to assess managed care beneficiaries' satisfaction with their health care services. The goal of the CAHPS Health Plan Survey is to provide performance feedback that is actionable and will aid in improving	According to the National Committee for Quality Assurance (NCQA) Healthcare Effectiveness Data and Information Set (HEDIS) Specifications for Survey Measures, if a measure has fewer than 100 responses, the measure is not reportable.	There are three categories of suppression types: Item Suppression, Program Type Suppression, and Reporting Category Suppression. 1. Item Suppression: If there are fewer than 20 valid responses

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
	overall beneficiary satisfaction. https://www.cms.gov/data-research/research/consumer-assessment-healthcare-providers-systems		available for any item, the item's results are suppressed. 2. Program Type Suppression: If there are fewer than 20 completed or partially completed surveys for a given program, the program is excluded from the Database. 3. Reporting Category Suppression: If there are fewer than 10 programs for a given characteristic (e.g., region), CAHPS does not show results for the characteristic. Given the limited number of programs in the 2024 Home and Community-Based Services CAHPS Database, no breakouts by program type or program characteristics are provided. https://www.ahrq.gov/sites/default/files/wysiwyg/cahps/cahps-

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			database/2024-hcbs-chartbook.pdf Health Services Advisory Group (HSAG) suppressed MCPs results with fewer than 11 cases in the numerator to satisfy the Health Insurance Portability and Accountability Act of 1996 Privacy Rule's de-identification standards.
Behavioral Risk Factor Surveillance System (BRFSS)	BRFSS is an ongoing, State-based system of health surveys conducted by telephone interview using random digit-dialed probability samples of adults ages 18 years and over.	Estimates are considered statistically unreliable and are suppressed if the denominator is based on fewer than 50 sample cases.	There are no published data-deidentification guidelines for the BRFSS. However, the CA BRFSS does adhere to guidelines set forth in the California Health Survey Data.
Medical Expenditure Panel Survey (MEPS)	MEPS is an annual, nationally representative subsample of respondents to the National Health Interview Survey (NHIS) and uses the stratified, multistage probability sample design of	Estimates are considered statistically unreliable and are suppressed if 1) the denominator is based on fewer than 70 sample cases, or	There is no general de-identification threshold. Small numbers are generally recoded to mask identifiable information.

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
	NHIS. The sample size is about 25,000 individuals.	2) the relative standard error (RSE) of the estimate (expressed as a percentage) is greater than 30 percent. Healthy People 2010 tracking data for American Indians or Alaska Natives and Asians or Pacific Islanders are suppressed because of their small numbers in the survey.	All person-level income amounts on the file, including both total income and the separate sources of income, are top-coded to preserve confidentiality. For each income source, top codes are applied to the top. Medical Conditions Rules: To ensure confidentiality, age of diagnosis was top-coded to 85. For confidentiality reasons, AGEDIAG is set to Inapplicable (-1) for cancer conditions. In order to preserve confidentiality, all of the conditions provided on this file have been collapsed to 3-digit diagnosis code categories rather than the fully-specified ICD-10-CM code. For confidentiality purposes, approximately 7% of ICD-10-CM codes were recoded to -15 (Cannot be Computed) for conditions where the frequency was fewer than 40 for the total

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			unweighted population in the file or less than 400,000 for the weighted population. Additional factors used to determine recoding include age and gender. Clinical Classification Software Refined (CCSR) are used alongside ICD-10-CM diagnosis codes to group medical conditions into clinically meaningful categories. For confidentiality purposes, less than two percent of the CCSR categories are collapsed into a broader code for the appropriate body system where the frequency is less than 40 for the total unweighted population in the file or less than 400,000 for the weighted population. Consolidated Data Rules: percentile of all cases (including negative amounts that exceeded

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			income thresholds in absolute value). In cases where less than one percent of all persons received a particular income source, all recipients are top-coded. Top-coded income amounts are masked using a regression-based approach. When missing, values are imputed for certain persons' hourly wages. Hourly wages greater than or equal to $105.77 are top-coded to -10 and the number of employees variable is top-coded at 500. Specific cancer diagnosis variables with a frequency count fewer than 20 and those considered clinically rare (i.e., appear on the National Institutes of Health's list of rare diseases), are removed from the file, and the corresponding variable CAOTHER, indicating diagnosis of a cancer that is not

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			counted individually, is recoded to "Yes" (1) as necessary. The age of diagnosis for arthritis (ARTHAGED) is included in this file and may be recoded in some cases to "Cannot be Computed" (-15). This variable is top-coded to 85 years of age. The annual Disability Days variables, which represent the number of days a person missed work (DDNWRK21 and OTHNDD21), are top-coded to mask values that exceed the top one-half of one percent of the population. NUMEMP indicates the number of employees at the location of the person's current main job. This variable is top-coded at 500 or more employees. Current main jobs are initially coded at the 4-digit level for both

