Data De-Identification Guidelines (DDG)

California Health and Human Services

September 23, 2016

Version 1.0

Revision History

Version	Date	Author	Brief Description of Changes
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 CHHS Data De-identification Workgroup meetings.
0.2	6/29/15	L. Scott	Additions made based on feedback: • CHHS Data De-identification Workgroup meetings on May 27, 2015 and June 8, 2015 • Department specific meetings
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.
0.4	1/22/16	L. Scott	Revisions based on recommendations from: • NORC • CHHS DDG Workgroup • CHHS Risk Management Subcommittee and associated Legal and Privacy Workgroup • Specific written comments from CDPH, CDSS
0.5	3/18/16	L. Scott	Revisions based on comments from CDPH, CDSS, OSHPD, DHCS.
0.6	4/4/16	L. Scott	Revisions based on feedback from and discussion with the Data Subcommittee
0.7	5/3/16	L. Scott	Revisions based on feedback from and discussion with the Data Subcommittee
0.8	6/17/16	L. Scott	Revisions based on direction from the CHHS Governance Advisory Council and input from the CHHS Risk Management Committee
0.9	7/5/16	P. Cervinka	Revisions based on clarification from the CHHS Governance Advisory Council
0.10	7/11/16	L. Scott	Formatting and citations edits to be consistent with previous version 0.8
1.0	9/23/16	L. Scott	Revisions based on direction from the CHHS Undersecretary. Approved as Version 1.0 for implementation.

Table of Contents

1) Purpose

The California Health and Human Services Agency (CHHS) Data De-identification Guidelines (DDG) describes a procedure to be used by departments and offices in the CHHS to assess data for public release. 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 DDG support CHHS 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 CHHS initiatives, such as the CHHS Open Data Portal.

2) Background

CHHS 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 CHHS. The Advisory Council consists of representatives of senior leadership from departments and offices in the Agency. There are five subcommittees that report to the Advisory Council, which include the Portfolio, Procurement, Infrastructure, Risk Management and Data Subcommittees. The Data De-identification Workgroup was convened by the Data Subcommittee with representation from all departments and offices in CHHS.

CHHS is engaged in improving transparency and public reporting through the Open Data Portal. As described in the CHHS Open Data Portal Handbook, not all data is suitable for use on the open data portal. 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

[1] Civ. Code § 1789 et seq.
[2] HIPAA Privacy Rule is located at 45 CFR Part 160 and Subparts A and E of Part 164
[3] Gov. Code 6250 et seq.

released if it is restricted due to the HIPAA, state or federal law. Data tables may fall into one of three categories:[4]

- 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. These data will NOT be accessible through the CHHS Open Data Portal.

Data can change from being Level 3 to Level 1 if appropriate de-identification processes are employed. The CHHS 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.

3) Scope

Data de-identification practices will be implemented by each department and office (further referred to as department) in the agency. This DDG is the default policy for CHHS departments. If a CHHS department wants to create a department DDG, it must have appropriate references to departmental processes and the department must file a copy of their DDG with the Office of the Agency Information Officer (OAIO). For example, the Legal Review process and the Departmental Release Procedures for De-Identified Data require additional information to describe these steps within each department. Additionally, a department with programs not covered by HIPAA will not require specific HIPAA references. A department must request DDG consultation from the CHHS peer review team (PRT), described in Section 8: DDG Governance prior to implementation. The PRT is available to review the department's documentation to ensure it is consistent with the principles of the CHHS DDG and meets requirements of the California IPA.

The CHHS DDG 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.

[4] CHHS' Open Data Portal Handbook, Version 2.1, October 2014, Data Levels Decision Tree, pages 91 and 92.

Departments are sometimes asked to release record level 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. However, summary data would include information from Jane Doe combined, or summarized, with data from other individuals. 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)[5] or Protected Health Information (PHI).[6] 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. 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.

CHHS 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 data sets or tables to assess risk for identification of the individuals to which they pertain.

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

[6] "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 (PHI)
Any information that identifies or describes an individual, including but not limited to:[7] • Name • Social security number • Physical description • Home address • Home telephone number • Education • Financial matters • Medical history • Employment history Electronically collected personal information:[8] • 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 • 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.	• 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 • 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

[7] California Civil Code 1798.3 (a)
[8] California Government Code 11015.5 (d) (1)

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 data set, then release of this data results in disclosure of personal information.

Less obvious qualities in data sets 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 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.[9]

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 6.

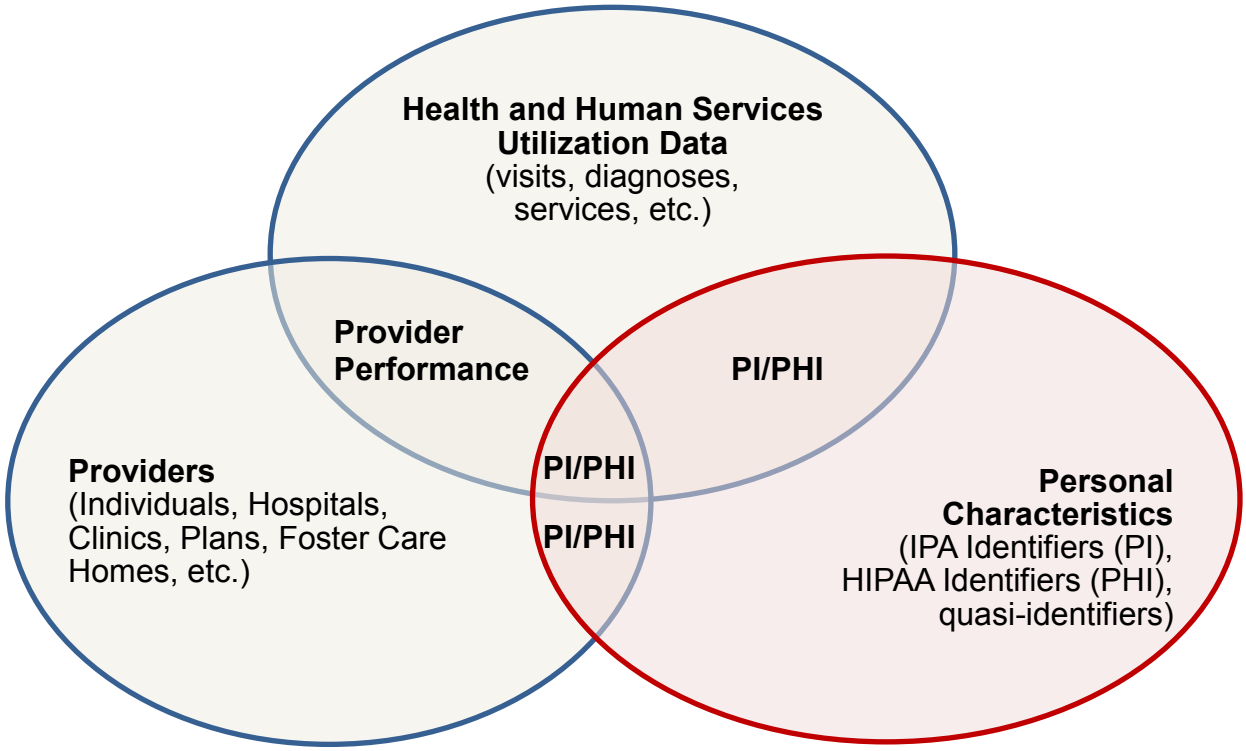
Assessment of risk in the context of maximizing the usefulness of the information presented

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

[9] Introduction to Statistical Disclosure Control, Temple et al. 2014

available characteristics. While patients may come to a provider from anywhere, they 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 puts 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



4) Statistical De-identification

The DDG describes a procedure, the Data Assessment for Public Release Procedure shown in Figure 5, to be used by departments in the CHHS to assess data for public release. This section, section 4, describes specific actions that may be taken for each step in the procedure with additional supporting information being described in sections 5, 6 and 7. 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:

1. Review the data to determine if it includes personal characteristics, directly or indirectly, that can be tied back to an individual;
2. If there is concern for personal characteristics, then assess the data for small numerators or denominators;
3. If there is concern for small numerators or denominators, assess potential risk of data release;
4. If there is potential risk identified, assess the need to apply statistical masking methods to de-identify the data;
5. Following statistical de-identification, the data release is reviewed by legal if indicated in departmental procedures; and,
6. 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.

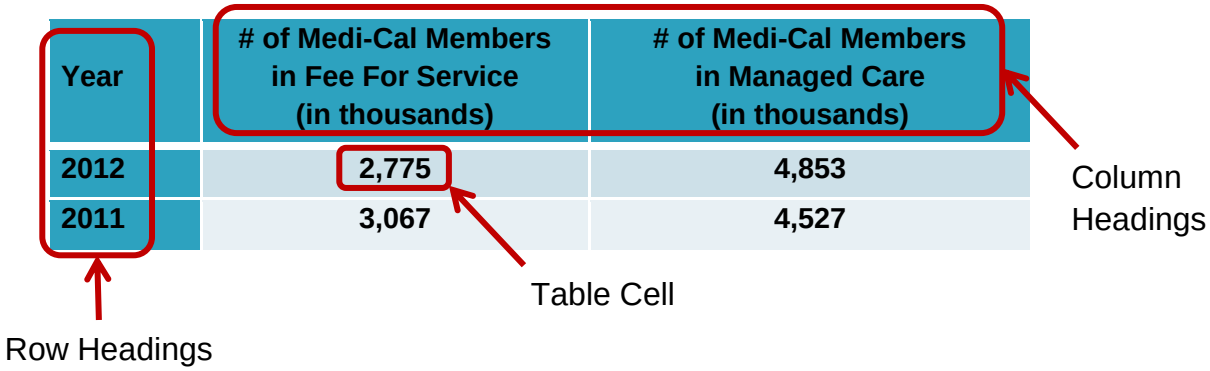
Data summaries that originate from data which includes personal identifiers must be de-identified before release to the public. Additionally, data summaries about conditions experienced by individuals must be adequately de-identified to prevent re-identification of individuals represented by the summarized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

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.



Numerator	# of Medi-Cal Members in Fee For Service (in thousands)	2,775
Denominator	# Medi-Cal Members in 2012 (in thousands)	7,628

Figure 3 shows an example table with the numerator and the denominator 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 2012 and 2011. The column headings are Year, # of Medi-Cal Members in Fee For Service (in thousands) and Number of Medi-Cal Members in Managed Care (in thousands). In Figure 3, "2,775" is the value in a table cell and represents a numerator. The sum of the row for year 2012 (2,775 + 4,853 = 7,628) represents a denominator. In this context, the denominator may represent row totals, column totals or the total occurrences in the data set released. Data in Figure 3 comes from the "Trend in Medi-Cal Program Enrollment by Managed Care Status - for Fiscal Year 2004-2012, 2004-07 - 2012-07." [10]

Figure 4 shows another type of table that contains rates. In this case, the numerator is the number of Salmonella cases for a sample of California Local Health Jurisdictions in 2014. The table also includes the rate of Salmonella for these jurisdictions. In order to

[10] Report Date: July 2013
http://www.dhcs.ca.gov/dataandstats/statistics/Documents/1_6_Annual_Historic_Trend.pdf