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			industry and occupation. Then, these codes are condensed into broader groups for release on the file. INDCAT31, INDCAT42, and INDCAT53 represent the condensed industry codes for a person's current main job at the interview date. OCCCAT31, OCCCAT42, and OCCCAT53 represent the condensed occupation codes for a person's current main job at the interview date. … bottom coded to a value of '1951' to preserve age confidentiality. https://meps.ahrq.gov/data_stats/download_data/pufs/h233/h233doc.pdf https://meps.ahrq.gov/data_stats/download_data/pufs/h231/h231doc.pdf

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
Monitoring the Future (MTF)	The MTF study uses self-administered questionnaires in annual national surveys of representative samples of 8th, 10th, and 12th graders in public and private schools in the continental United States. The sample size is about 45,000–50,000 students.	Data may be considered statistically unreliable when there are fewer than 100 cases in the denominator or if data values are less than 0.05 percent. For all objectives tracked by this survey, data for American Indians or Alaska Natives and Asians are statistically unreliable and are suppressed. MTF provides threshold information for the restricted data available through NAHDAP (National Addiction & HIV Data Archive Program). All analyses, output, tables, and figures are examined for disclosure risk; any estimate to be reported by a researcher must be based on at least an N=30, especially for subgroups. MTF is in the process of revisiting this threshold of N=30 based on internal work and feedback it is	There is no general threshold. For MTF public use data, MTF routinely does the following to minimize disclosure risk: • reports race/ethnicity only as Black, White, Hispanic, Other • reports respondent's age in terms of number of months • omits or truncates other sensitive questions. Internally, MTF uses its best judgment of the reliability and replicability of the estimates for any analyses, especially as related to questions new to a survey in a given year. MTF also considers cell size/sample size as it breaks the data into subgroups (sex, race/ethnicity, etc.)

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
		receiving from users of the restricted MTF data.	
National Ambulatory Medical Care Survey (NAMCS) and National Hospital Ambulatory Medical Care Survey (NHAMCS)	NAMCS and NHAMCS are the National Center for Health Statistics' (NCHS) ambulatory healthcare surveys. The National Center for Health Statistics is part of the Centers for Disease Control and Prevention (CDC), Department of Health and Human Services. NCHS collects, analyzes, and disseminates timely, relevant, and accurate health data and statistics. The annual NAMCS and NHAMCS use multistage probability sample designs. NAMCS collects data from over 3,000 non-federally employed office-based physicians and NHAMCS	Per the Healthy People 2010 Criteria for Data Suppression, estimates are considered statistically unreliable if: 1) the numerator is less than 30, or 2) the RSE of the estimate is greater than 30 percent. For all objectives tracked by these surveys, data for American Indians or Alaska Natives, Asians or Pacific Islanders, and Hispanics are suppressed. For American Indians or Alaska Natives and Asians or Pacific Islanders, the number of visits is too small. For the Hispanic origin variable, item nonresponse is too high. Also, see the examples of the application of NCHS data presentation standards for	NCHS follows standard statistical disclosure limitation (SDL) methods when they create public-use files. Their SDL threshold rule (Page 15) indicates that a cell in a table of frequencies is defined as sensitive if the number of respondents is less than some specified number. Some agencies require at least five (5) respondents in a cell, while others require three (3). Under certain circumstances, the number may be much larger. The choice of the minimum number is generally made in consideration of: (a) the sensitivity of the information that the agency is considering to publish, (b) the amount of protection the agency determines to be

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
	collects data on office visits from about 400 emergency departments and about 230 outpatient departments on samples of their patient visits during an assigned reporting period.	proportions for the malignant neoplasms of colon and rectum visits.	necessary given the degree of precision required to achieve disclosure. A more recent publication from the National Institute of Standards and Technology (NIST) explains cells in contingency tables with counts lower than a predefined threshold can be suppressed to prevent the identification of attribute combinations with small numbers. NIST referenced the State of Washington's small numbers standards. Accordingly, department staff who are preparing confidential data for public presentation must: 1. Suppress all non-zero counts which are less than ten, unless they are in a category labeled “unknown.”

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			2. Suppress rates or proportions derived from those suppressed counts. 3. Use secondary suppression as needed to assure that suppressed cells cannot be recalculated through subtraction. 4. When possible, aggregate data to minimize the need for suppression. 5. Individuals at the high or low end of a distribution (e.g., people with extremely high incomes, very old individuals, or people with extremely high body mass indexes) might be more identifiable than those in the middle. If needed, analysts need to top- or bottom-code the highest and lowest categories within a distribution to protect confidentiality.