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

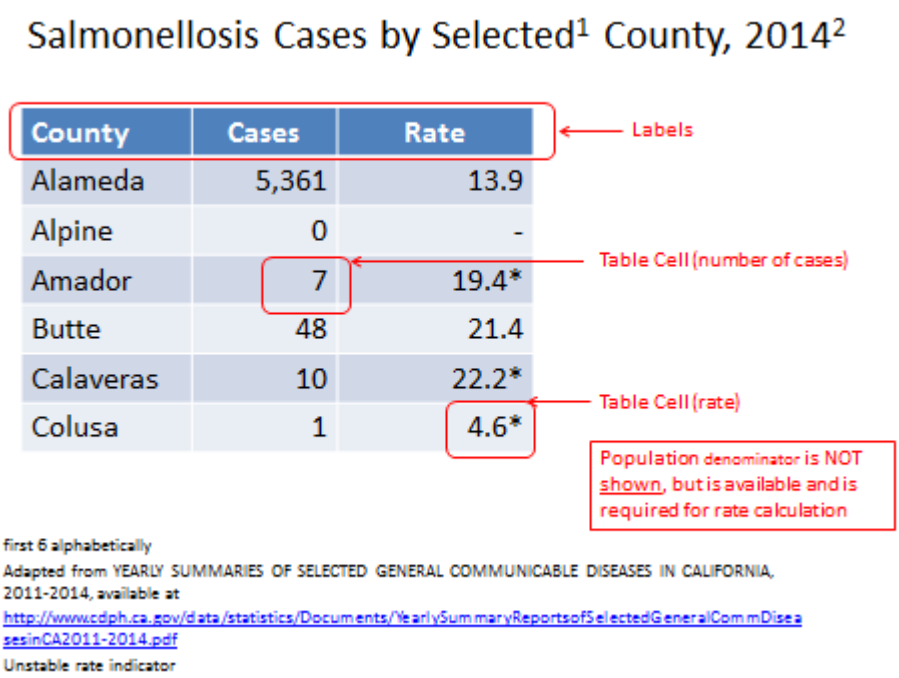
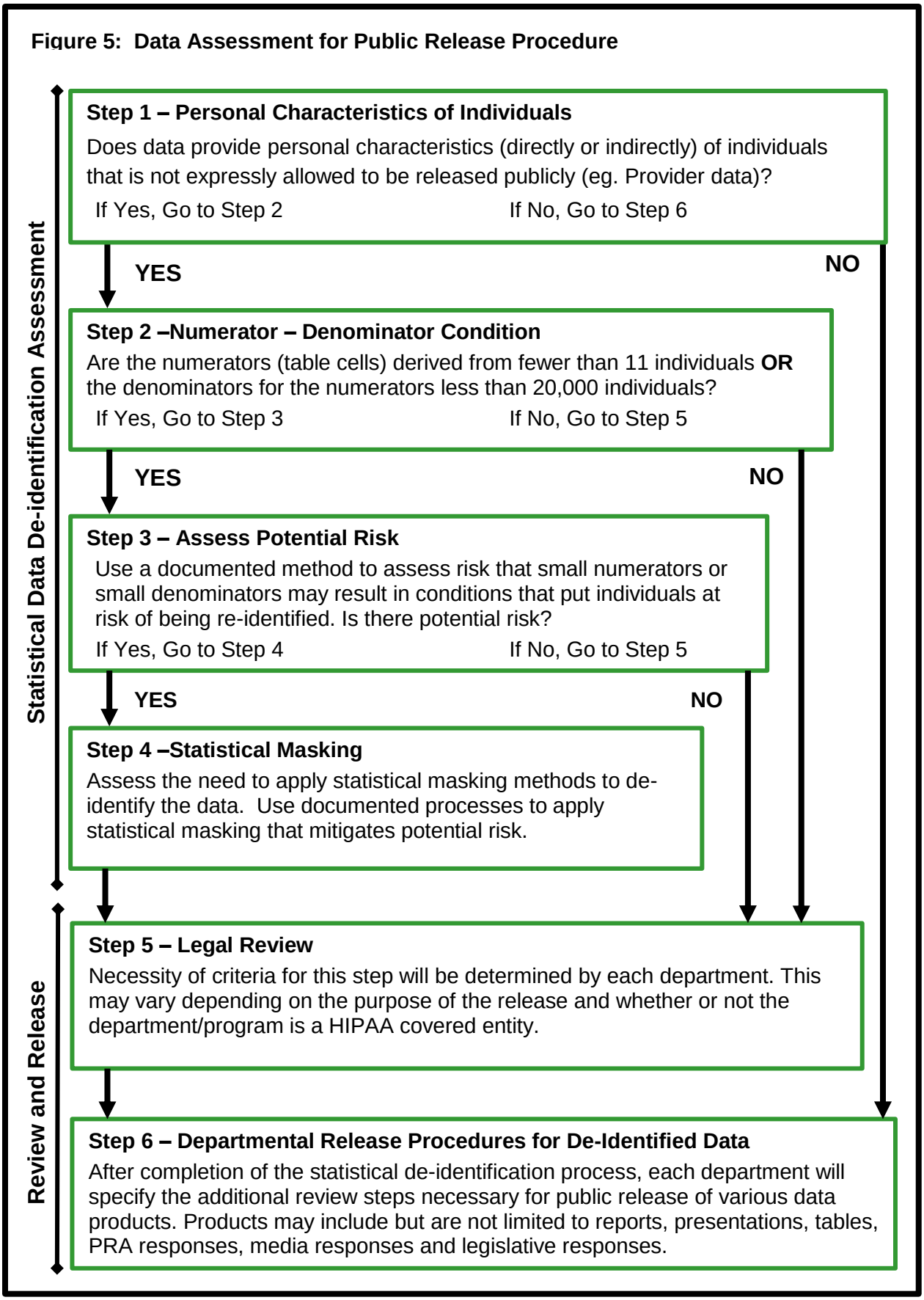


Figure 5: Data Assessment for Public Release Procedure



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 data set. The following are examples of personal characteristics.

- Identifiers as defined in CA IPA
- Identifiers as defined in HIPAA
- Demographics typically reported in census and other reporting
    - Race
    - Ethnicity
    - Language Spoken
    - Sex
    - 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 data set.

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

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 of 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 data set must take into account a range of contributing considerations. This includes understanding particular characteristics of a given data set that is being released. For example, if the potential values for a specific personal characteristic, such as race, results in many small numbers in data set A but does not in data set B, then the risk may be low for data set 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 data sets of interest.

There are many methods 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 15. While each department will document the method(s) chosen for use, the following description of the Publication Scoring Criteria is provided as an example and may be adopted by departments as a method to assess potential risk.

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 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.

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 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.

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 data set 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 6.4, decreasing the granularity or the number of variables are both techniques for increasing the values for the numerators. The final criteria in Figure 6 is that for Variable Interactions. 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. This is discussed in more detail in Section 6.2.

In assessing risk, the 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 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

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

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

4.4 Statistical Masking

If Step 3 determined that the data set has a risk that small numerators or small denominators may result in conditions that put individuals at risk of being re-identified, then the data set must be assessed to determine the need for statistical masking of those small values and complimentary 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. Initial considerations for statistical masking are described below. For additional methods related to statistical masking, please see Section 6.4.

Reduce Table Dimensions

If there are more dimensions present in the table than necessary for the vast majority of analysis, the data producer should consider reducing the number of dimensions in a single table and produce multiple tables each with a subset of the dimensions in the table that resulted in small cells. For example, if there are six dimensions of interest for study, but a table that crosses all six dimensions produces a large number of small cells, the data producer could consider producing several tables each of which crosses four dimensions. This is especially effective if there are very few analytic questions requiring a cross section of all six variables.

Reduce Granularity of Variable(s), aka Recoding or Aggregation

An alternative approach to addressing small cells in a table is to reduce the number of levels of a particular dimension. This is especially useful for dimensions with a large number of levels that can be easily aggregated to fewer levels and maintain much of their utility. Geographic variables such as state or county can often be recoded into regional variables 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 dimensions which have limited opportunities for table dimension reduction.

It should be noted that these actions can be used alone or in tandem to reduce, or completely eliminate, small cells within a table.

Cell Suppression and Complementary Cell Suppression

There will be cases where not all small cells can be eliminated by reducing granularity of dimensions or the number of dimensions present in a table. 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’.
3. 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.

4.5 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 or program is a HIPAA covered entity or not. See Section 7 for further discussion.

4.6 Departmental Release Procedure 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, presentation, tables, PRA responses, media responses and legislative responses. See Section 7 for further discussion.

5) Types of Reporting

CHHS 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 CHHS Open Data Portal, which include:

- Diseases and Conditions
- Facilities and Services
- Healthcare
- Workforce
- Environmental
- 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 list of variables is important to consider when preparing data for release.

Personal characteristics	Event characteristics
Age Sex Race Ethnicity Language Spoken Location of Residence Education Status Financial Status	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 data set, 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 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.

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 the 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), healthcare 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 Study) 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 protection of security and confidentiality. In particular, the main “population denominator” for strict confidentially considerations remains the whole target population, not the sampled population. But, if persons have special or external knowledge of the sampled populations (e.g. that a family member participated in the survey), further considerations may be required. Also, it is in the context of surveys that issues of statistical reliability often arise—which are distinct from confidentially issues, but often arise in related discussions.

Of particular note, small numbers (e.g. less than 11) of individuals reported in surveys do not generally lead to the same security/confidentiality concern as in population-wide

data, and as such should be treated differently than is described within the Publication Scoring Criteria and elsewhere. In this case a level of de-identification occurs based on the sampling methodology itself.

5.3 Budgets and Fiscal Estimates

Budget reporting may include both actuals 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, then the budgets 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, if the actual amounts report caseloads based on personal characteristics, such as age, sex, race or ethnicity, then the budget reporting needs to be assessed for de-identification.

5.4 Facilities, Service Locations and Providers

Many CHHS 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.

- Characteristics of the entity are typically public information, such as location, type of service provided, type of license and the license status.
- 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

CHHS 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.

6) Justification of Thresholds Identified

6.1 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]." [11] 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 unique for over 50% of residents in the United States.[12,13] 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

[11] "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.
[12] 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.
[13] See L. Sweeney. K-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*. 2002; 10(5): 557-570.

combination of Year of Birth, Sex, and 3-Digit ZIP Code is unique for approximately 0.04% of residents in the United States.[14] 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 data set. If this condition is met, additional analysis is not necessary. If the condition is not met, then the analysis proceeds to Step 3.

6.2 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 two identification risks: size of potential population and variable specificity. The Publication Scoring Criteria is used to assess the need to suppress small cells as a result of a small numerator, small denominator, or both small numerator and small denominator where a small numerator is less than 11 and a small denominator is less than 20,001. That is why the Publication Scoring Criteria takes into account both numerator (e.g., Events) and denominator (e.g., Geography) variables.

The Publication Scoring Criteria is based on a framework that has been in use by the Illinois Department of Public Health, Illinois Center for Health Statistics. 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 6.3.

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

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

Events

Variable	Characteristics	Score
Events	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.

Sex

Variable	Characteristics	Score
Sex	Male or Female	+1

Sex is commonly represented as two categories: male and female. Because the number of categories is small, just knowing a person's reported sex is not enough to pose a risk of identifying that person. The score of +1 reflects that inclusion of the variable in a table introduces increased specificity; however, that it only has two potential values gives it a low risk.

In cases where an additional stratification of other/unknown is used for sex, the reviewer will need to assess potential for increased risk based on the inclusion of the additional stratification.

Although the variable "Sex" is often called "Gender", it should not be confused with the variables "sexual orientation" and "gender identity." According to definitions from the American Psychological Association, "Sexual orientation refers to the sex of those to whom one is sexually and romantically attracted" and "Gender identity refers to "one's sense of oneself as male, female, or transgender."[15]

[15] Definition of Terms: Sex, Gender, Gender Identity, Sexual Orientation; Excerpt from: The Guidelines for Psychological Practice with Lesbian, Gay, and Bisexual Clients, adopted by the APA Council of Representatives, February 18-20, 2011. http://www.apa.org/pi/lgbt/resources/sexuality-definitions.pdf

Additional information is provided from San Francisco County at https://www.sfdph.org/dph/files/hc/HCFinance/agendas/2014/August%205/pdf%20rev%20072514%20re%20age%20adopted%20090313%20-%20SFDPH%20Sex%20and%20Gender%20Guidelines.pdf.

Age Range

Variable	Characteristics	Score
Age Range	>10-year age range	+2
	6-10 year age range	+3
	3-5 year age range	+5
	1-2 year age range	+7

Age ranges receive a higher score for smaller ranges of years due to the increased risk for identification.

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 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.

Race Group and Ethnicity

Race Group	White, Asian, Black or African American	+2
	White, Asian, Black or African American, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Mixed	+3
	Detailed Race	+4
Ethnicity	Hispanic or Latino - yes or no	+2
	Detailed ethnicity	+4
Race/Ethnicity Combined	This applies when race and ethnicity are collected in a single data field	
	White, Asian, Black or African American, Hispanic or Latino	+2
	White, Asian, Black or African American, Hispanic or Latino, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Mixed	+3
	Detailed Race/Ethnicity	+4

Race and Ethnicity are collected in a number of different ways on the different state and federal data collection tools. At the federal level, starting in 1997, Office of Management and Budget required federal agencies to use a minimum of five race categories:

- White,
- Black or African American,
- American Indian or Alaska Native,
- Asian, and
- Native Hawaiian or Other Pacific Islander.

Ethnicity asks individuals if they are Hispanic or Latino. Additional specificity for Ethnicity may be requested.

The California population in general is approximately:[16]

- 40% White
- 13% Asian
- 6% Black or African American
- <1% American Indian
- <1% Native Hawaiian and other Pacific Islander
- 37% Hispanic or Latino

Based on these percentages, Race Group at the level of White, Asian and Black or African American is given a score of +2 because the Asian and Black or African American groups are relatively small. If the reporting is for the OMB standard categories, White, Asian, Black or African American, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and Mixed, then the score is +3. If more specificity is requested for Race Groups the score is +4 because the other groups are much smaller at less than 1% of the overall population. Similarly, for the Hispanic or Latino Ethnicity the score is a +2 for a yes or no answer, whereas more detailed ethnicity results in a higher score of +4.

For Race/Ethnicity Combined fields, the scoring is +2 for the groups White, Asian, Black or African American, Hispanic or Latino. The score is +3 for the OMB standard categories with Hispanic or Latino, White, Asian, Black or African American, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and Mixed. The score is +4 for more detailed categories.

[16] Based on Year 2010 from the State of California, Department of Finance, Report P-1 (Race): State and County Population Projections by Race/Ethnicity, 2010-2060. Sacramento, California, January 2013

Race and Ethnicity 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. Appendix C provides the county specific demographics produced by Department of Finance for reference.

Three scenarios are presented to help demonstrate how to use the three race group and ethnicity scoring criteria.

First Scenario – Complete Cross-Tabulation between Race and Ethnicity

Consider this table:

	Hispanic	Non-Hispanic	
Black	50	250	300
White	200	1000	1200
Asian	5	95	100
	255	1345	1600

This is the most granular you can get, so you would add both the Race and Ethnicity score 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.” So the above table would become:

- Black 250
- White 1000
- Asian 95
- Hispanic 255

This 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

If you did this, the above table would become:

- Black 300
- White 1200
- Asian 100

- Hispanic 255

Note that this is the only scenario where you can't add up all the categories to get a total population. 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.

Language Spoken

Variable	Characteristics	Score
Language Spoken	English, Spanish, Other Language	+2
	Detailed Language	+4

Language spoken is captured in a variety of data systems to support individuals in receiving services in the language they speak. The following table is taken from the report: Medi-Cal Beneficiaries by Primary Language Report of October, 2010.[17] This frequency distribution was used to determine the groupings for the scoring above.