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			Note: Due to confidentiality reasons, NCHS cannot share specific information on their cell size suppression criteria as their survey data. For their vital statistics data, the cell suppression rule is to suppress cell sizes less than 10 at the subnational level. For more information, please follow the links below: https://www.nber.org/research/data/national-hospital-ambulatory-medical-care-survey https://www.cdc.gov/nchs/about/organization.html https://www.fcsm.gov/assets/files/docs/spwp22WithFrontNote.pdf https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-188.pdf

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			https://doh.wa.gov/sites/default/files/legacy/Documents/1500//SmallNumbers.pdf
National Crime Victimization Survey (NCVS)	NCVS is an annual, nationally representative survey of the civilian, noninstitutionalized population ages 12 years and over. • A stratified, multistage cluster sampling strategy is used to select approximately 50,000 households for a series of telephone and in-person interviews. •The survey obtains information on the frequency, nature, and consequences of criminal victimizations (including those not reported to police).	Estimates are considered statistically unreliable and are suppressed if they are based on 10 or fewer sample cases in the numerator.	The Bureau of Justice Statistics (BJS) sponsors the NCVS, and the U.S. Census Bureau collects the data and completes all data processing. Per the BJS: 1. The data confidentiality requirements are defined by the Census Bureau's Disclosure Review Board. The detailed guidelines/thresholds for the NCVS are not publicly available. 2. But the Census Bureau cell size threshold often requires a minimum unweighted count to be at least three (3) for each cell. 3. Counts of zero are not considered disclosures, but very small counts are.

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			4. Cell suppression may be used for frequency count data. Utilizing Census Bureau software, sensitive cells (those with small unweighted counts) are recognized and then suppressed (the estimate is replaced with the letter ”D”) from the published data. For more information, please follow the links below: https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-McKennaHaubach-Legacy%20Techniques.pdf https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/FSRDC-Disclosure-Avoidance-Methods-Handbook.pdf
National Health Interview	The National Health Interview Survey (NHIS) is the oldest ongoing national household health survey in	According to the Health People 2010 Criteria for Data Suppression, estimates are considered	The small cell size suppression policy will follow the NCHS guidelines as stated at the NAMCS and NHAMCS section

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
Survey (NHIS)	the United States. The survey is conducted by the NCHS, which is part of the CDC. NHIS uses a stratified, multistage probability sample design. It collects data annually on the civilian, non-institutionalized population by computer-assisted personal interview. The expected sample of 43,000 occupied respondent households yields a probability sample of about 111,000 persons.	statistically unreliable and are suppressed if: 1) the denominator is based on fewer than 50 sample cases, or 2) the RSE of the estimate is greater than 30 percent. However, the examples of the application of NCHS data presentation standards for proportions use slightly different criteria.	above. For more information, please follow the link below: https://www.cdc.gov/nchs/nhis-participants/what-to-expect/index.html
National Health and Nutrition Examination Survey (NHANES) & Continuing Survey of	NHANES is an annual, nationally representative examination survey of the U.S. civilian, noninstitutionalized population. A stratified, multistage probability sampling scheme is used to	Data from NHANES may be considered statistically unreliable for two reasons: sampling design and/or small sample size. -Estimates for Healthy People 2010 objectives that are based on fewer than 30 sample events in the	The National Health and Nutrition Examination Survey (NHANES), now combined with the Continuing Survey of Food Intake by Individuals (CSFII), is a major program of the NCHS. NCHS is part of the CDC and has the responsibility for producing vital

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
Food Intake by Individuals (CSFII)	select approximately 5,000 persons for personal interview and examination. CSFII is a nationally representative periodic survey, which uses a stratified multistage probability sample of the U.S. noninstitutionalized civilian population. It includes the collection of data on the kinds and amounts of foods consumed on each of two nonconsecutive days, sources of foods, time, and name of each eating occasion. CSFII has been integrated with the National Health and Nutrition Examination Survey (NHANES); dietary data collection for the integrated	denominator or that have a RSE greater than 30 percent are also considered statistically unreliable and are suppressed. Other sources may use different criteria to assess the statistical reliability of NHANES data. -The NHANES sample size and design does not support producing estimates for certain racial and ethnic groups, including American Indians or Alaska Natives and Asians and Pacific Islanders. Healthy People 2010 tracking data for these populations are considered statistically unreliable and are suppressed. https://www.cdc.gov/nchs/data/statnt/statnt24.pdf Per CSFII estimates between 25 percent and 75 percent are considered unreliable and are suppressed if the coefficient of	and health statistics for the Nation. Therefore, the small cell size suppression policy will follow the NCHS guidelines as stated at the NAMCS and NHAMCS section above. For more information, please follow the links below and under the NAMCS and NHAMCS: https://www.cdc.gov/nchs/nhanes/about/