Language Spoken	Count of Medi-Cal Members	Percent of Count
Total	**7,835,022**	**100.00**
English	4,135,060	52.78
Spanish	2,840,758	36.26
Vietnamese	141,289	1.80
Cantonese	85,750	1.09
Armenian	65,096	0.83
Russian	41,252	0.53
Tagalog	39,361	0.50
Mandarin	35,330	0.45
Hmong	33,594	0.43
Korean	27,814	0.35
Farsi	26,123	0.33
Arabic	23,929	0.31
Cambodian	20,476	0.26
Lao	8,355	0.11
Other Chinese	7,483	0.10
Mien	3,803	0.05
Sign Language	2,637	0.03
Thai	1,940	0.02
Portuguese	1,666	0.02
Ilocano	1,661	0.02

[17] http://www.dhcs.ca.gov/services/MH/InfoNotices-Ltrs/Documents/InfoNotice-PrimaryLang-Enclosure1.pdf

Language Spoken	Count of Medi-Cal Members	Percent of Count
Samoan	1,306	0.02
Japanese	1,215	0.02
French	653	0.01
Turkish	376	0.00
Hebrew	367	0.00
Polish	275	0.00
Italian	252	0.00
Other and unspecified	287,201	3.67

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 +2 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.

Time – Reporting Period

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

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 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.

Geography

Variable	Characteristics	Score
Residence Geography*	State or geography with population >2,000,000	-5
	Population 1,000,001 - 2,000,000	-3
	Population 560,001 - 1,000,000	-1
	Population 250,000 - 560,000	0
	Population 100,000 - 250,000	+1
	Population 50,001 - 100,000	+3
	Population 20,001 - 50,000	+4
	Population ≤ 20,000	+5
Service Geography*	State or geography with population >2,000,000	-5
	Population 1,000,001 - 2,000,000	-4
	Population 560,001 - 1,000,000	-3
	Population 250,000 - 560,000	-1
	Population of reporting region 20,001 - 250,000	0
	Population of reporting region ≤20,000	+1
	Address (Street and ZIP)	+3

* 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”.

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). 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 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 not 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 results in the highest score and thus the highest risk.

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 Office of Statewide Health Planning and Development (OSHPD) [18] 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.

Variable Interactions

Variable	Characteristics	Score
Variable Interactions	Only Events (minimum of 5), Time, and Geography (Residence or Service)	-5
	Only Events (minimum of 3), Time, and Geography (Residence or Service)	-3
	Only Events (no minimum), Time, and Geography (Residence or Service)	0

[18] Office of Statewide Health Planning and Development (OSHPD), Patient Origin & Market Share Reports, Retrieved from http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PivotTables/PatOrginMkt/default.asp on January 22, 2016.

	Events, Time, and Geography (Residence or Service) + 1 variable	+1
	Events, Time, and Geography (Residence or Service) + 2 variables	+2
	Events, Time, and Geography (Residence or Service) + 3 variables	+4

This criteria specifically addresses the interaction of the variables in a given data presentation and requires the analyst to identify dependent as opposed to independent variables. This criteria is 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.

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.

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

This criteria is 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.[19] 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.[20] For this reason, additional points are added because of the inclusion of multiple dependent variables.

Other Variables

Variables other than those specified in the Publication Scoring Criteria can be released only after an additional review by the department's Statistical Expert on a case by case basis. A guideline that can be considered in performing this review is the following scoring.

[19] NORC, "NORC Recommendations for California Department of Health Care Services (DHCS) Data De-Identification Guidelines (DDG)," January 8, 2016.
[20] NORC, "Case Study: The Disclosure Risk Implications of Small Cells Combined with Multiple Tables or External Data," January 8, 2016.

Variable	Characteristics	Score
Other Variables	<5 groups or categories	+3
	5-9 groups	+5
	10+ groups	+7

Considerations include not just the number of groups, but also the characteristics of the variables. Consider whether the variable represents an aggregation (Diagnosis Related Groups) or a specific item (ICD-10 Code). Also consider the availability of the variable to the public when also associated with other information, in particular with variables that may be personal characteristics.

6.3 Assessing Potential Risk – Alternate Methods

As noted in Section 6.2, the Publication Scoring Criteria is based on a framework that has been in use by the Illinois Department of Public Health, Illinois Center for Health Statistics. 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. Some of those methods are highlighted here.

- Ohio Department of Health published a Data Methodology Standards for Public Health Practice.[21] This method is framed around the concept that a Disclosure Limitation Standard for tabulations of confidential Ohio Department of Health data shall be suppressed when the table denominator value minus the table numerator value is less than 10.
- Washington State Department of Health published Guidelines for Working with Small Numbers[22] that highlights many topics covered in the CHHS DDG but also discusses the use of relative standard error (RSE) to assess reliability of data in addition to steps to take protect confidentiality.
- Colorado Department of Public Health and Environment published Guidelines for Working with Small Numbers[23] which also addresses many of the same topics.

The size of numerators and denominators vary in each of the documents above although the principles are consistent.

[21] 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.
[22] Washington State Department of Health. "Guidelines for Working with Small Numbers." N.p., 15 October 2012. Retrieved from http://www.doh.wa.gov/Portals/1/Documents/5500/SmallNumbers.pdf.
[23] Colorado Department of Public Health and Environment. "Guidelines for Working with Small Numbers." Retrieved from http://www.cohid.dphe.state.co.us/smnumguidelines.html

6.4 Statistical Masking

Statistical masking provides an extensive set of tools that can be used to mitigate potential risk in a given data presentation. As discussed in Section 4.4, the data releaser will assess the need for statistical masking when the assessment in Step 3 identified potential risk. Each department will document statistical masking processes that are routinely used in data preparation for public release.

As discussed in section 4.4, initial methods to address sensitive or small cells, as well as complimentary cells include the following:

- Reduce Table Dimensions
- Reduce Granularity of Variable(s), aka Recoding or Aggregation
- Cell Suppression and Complementary Cell Suppression

Small cell sizes are typically encountered 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 county, stratified by sex, stratified by 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. Examples of incidents may result from 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 for the reporting so that the method chosen for masking can still maximize the usefulness of the data provided. Choices for each condition 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 suppressing the small cell with an indicator that it is less than 11.
b) Granular variables. A common approach to this situation would be to decrease the granularity of the variables although suppressing the small cell with an indicator that it is less than 11 is also an option.

c) Rare events. In these cases it becomes very challenging to suppress the value in a way that it will not be able to be used with other public information to identify individuals. Additionally, with rare events, there is more significance in the variance of small numbers.

In addition to small cells, complementary cells must also be suppressed. Complementary cells are those which must be suppressed to prevent someone from being able to calculate the suppressed cell based on row or column totals in combination with other data in that row or column.

Suppressing small cell values and complimentary cells can be done in two ways.

1) Use a symbol to indicate the cell has been suppressed. Identify any other cells (complimentary cells) that can be used to calculate the small cell and use a symbol to indicate the cell has been suppressed.
2) Use a symbol to indicate the cell has been suppressed or leave the cell blank and remove the value from all pertinent row and column totals so that the cell cannot be calculated. This negates the need for evaluation of complementary cells. This method must be used with great caution because the totals may actually be published in other non-related tables. For this reason the method is not recommended.

When suppressing values, the following footnote to indicate the suppression is recommended:

> "Values are not shown to protect confidentiality of the individuals summarized in the data."

In addition to the above, there are a number of other methods that may be used for Statistical Masking. Methods discussed in the "Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology" include the following for tables of counts or frequencies and for magnitude data.[24]

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

[24] 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.

    - Random Rounding
    - Controlled Rounding
    - Controlled Tabular Adjustment
- Protecting Sensitive Cells Before Tabulation

Tables of Magnitude Data

- Defining Sensitive Cells – Linear Sensitivity Rules
- Protecting Sensitive Cells After Tabulation
- Protecting Sensitive Cells Before Tabulation

7) Approval Processes

After completion of the statistical de-identification process, each department will specify the additional review steps necessary for public release. This may vary depending on the purpose of the release and whether or not the department/program is a HIPAA covered entity.

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.

Departments and offices may consider the following components for reviews related to data that has been de-identified.

- Statistical Review to Assess De-identification
  (for HIPAA entities this may be an Expert Determination Review)
- Legal Review
- Departmental Release Procedures

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

The department or office may designate individuals within the department to provide a 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.

For HIPAA covered entities, this will be performed by individuals who 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"[25]

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 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 A. 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.

Legal Review (Step 5)

Step 5 in the Data Assessment for Public Release Process provides for a legal review within the department. This may vary depending on the purpose of the release and whether or not the department or program is a HIPAA covered entity or not. This review may 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 may review the expert

[25] 45 CFR section 164.514 (b)

determination documentation to ensure compliance with the HIPAA Privacy Rule as applicable.

Departmental Release Procedures (Step 6)

Step 6 in the Data Assessment for Public Release Process provides for 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, presentation, tables, PRA responses, media responses and legislative responses.

Potential reviews include Public Affairs. Public Affairs is often designated 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 Agency Approval or Governor's Office approval. Public Affairs may also be designated to review content to assess the data table for compliance with the Americans with Disabilities Act of 1990[26] (ADA).

Departments may also consider processes for quality assurance reviews: The may apply to data products being added to the web sites to ensure that they have had appropriate reviews and de-identification steps. It may also include 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.

[26] 42 U.S.C 12101 et seq.

8) 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 CHHS governance structure as described in the CHHS Information Strategic Plan.[27] Governance for the CHHS DDG will provide the following support for departments and offices.

- Maintain the CHHS DDG, which will include updates and revisions to the document as well as annual reviews for currency.
- Coordinate integration of the CHHS DDG into the Statewide Health Information Policy Manual (SHIPM), Section 2.5.0 De-identification[28] and the CHHS Open Data Handbook.
- Convene a Peer Review Team (PRT).
- Provide for escalation of issues that cannot be resolved by the PRT.

The CHHS 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 CHHS 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 CHHS DDG.
- Provide for escalation and review of data de-identification questions or issues that a department is not comfortable resolving independently.
- Develop training tools to be used by departments when developing and implementing department specific DDGs based on the content of the CHHS DDG.

The PRT will not review all disclosures or data released by each department.

[27] California Health and Human Services Agency, Information Strategic Plan 2016.
[28] http://www.ohi.ca.gov/calohi/ohii-shipm-manual.htm

9) 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 data sets that are publicly available that may be used in combination with CHHS 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.

Vital Records Data

Another common data set 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 $25 fee for each certified copy of a birth certificate and $21 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 amount of certificates may generate an investigation among vital records staff as to why so many certificates were requested at once.

CHHS 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 data set. The CHHS Open Data Workgroup will be providing easier access to both lists of data currently on the portal as well as data sets planned for addition to the porta. While significant with over 100 data sets, 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. CHHS 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.

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

- Fair Hearing Decisions include appellant's initials and possibly other information, depending on issue 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.

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.[29] The DRU produces the following products which serve as the basis for

[29] http://www.dof.ca.gov/research/demographic/dru/index.php

understanding the population characteristics and distributions that frequently make up the denominators in the review of data sets.

- 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.

Commonly Shared Information

With the growth of social media, people frequently share information through tools such as Facebook, Linked In, and Tweets. 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.

Geographic Information

Geographic information is particularly suited to being combined with other geographic information given the relatively standardized was 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.[1] 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."[30]

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

10) Development Process

The CHHS Data Subcommittee requested the convening of the CHHS Data De-Identification Workgroup to develop the DDG.

The DDG Workgroup began with an orientation to the topic of data de-identification and presentations by the DHCS, OSHPD and California Department of Public Health (CDPH) regarding current practices and activities related to data de-identification. The DDG Workgroup used the Public Aggregate Reporting for DHCS Business Reports (PAR-DBR) as a starting point for initial drafts. The PAR-DBR had been developed between April and August, 2014 through a workgroup processes within DHCS with input and presentations from OSHPD, CDPH, and University of California, Los Angeles California Health Interview Survey. The PAR-DBR served as a basis for this document, including the literature review conducted as part of the development of the PAR-DBR.

The development process was designed to include an updated literature review, case examples and broad discussion among CHHS programs. Publishing data publicly is always a balance between the protection of confidentiality and the usability of the data.

The project timeline for the CHHS DDG Workgroup is below:

3/15/15	Planning Meeting Part 1 – Participants included DHCS, CDPH, OSHPD, OHII
3/20/15	Planning Meeting Part 2 – Participants included DHCS, CDPH, OSHPD, OHII
4/7/15	Present Objectives for the project and use the DHCS PAR-DBR as an example
4/23/15	Presentations from OSHPD and CDPH regarding current processes and approach to small cell sizes
5/5/15	Discuss concept of uniqueness as a way to measure risk for re-identification and gather input from Departments/Offices regarding DDG variables and topics
5/27/15	Review initial draft DDG – Focus on new sections of the document
6/8/15	Review initial draft DDG – Focus on Data Assessment for Public Release Procedure
May & June, 2015	Meet with each department/office individually

6/30/15	Review draft DDG version 0.2
July 2015	Departments/offices vet the DDG within their departments/offices
8/21/15	Received input from the CHHS Risk Management Committee
8/6/15	Review draft DDG version 0.3
9/14/15	Progress update for DDG Workgroup and discussion of additional topics
12/18/15	Presentation from NORC to review their findings of the draft DDG
1/8/16	Receive final recommendations from NORC
Jan. 2016	Provide DDG version 0.4 to DDG Workgroup
2/18/16	Review and discussion of draft DDG version 0.4 with the DDG Workgroup
3/18/16	Provide DDG version 0.5 with outstanding comments from the DDG Workgroup to the Data Subcommittee
4/18/16	Provide revised draft DDG to the Data Subcommittee.
5/24/16	Provide draft DDG version 0.7 from the CHHS Data Subcommittee to the CHHS Advisory Council. The Advisory Council shared the DDG version 0.7 with the other subcommittees and discussed the version 0.7 at the 6/8/16 meeting and the version 0.8 at the 7/6/16 meeting.
7/7/16	Provide draft DDG version 0.10 to the Undersecretary.
9/23/16	DDG approved by CHHS Undersecretary as Version 1.0.

The final document will be incorporated into the Open Data Handbook and made publicly available.