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
	survey began in 2002. For 1994–96, the sample size was about 16,000 individuals of all ages, and for 1998, an additional sample was collected from about 5,500 children aged 0–9 years. In 2002, the Continuing Survey of Food Intakes by Individuals and the NHANES dietary component were merged, forming a consolidated dietary data collection. For more information please follow the link below: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4717880/	variation of the mean (CV) is greater than 30 percent or the sample size is less than 30 times an average or generalized design effect. A variance inflation factor (VIF), equal to the ratio of the mean of the squared sampling weights to the square of the mean of the weights, is used in the generalized design. Estimates of 25 percent or less are considered unreliable and are suppressed if the sample size is less than eight times a generalized design effect divided by p, where p is the proportion expressed as a fraction. - Estimates of 75 percent or greater are considered unreliable and are suppressed if the sample size is less than eight times a generalized design effect divided by (1–p). -Data for American Indians or Alaska Natives and Asians or Pacific	

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
		Islanders are suppressed because the size of the sampled populations for these groups is too small.	
National Hospital Discharge Survey (NHDS)	NHDS is an annual survey that collects data from medical records to provide national estimates on hospital discharges from short-stay, noninstitutional hospitals and general and children's general hospitals regardless of length of stay. Annually, the national estimate is based on a sample of about 300,000 records.	Population based rates are considered unreliable and are suppressed if the numerator is based on fewer than 30 records or if they have a RSE greater than 30 percent. Estimates based on 30–59 patient records are flagged to indicate they also have low reliability. For all objectives tracked by this survey, data for American Indians or Alaska Natives and Asians or Pacific Islanders are statistically unreliable because of their small numbers in the survey. Data for Hispanics are suppressed because of high item nonresponse to the Hispanic origin variable.	The National Hospital Care Survey (NHCS) is a new survey that integrates inpatient data formerly collected by the NHDS with the emergency department (ED), outpatient department (OPD), and ambulatory surgery center (ASC) data collected by the National Hospital Ambulatory Medical Care Survey (NHAMCS). It is being collected by the CDC's NCHS. The small cell size suppression policy will follow the NCHS guidelines as stated at the NAMCS and NHAMCS section above. For more information, please follow the links below and under the NAMCS and NHAMCS section:

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			https://www.cdc.gov/rdc/restricted-nchs-variables/nhds.html
National Household Survey on Drug Abuse (NHSDA)	NHSDA is an annual, multistage national probability sample survey of the civilian, noninstitutionalized population (ages 12 years and over). In 1999, the sample size increased from about 25,000 to about 70,000 persons.	Estimated proportions are considered statistically unreliable and are suppressed if their RSE is greater than 17.5 percent. Estimated proportions are also considered statistically unreliable if p < 0.0005 or p ≥ 0.99995.	Data collection progress of the NHSDA was monitored during each quarterly survey by state. Small reserve samples were held back each quarter so that the assigned sample size could be adjusted if necessary during the course of data collection. To protect the confidentiality of respondents, for example, in the 2019 National Survey on Drug Use and Health (NSDUH), the full analytic file of the individuals was treated using a statistical disclosure limitation method called MASSC, which consists of the following four major steps: Micro Agglomeration, optimal probabilistic Substitution, optimal probabilistic Subsampling, and

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			optimal sampling weight Calibration. -All directly identifying information (such as name, phone number, and address) on the file was eliminated. In addition, census region, state, and other geographic identifiers were removed. Moreover, the household link between respondents from the same household was not included in the public use file. Substance Abuse and Mental Health Services Administration (SAHMSA) reported that they are unable to provide detailed information on their disclosure avoidance techniques apart from that in their public-facing documentation. But according to their suppression criteria, for confidentiality protection, survey

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
			sample sizes greater than 100 were rounded to the nearest ten (10), and sample sizes less than 100 were not reported (i.e., are shown as "<100" in tables). For more information, please follow the links below: https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH-2019/NSDUH-2019-datasets/NSDUH-2019-DS0001/NSDUH-2019-DS0001-info/NSDUH-2019-DS0001-info-codebook.pdf https://www.samhsa.gov/data/sites/default/files/reports/rpt41913/2021NSDUHmrbStatInference.pdf
National Survey of Family Growth (NSFG)	NSFG is a periodic survey based on a multistage probability design. It collects data on civilian, noninstitutionalized females	The Healthy People 2010 Criteria for Data Suppression describes data as statistically unreliable when RSE is greater than 30 percent or the	NSFG is conducted by the CDC NCHS with the support and assistance of a number of other programs and agencies within the