11) Legal Framework

The overarching legal framework for the CHHS Data De-identification Guidelines 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 healthcare-related information. HIPAA requirements apply in combination with IPA requirements.

"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. (http://www.oshpd.ca.gov/Boards/CPHS/)

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

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 4744
- Welf. & Inst. Code 4659.22

Environmental Health Hazards

- Health & Saf. Code 59016

General Public Health Records

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

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 121022
- Health & Saf. Code 121023
- Health & Saf. Code 121025
- Health & Saf. Code 121075
- Health & Saf. Code 121085
- Health & Saf. Code 121110
- Health & Saf. Code 121125
- Health & Saf. Code 121010
- Health & Saf. Code 120820
- Health & Saf. Code 120980
- Health & Saf. Code 121280
- Health & Saf. Code 120962

- Health & Saf. Code 120975
- Health & Saf. Code 121080
- Health & Saf. Code 121090
- Health & Saf. Code 121095
- Health & Saf. Code 121120
- 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 5328 through 5328.9
- Welf. & Inst. Code 5329 (aggregation and publication of data)
- Welf. & Inst. Code 5540
- Welf. & Inst. Code 5610
- Welf. & Inst. Code 4135
- Educ. C. 56863

Medi-Cal Data

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

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

Federal Laws Governing Public Data Release
(please note that 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)

Data De-identification

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.

The California Office of Health Information Integrity (CalOHII) 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. The 2014 assessment that was revised July 2015, identified programs and departments in CHHS

that are considered covered entities under HIPAA as a Health Care Provider, Health Care Plan, Health Care Clearinghouse, Hybrid Entity or Business Associate. Detail is provided in Appendix B. One difference between CA IPA and HIPAA is the documentation requirement in HIPAA for data de-identified using the Expert Determination method. Each of the following departments will need to identify which programs within the department are impacted by HIPAA as part of the department specific DDG.

- Department of Aging
- Department of Developmental Services
- Department of Health Care Services
- Department of Managed Health Care
- Department of Public Health
- Department of Social Services
- Department of State Hospitals
- Health and Human Services Agency
- Office of Systems Integration

For programs and departments that are covered by HIPAA, de-identification must meet the HIPAA standard. The DDG serves as a tool to make and document an expert determination consistent with the HIPAA standard. The following comes from federal guidance for HIPAA that provides more detail regarding Safe Harbor and Expert Determination under the HIPAA standard.

The HIPAA Standard[31] for de-identification of protected health information (PHI)[32] 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: (http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-

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

identification/guidance.html) (Formatting of text may be different than the original document.)

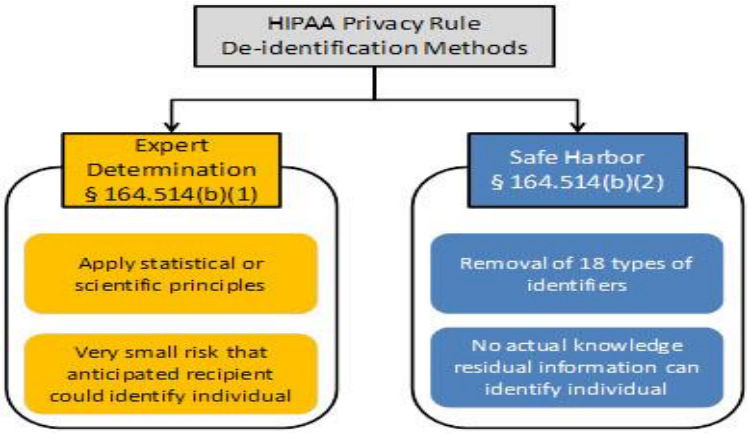
The HIPAA De-identification Standard

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.
>
> (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 1, the Privacy Rule provides two methods by which health information can be designated as de-identified.

Figure 1. Two methods to achieve de-identification in accordance with the HIPAA Privacy Rule.



The first is the "Expert Determination" method:

> (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.

(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.

12) Abbreviations and Acronyms

CalOHII ..... California Office of Health Information Integrity
CDC ..... Centers for Disease Control and Prevention
CDPH ..... California Department of Public Health
CDSS ..... Department of Social Services
CHHS ..... 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
HIPAA ..... Health Insurance Portability and Accountability Act
IPA ..... Information Practices Act
MHSOAC .... Mental Health Services Oversight and Accountability Commission
OSHPD ..... Office of Statewide Health Planning and Development
PAR-DBR .... Public Aggregate Reporting - DHCS Business Reports
PHI ..... Protected Health Information
PI..... Personal Information
PRA..... Public Records Act
PRT ..... Peer Review Team

13) 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 later in this handbook).[33]

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

[33] http://chhsopendata.github.io/

14) References

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

CHHS Open Data Handbook - http://chhsopendata.github.io/

CHHS, Information Strategic Plan 2016.

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

Committee for the Protection of Human Subjects (CPHS), CPHS Bulletin & Update, January, 2005.

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.

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.

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.

NAHDO-CDC Cooperative Agreement Project CDC Assessment Initiative. "Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk." December 2004. Retrieved from http://api.ning.com/files/sCi4ZnrAubmkUqLO5Zfm3XYlq*7jctjEJXwGDDMepE4_/Statapproachesforsmallnumbers.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 from http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

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.

Privacy Technical Assistance Center, U.S. Department of Education. “Data De-identification: An Overview of Basic Terms.” May 2013. Retrieved from http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf

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

State of California, Department of Health Care Services, Trend in Medi-Cal Program Enrollment by Managed Care Status - for Fiscal Year 2004-2012, 2004-07 - 2012-07, Report Date: July 2013. Retrieved from

http://www.dhcs.ca.gov/dataandstats/statistics/Documents/1_6_Annual_Historic_Trend.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.

The Centers for Medicare and 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.

Washington State Department of Health. "Guidelines for Working with Small Numbers." N.p., 15 October 2012. Retrieved from http://www.doh.wa.gov/Portals/1/Documents/5500/SmallNumbers.pdf.