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
	15–44 years of age by computer-assisted personal interview with a self-administered audio section for more sensitive topics. The sample size was 13,795 females in 1995. In 2002 civilian, noninstitutionalized males 15–44 years of age were added to the sample.	denominator is based on fewer than 50 sample cases. For all objectives tracked by this survey, data for American Indians or Alaska Natives and Asians or Pacific Islanders are statistically unreliable and are suppressed because the size of the sampled populations for these groups is too small. The NCHS Data Presentation Standards for Proportions uses different application examples (see footnote 78).	U.S. Department of Health and Human Services (HHS). The small cell size suppression policy will follow the NCHS guidelines as stated at the NAMCS and NHAMCS section above. For more information, please follow the links below and under the NAMCS and NHAMCS: https://www.cdc.gov/nchs/nsfg/about_nsfg.htm#:~:text=NSFG%20is%20conducted%20by%20the,and%20Human%20Services%20(HHS)
School Health Policies and Programs Study (SHPPS)	The periodic study conducted every 6 years consists of a census of all State education agencies; a national probability sample of public and private school districts; a national sample of public and private elementary, middle/junior	Data based on fewer than 30 schools in the denominator are considered statistically unreliable and are suppressed.	The School Health Policies and Practices Study (SHPPS) is a national survey periodically conducted by the CDC to assess school health policies and practices at the state, district, school, and classroom levels every 6 years.

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
	high, and senior high schools; and a random sample of required health education and physical education classes.		The small cell size suppression policy will follow the NCHS guidelines as stated at the NAMCS and NHAMCS section above. For more information, please follow the links below and under the NAMCS and NHAMCS: SHPPS 2016 "Results from the School Health Policies and Practices Study" was originally retrieved December 2024 from the CDC website (removed in 2025), Copy is available through the National Coalition of STD Directors website: https://ncsddc.org/wp-content/uploads/2017/09/shpps-results_2016.pdf :
Youth Risk Behavior Surveillance System (YRBSS)	The national Youth Risk Behavior Survey (YRBS) is a part of the YRBSS and is a biennial, school-based survey administered to	Data based on a denominator of fewer than 100 students are considered statistically unreliable and are suppressed.	The CDC's Division of Adolescent and School Health (DASH) routinely monitors youth health behaviors and experiences.

Survey Name	Brief Details	Statistical Unreliability Rules	Data De-Identification/ Protection Rules
	samples of students in grades 9–12. In 1999 about 15,300 students participated.	The national YRBS does not sample enough Asian or Pacific Islander or American Indian or Alaska Native adolescents from a single data year to present estimates for these race groups.	The small cell size suppression policy will follow the NCHS guidelines as stated at the NAMCS and NHAMCS section above. For more information, please follow the links below and under the NAMCS and NHAMCS: Survey documentation was originally retrieved December 2024 from the CDC website: https://www.cdc.gov/healthyyouth/data/yrbs/pdf/trendsreport.pdf

21) APPENDIX I: SCORING SCENARIOS FOR PUBLIC HEALTH CARE COVERAGE

This section addresses scoring scenarios involving the overlap of Insurance Coverage, Expected Payer/Public Assistance and Means-Tested Programs, and Geography. Below are three key points that summarize all the scenarios:

1) If the data is ONLY related to Residence or Service Geography, then DO NOT USE Insurance Coverage or Means-Tested Tables.
2) Means-Tested Programs—Only add interaction if enrollment in the Public Assistance program is 10 million or fewer people. No interaction is needed for Medi-Cal as the current enrollment is approximately 14 million, which exceeds 10 million.
3) If the number of members enrolled in Insurance Coverage is less than the population of the geographic subdivision, then use the Insurance Table. If the number of members enrolled in Insurance Coverage is greater than or equal to the population of the geographic subdivision, then use the Geography Table.

Table 36: Scoring Scenarios for Insurance Coverage and Geography Data that Includes Public Programs

In the table below, the examples assume annual data (+0 score) with under 11 events (+7 score) along with other dataset characteristics described under the Dataset Example column. A Yes entry indicates that the applicable score for that variable should be added in the Total Risk Score.

Dataset Example: Annual Data (+0) <11 events (+7)	Insurance Coverage	Expected Payer/ Public Assistance and Means-Tested Programs	Geography	Interaction	Total Risk Score
Events, Statewide data					**Safe Harbor if Statewide and Annual data**
Events, County-level data (Includes regional and other sub-state geographies)			Yes		**Base score*: +7** *Examples based on county population:* • *Alpine (population <4,000): 7+7=14* • *Pop size of 10,000: 7+5=12* • *Pop size of 30k: 7+4=11*

Dataset Example: **Annual Data (+0)** **<11 events (+7)**	**Insurance Coverage**	**Expected Payer/ Public Assistance and Means-Tested Programs**	**Geography**	**Interaction**	**Total Risk Score**
Events, Means-Tested Programs, Statewide		Yes	Yes	+1, if enrollment in Means-Tested Programs is less than 10 million people	**Base score*: +7** *Examples based on public assistance program size:* • *Medi-Cal (~14 million) = 7-5+0+0= 2* • *Self-Pay/Uninsured (~2.7 million) = 7-5+2+1=5* • *CalFresh (5.5 million) = 7-5+1+1=4* • *WIC (1 million) = 7-5+2+1=5* • *CalWORKS, HSP (13,000) = 7-5+7+1= 10*
Events, Means-Tested Programs, County-level data (Includes regional		Yes	Yes *Example:* *Alpine- 1,163 (+7)*	+1, if enrollment in Means-Tested Programs is	**Base score*: +7** *Examples based on public assistance program size:* • *Medi-Cal (~14 million)= 7+7+0+0= 14*

Dataset Example: Annual Data (+0) <11 events (+7)	Insurance Coverage	Expected Payer/ Public Assistance and Means-Tested Programs	Geography	Interaction	Total Risk Score
and other sub-state geographies)				less than 10 million people	- *Self-Pay/Uninsured (~ 2.7 million) = 7+7+2+1=17* - *CalFresh (5.5 million) = 7+7+1+1=16* - *WIC (1 million) = 7+7+2+1=17* - *CalWORKS, HSP (13,000) = 7+7+7+1= 22*
Health plans / Insurers	Yes				**Base score*: +7** *Examples based on Insurance Coverage enrollment:* - *Enrollment <4,000): 7+5=12* - *Enrollment of 10,000: 7+5=12* - *Enrollment of 30k: 7+4=11*

Dataset Example: Annual Data (+0) <11 events (+7)	Insurance Coverage	Expected Payer/ Public Assistance and Means-Tested Programs	Geography	Interaction	Total Risk Score
Health plans / Insurers, **including Medi-Cal,** and no specific geographic coverage and no other public coverage	Yes	+0, since enrollment in Medi-Cal is approx.14 million people		+0, since enrollment in Medi-Cal is approx.14 million people	**Base score*: 7+0+0 = +7** *Examples based on Insurance Coverage enrollment:* • *Enrollment <4,000): 7+5 =12* • *Enrollment of 10,000: 7+5=12* • *Enrollment of 30k: 7+4=11*
Health plans / Insurers, **including Medi-Cal**, and no other public coverage, County-level data (Includes regional and other sub-state geographies)	Yes, if the Insurance Coverage population is less than the geography population	+0, since enrollment in Medi-Cal is approx.14 million people	Yes, if the Insurance Coverage population greater than or equal to the geography population	+0	**Base score*: 7+0+0 = 7** If Ins Cov. Population < Geography Population, then use Ins Cov. Table: *Examples based on Insurance Coverage enrollment:* • *Enrollment <4,000):7+5=12*

Dataset Example: Annual Data (+0) <11 events (+7)	Insurance Coverage	Expected Payer/ Public Assistance and Means-Tested Programs	Geography	Interaction	Total Risk Score
					• *Enrollment of 10,000: 7+5=12* • *Enrollment of 30k: 7+4=11* Else, use Geography Table *Examples based on Residence Geography:* • *Population <4,000): 7+7=14* • *Population of 10,000: 7+5=12* • *Population of 30k: 7+4=11*
Health plans / Insurers, Means-Tested Programs with less than 10 million enrolled	Yes	Yes *Example: WIC (1 million) (+2)*		+1 enrollment in Means-Tested Program is	**Base score*: 7+2+1 = +10** *Examples based on Insurance Coverage enrollment:* • *Enrollment of 10,000: 10+5=15*

Dataset Example: Annual Data (+0) <11 events (+7)	Insurance Coverage	Expected Payer/ Public Assistance and Means-Tested Programs	Geography	Interaction	Total Risk Score
				less than 10 million people	- *Enrollment of 30k: 10+4=14* - *Enrollment of 150k: 10+1=11*
Health plans / Insurers, Means-Tested Programs with less than 10 million enrolled, County-level data (Includes regional and other sub-state geographies)	Yes, if the Insurance Coverage population is less than the geography population	Yes *Example: WIC (1 million)* *(+2)*	Yes, if the Insurance Coverage population greater than or equal to the geography population	+1 enrollment in Means-Tested Program is less than 10 million people	**Base score*: 7+2+1 = +10** If Ins Cov. < Geography, then use Ins Cov. Table: *Examples based on Insurance Coverage enrollment:* - *Enrollment < 10,000: 10+5=15* - *Enrollment of 30k: 10+4=14* - *Enrollment of 150k: 10+1=11* Else, use Geography Table

Dataset Example: **Annual Data (+0)** **<11 events (+7)**	**Insurance Coverage**	**Expected Payer/ Public Assistance and Means-Tested Programs**	**Geography**	**Interaction**	**Total Risk Score**
					Examples based on Residence Geography: - *Population <4,000: 10+7=17* - *Population of 10,000: 10+5=15* - *Population of 30k: 10+4=14* - *Population of 150k: 10+1=11*

* Base Score includes reporting time (annual), events, Expected Payer / Public Assistance or Means-Tested programs, and interaction.