15) Appendix A: Expert Determination Template

HIPAA covered entities in CHHS 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 CHHS 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 CHHS 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"*	

16) Appendix B: 2015 HIPAA Reassessment Results

The CalOHII 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. To help ensure full compliance with HIPAA, CalOHII conducted a reassessment with all State Departments in January 2014 and updated as of July 27, 2015.[34] The following are the self-reported results of this reassessment:

	DEPARTMENTS	COVERED ENTITIES					IMPACTED ENTITIES		
		Health Care Provider	Health Care Plan	Health Care Clearinghouse	Hybrid Entity	Business Associate	Trading Partner	Impacted by Data Content	Health Oversight Agency
	COVERED ENTITIES & BUSINESS ASSOCIATES								
1	Aging, Department of					X			X
2	Controllers Office, State					X			
3	Corrections and Rehabilitation, CA Dept. of,	X			X				
4	Developmental Services, Dept. of	X		X	X	X	X	X	X
5	Forestry and Fire Protection, Dept. of					X			
6	Health and Human Services Agency					X	X	X	
7	Healthcare Services, Department of		X				X	X	X
8	Justice, Department of					X			
9	Managed Health Care, Dept. of					X			X
10	Public Employees' Retirement System		X		X		X	X	
11	Public Health, Department of	X	X		X			X	X
12	Social Services, Dept. of					X			
13	State Hospitals, Dept. of	X			X	X	X	X	
14	Systems Integration, Office of					X			
15	Veterans Affairs, Dept. of (CalVET)	X			X				
	IMPACTED ENTITIES								
1	Health Information Integrity, California Office of								X
2	Health Planning and Development, Office of Statewide							X	
3	Industrial Relations, Dept. of							X	X
4	Insurance, Dept. of								X
5	Inspector General, Office of								X

[34] http://www.ohi.ca.gov/calohi/download2011-HIPAA%20Assessment%20Results%207-27-2015.pdf

17) Appendix C: State and County Population Projections

The following table is provided for reference related to the race and ethnicity composition at the county level. It is *State of California, Department of Finance, Report P-1 (Race): State and County Population Projections by Race/Ethnicity, 2010-2060.* Sacramento, California, January 2013. The table is for year 2010.

State/ County	Race/Ethnicity							
	Total (All race groups)	White, not Hispanic or Latino	Black, not Hispanic or Latino	American Indian, not Hispanic or Latino	Asian, not Hispanic or Latino	Native Hawaiian and other Pacific Islander, not Hispanic or Latino	Hispanic or Latino	Multi-Race, not Hispanic or Latino
California	**37,309,382**	**15,024,945**	**2,188,296**	**163,040**	**4,827,438**	**131,415**	**14,057,596**	**916,651**
Alameda	1,513,236	514,086	186,737	4,098	395,898	12,337	343,141	56,939
Alpine	1,163	869	0	204	2	0	71	17
Amador	37,853	30,091	950	539	447	53	4,859	913
Butte	219,990	164,870	3,139	3,376	9,458	397	31,670	7,080
Calaveras	45,462	37,999	353	518	526	59	4,779	1,227
Colusa	21,478	8,601	153	284	247	50	11,892	251
Contra Costa	1,052,211	508,220	93,096	3,033	149,853	4,532	256,047	37,431
Del Norte	28,544	18,522	1,060	1,928	933	21	5,126	953
El Dorado	180,921	143,909	1,289	1,543	6,739	248	22,443	4,750
Fresno	932,377	307,295	45,680	6,080	86,637	1,067	469,935	15,682
Glenn	28,143	15,688	181	463	663	17	10,664	467
Humboldt	134,663	103,996	1,404	6,940	3,127	320	13,560	5,316
Imperial	175,389	24,406	5,359	1,639	1,954	75	140,945	1,010
Inyo	18,528	12,309	102	1,895	184	12	3,629	396
Kern	841,146	325,711	45,798	5,933	33,266	996	414,414	15,028
Kings	152,656	54,303	10,686	1,305	5,343	216	77,595	3,208
Lake	64,599	47,973	1,186	1,531	647	81	11,165	2,016
Lassen	35,136	23,452	2,999	992	427	153	6,243	870
Los Angeles	9,824,906	2,746,305	821,829	19,527	1,336,086	23,152	4,694,972	183,035
Madera	151,328	57,494	5,204	1,818	2,661	98	81,807	2,246
Marin	252,731	184,377	7,069	520	14,004	423	39,459	6,879
Mariposa	18,193	15,224	118	456	158	21	1,677	539
Mendocino	87,924	60,398	544	3,433	1,469	79	19,691	2,310
Merced	255,937	83,475	8,742	1,134	17,363	466	140,472	4,286
Modoc	9,648	7,677	69	280	53	17	1,344	208
Mono	14,240	9,731	36	217	206	9	3,815	226
Monterey	416,259	136,348	11,334	1,372	24,430	1,882	231,700	9,193
Napa	136,811	77,088	2,457	533	9,377	299	44,235	2,823
Nevada	98,639	85,120	331	787	1,295	83	8,703	2,320
Orange	3,017,327	1,336,843	45,894	6,247	540,485	8,507	1,010,752	68,599

State/ County	Race/Ethnicity							
	Total (All race groups)	White, not Hispanic or Latino	Black, not Hispanic or Latino	American Indian, not Hispanic or Latino	Asian, not Hispanic or Latino	Native Hawaiian and other Pacific Islander, not Hispanic or Latino	Hispanic or Latino	Multi-Race, not Hispanic or Latino
Placer	350,275	263,747	4,448	2,063	22,443	685	46,677	10,214
Plumas	19,911	16,989	173	453	98	14	1,602	581
Riverside	2,191,886	874,405	133,791	10,951	127,558	5,891	993,930	45,361
Sacramento	1,420,434	691,338	140,694	7,973	200,201	13,795	307,513	58,920
San Benito	55,350	20,573	380	215	1,542	54	31,721	865
San Bernardino	2,038,523	684,856	172,602	8,660	122,187	5,970	1,003,256	40,991
San Diego	3,102,745	1,501,675	148,728	14,121	333,728	13,606	999,392	91,494
San Francisco	806,254	338,874	46,758	1,808	268,020	3,145	122,869	24,780
San Joaquin	686,588	248,202	49,199	3,220	94,812	3,315	267,086	20,752
San Luis Obispo	269,713	191,725	5,392	1,367	8,622	334	56,309	5,965
San Mateo	719,729	303,475	19,474	1,134	178,665	10,225	184,420	22,337
Santa Barbara	424,050	201,823	7,507	1,817	20,281	675	183,511	8,436
Santa Clara	1,786,429	627,438	43,926	4,085	573,622	6,413	481,108	49,838
Santa Cruz	263,260	156,796	2,357	972	11,260	288	84,804	6,783
Shasta	177,472	145,533	1,429	4,150	4,893	216	15,410	5,841
Sierra	3,230	2,883	4	34	3	2	258	48
Siskiyou	44,893	35,691	537	1,547	548	58	4,663	1,848
Solano	413,117	170,275	58,396	1,853	59,126	3,304	99,759	20,405
Sonoma	484,084	321,695	7,009	3,560	17,581	1,404	120,414	12,422
Stanislaus	515,205	243,208	12,534	2,894	24,168	3,170	216,228	13,003
Sutter	94,669	48,033	1,734	925	13,582	251	27,326	2,818
Tehama	63,487	45,708	347	1,213	548	53	14,010	1,610
Trinity	13,713	11,307	38	536	183	12	1,080	557
Tulare	443,066	145,549	5,505	3,319	13,543	370	269,012	5,767
Tuolumne	55,144	45,279	1,161	831	546	51	5,950	1,327
Ventura	825,077	402,144	13,216	2,363	55,015	1,351	333,230	17,758
Yolo	201,311	100,679	5,025	1,094	26,065	842	61,057	6,549
Yuba	72,329	42,666	2,134	1,260	4,659	256	18,192	3,162