22) APPENDIX J: CALHHS DDG TEMPLATE DEVELOPMENT AND REVISION HISTORY

This section outlines the development, review, and approval stages for the CalHHS DDG Template Editions 1.0 and 2.0.

In 2015, the CalHHS Data Subcommittee requested the convening of the CalHHS Data De-Identification Workgroup to develop the DDG.

Edition	Date	Author/Lead	Brief Description of Development Activity or Edition Change
Pre-draft	3/15/15	L. Scott	Planning Meeting Part 1 – Participants included DHCS, CDPH, OSHPD, OHII
Pre-draft	3/20/15	L. Scott	Planning Meeting Part 2 - Participants included DHCS, CDPH, OSHPD, OHII
Pre-draft	4/7/15	L. Scott	Present Objectives for the project and use the DHCS PAR-DBR as an example
Pre-draft	4/23/15	L. Scott	Presentations from OSHPD and CDPH regarding current processes and approach to small cell sizes
Pre-draft	5/5/15	L. Scott	Discuss concept of uniqueness as a way to measure risk for re-identification and gather input from Departments/Offices regarding DDG variables and topics
0.1	5/26/15	L. Scott	Initial draft for review which was based on the DHCS PAR-DBR Guidelines dated 8/25/14 and conversations at the CalHHS Data De-Identification Workgroup meetings
0.1	5/27/15	L. Scott	Review initial draft DDG – Focus on new sections of the document
0.1	6/8/15	L. Scott	Review initial draft DDG – Focus on Data Assessment for Public Release Procedure
0.1	May-June 2015	L. Scott	Meet with each department/office individually

Edition	Date	Author/Lead	Brief Description of Development Activity or Edition Change
0.2	6/29/15	L. Scott	Additions made based on feedback: • CalHHS Data De-Identification Workgroup meetings on May 27, 2015, and June 8, 2015 • Department-specific meetings
0.2	6/30/15	L. Scott	Review draft DDG 0.2
0.2	July 2015	L. Scott	Departments/offices vet the DDG within their departments/offices
0.3	8/5/15	L. Scott	Additions and changes based on feedback from all departments with specific written comments from CDPH, OSHPD, DCSS, CDSS, MHSOAC
0.3	8/6/15	L. Scott	Review draft DDG 0.3
0.3	8/21/15	L. Scott	Received input from the CalHHS Risk Management Committee
0.3	9/14/15	L. Scott	Progress update for DDG Workgroup and discussion of additional topics
0.3	12/18/15	L. Scott	Presentation from NORC to review their findings of the draft DDG
0.3	1/8/16	L. Scott	Receive final recommendations from NORC
0.4	1/22/16	L. Scott	Revisions based on recommendations from: • NORC • CalHHS DDG Workgroup • CalHHS Risk Management Subcommittee and associated Legal and Privacy Workgroup Specific written comments from CDPH, CDSS
0.4	Jan. 2016	L. Scott	Provide DDG 0.4 to DDG Workgroup
0.4	2/18/16	L. Scott	Review and discussion of draft DDG 0.4 with the DDG Workgroup
0.5	3/18/16	L. Scott	Revisions based on comments from CDPH, CDSS, OSHPD, DHCS

Edition	Date	Author/Lead	Brief Description of Development Activity or Edition Change
0.5	3/18/16	L. Scott	Provide DDG 0.5 with outstanding comments from the DDG Workgroup to the Data Subcommittee
0.6	4/4/16	L. Scott	Revisions based on feedback from and discussion with the Data Subcommittee
0.6	4/18/16	L. Scott	Provide revised draft DDG to the Data Subcommittee
0.7	5/3/16	L. Scott	Revisions based on feedback from and discussion with the Data Subcommittee
0.7	5/24/16	L. Scott	Provide draft DDG 0.7 from the CalHHS Data Subcommittee to the CalHHS Governance Advisory Council. The Advisory Council shared DDG 0.7 with the other subcommittees and discussed DDG 0.7 at the 6/8/16 meeting.
0.8	6/17/16	L. Scott	Revisions based on direction from the CalHHS Governance Advisory Council and input from the CalHHS Risk Management Committee
0.8	7/6/16	L. Scott	CalHHS Governance Advisory Council discussed DDG 0.8 at its meeting
0.9	7/6/16	P. Cervinka	Revisions based on clarification from the CalHHS Governance Advisory Council
0.10	7/7/16	L. Scott	Provide draft DDG 0.10 to the CalHHS Undersecretary
0.10.1	7/11/16	L. Scott	Formatting and citations edits to be consistent with previous DDG 0.8
1.0	9/23/16	L. Scott	Revisions based on direction from the CalHHS Undersecretary. Approved as Edition 1.0 for implementation.
1.0	5/24/23	L. Scott	DDG Template Workgroup convenes a Scoring Criteria subgroup to begin DDG updates

Edition	Date	Author/Lead	Brief Description of Development Activity or Edition Change
1.0.1	May - July 2023	D. Aggarwal	DDG Template Workgroup – Scoring Criteria subgroup conducts a literature search to address scoring update topics and holds weekly meetings to develop proposals
1.0.2	7/19/23	D. Aggarwal	Scoring Criteria subgroup scoring proposals and language revisions prepared for presentation to the DDG Template Workgroup
1.1	7/20/23	D. Aggarwal	Revisions based on DDG Template Workgroup input on proposals developed by the Scoring Criteria subgroup during weekly meetings from May to July 2023
1.1	August - September 2023	D. Aggarwal	Weekly scoring criteria subgroup meetings to develop proposals based on DDG Template Workgroup feedback
1.2	9/21/23	D. Aggarwal	Revisions based on DDG Template Workgroup input on proposals developed by a Scoring Criteria subgroup during weekly meetings, including to address topic High-Risk Populations
1.2	October 2023	D. Aggarwal	Scoring Criteria subgroup meets to prepare proposed updates for review by the DDG Template Workgroup
1.3	10/19/23	D. Aggarwal	Revisions based on DDG Template Workgroup input on proposals developed by a Scoring Criteria subgroup during weekly meetings, including to address topic Language Spoken
1.3	November-December 2023	D. Aggarwal	Scoring Criteria subgroup works on language and scoring updates for additional topics based on feedback from the DDG Template Workgroup
1.4	12/21/23	D. Aggarwal	Revisions based on DDG Template Workgroup input on proposals developed by the Scoring Criteria subgroup, including to address topics Age Range, Assess Potential Risk, and Immigration Status

Edition	Date	Author/Lead	Brief Description of Development Activity or Edition Change
1.4	January-October 2024	D. Aggarwal	Scoring Criteria subgroup works on language and scoring updates based on feedback from the DDG Template Workgroup
1.5	12/19/24	D. Aggarwal	Revisions based on DDG Template Workgroup input on proposed updates developed by the Scoring Criteria subgroup. The Workgroup reviewed suggested revisions in November and December for topics including Artificial Intelligence, Data with More Specificity, Expected Payer, Geography, High-Risk Populations, Other Variables, Race/Ethnicity details based on new OMB/AB 91 standards, Risk with Increased Granularity, SOGI, Special Scenarios, Suppression Rules, Survey Data, and Statistical Masking. The Workgroup approved the DDG draft.
1.5	1/23/25	D. Aggarwal	Draft DDG 1.5 presented to the Peer Review Team
1.5	February-March 2023	D. Aggarwal	DDG Template Workgroup - Scoring Criteria subgroup develops updates based on Peer Review Team feedback
1.5	4/14/25	D. Aggarwal	DDG Template Workgroup reviews Scoring Criteria subgroup proposed updates
1.6	4/18/25	D. Aggarwal	DDG Template Workgroup approves Scoring Criteria subgroup's revisions based on Peer Review Team feedback and discussion
1.6	4/24/25	D. Aggarwal	DDG 1.6 updates presented to the Peer Review Team
1.6	5/9/25	D. Aggarwal	Peer Review Team approves DDG 1.6
1.6	May 2025	D. Aggarwal	DDG 1.6 updates presented to the Data, Risk Management, and JEDI Subcommittees
1.6	June-August 2025	D. Aggarwal	DDG Template Workgroup addresses feedback from the Data, Risk Management, and JEDI Subcommittees and develops updates

Edition	Date	Author/Lead	Brief Description of Development Activity or Edition Change
1.7	9/8/25	D. Aggarwal	Revisions based on feedback from the Data, Risk Management, and JEDI Subcommittees, including addressing the topics: Insurance Coverage and Expected Payer/ Public Assistance and Means-Tested Programs
1.7	September 2025	D. Aggarwal	Peer Review Team reviews updates based on the Subcommittees' feedback and approved DDG 1.7 in early October
1.8	10/3/25	J. Schwartz/ A. Rykaczewska	Revisions based on feedback from the Agency General Counsel and Chief Equity Officer
1.8.1	10/9/25	D. Aggarwal	Addition of Open Data Portal resource link for population tables
1.9	10/16/25	D. Aggarwal	CalHHS Interdepartmental Advisory Council (IDAC) reviewed and approved
2.0	10/24/25	D. Aggarwal	CalHHS Undersecretary reviewed and approved Edition 2.0 for implementation