Public Aggregate Reporting – Guidelines Development Project (PAR-GDP)

Public Aggregate Reporting – DHCS Business Reports Guidelines

California Department of Health Care Services

August 25, 2014

Version 1.6

Revision History

Version	Date	Author	Brief Description of Changes
1.0	5/5/14	L. Scott	Initial draft for review.
1.1	5/12/14	L. Scott	Updates based on feedback at PAR-GDP Development Team Meeting May 6, 2014
1.2	5/21/14	L. Scott	Updates based on feedback at PAR-GDP Development Team Meeting May 14, 2014
1.3	6/9/14	L. Scott	Updates based on feedback from PAR-GDP Development Team and Office of Legal Services Privacy Team
1.4	6/30/14	L. Scott	Updates based on feedback from Office of Legal Services and Executive Sponsor
1.5	7/7/14	L. Scott	Updates based on feedback from the Review Team
1.6	8/25/14	L. Scott	Updates based on feedback from the Office of Legal Services and Executive Sponsor

Table of Contents

1) Purpose

Describe Public Aggregate Reporting – DHCS Business Reports (PAR-DBR) Guidelines (Guidelines) created by the "Public Aggregate Reporting – Guidelines Development Project (PAR-GDP)" and provide documentation of the methods described in the Guidelines with the associated analysis that justifies the use of the Guidelines.

2) Background

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

The HIPAA Standard[1] for de-identification of protected health information (PHI)[2] 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.

3) Scope

The PAR-GDP focused on the assessment of aggregate data for purposes of de-identification per 45 C.F.R. section 164.514.(b)(1) [Expert Determination], and for the

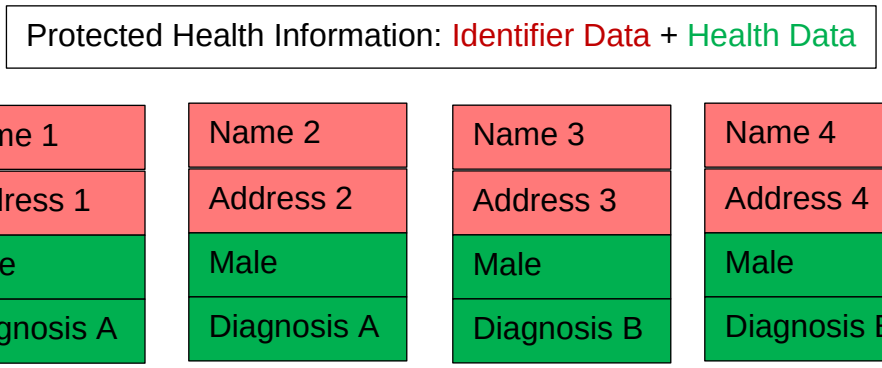
[1] The Standard is found in the HIPAA Privacy Rule, 45 CFR section 164.514(a).
[2] "PHI" is defined as information which relates to the individual's past, present, or future physical or mental health or condition, the provision of health care to the individual, or the past, present, or future payment 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)

potential release in reports for the public that DHCS generates for its own DHCS business needs (DHCS business reports). The project resulted in the PAR-DBR Guidelines represented in this document. Assessment of aggregate data for potential release in response to external requests (e.g., but not limited to, Public Records Act (PRA) requests), is subject to the Public Aggregate Reporting - For External Requests (PAR-FER) Guidelines.

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 (sum of the components), percentages, rates, averages, or other statistical groupings.

The Guidelines also do not address data releases that involve record level data, whether or not the records include PHI or Personal Information (PI).[3] Figure 1 highlights data fields that make up PHI which include identifiers combined with health information. Aggregation involves grouping of information about individuals.

Figure 1: Graphic showing the two components of Protected Health Information



Record or Person based reporting: Include information for each person that may have any combination of variables (Name, Address, Gender, Diagnosis). There will be one entry for each individual. If Name or Address is included with Diagnosis it is considered PHI.

Aggregation: Summarize that there are two male individuals with Diagnosis A and two male individuals with Diagnosis B.

The Guidelines provide a structured and documented set of procedures to follow to prepare aggregate data for public reporting in DHCS business reports. The current DHCS document review processes are documented in the Health Administrative

[3] Personal Information is defined by California Civil Code section 1798.3.

Manual (HAM).[4] These processes have been supplemented as described in the Guidelines.

4) Reporting Assessment Decision Tree

The Reporting Assessment Decision Tree, Figure 3, is to be applied to aggregate data tables to assess risk for data release. Of note, the decision tree does not include steps required for release of record level data by DHCS, which will continue to follow processes that have already been established.[5] The PAR-GDP has addressed the five steps to be used for assessment of aggregate data when considering public release of such data in DHCS business reports.

Figure 2 shows an example table with the numerator and the denominator highlighted. Data in Figure 2 comes from the "Trend in Medi-Cal Program Enrollment by Managed Care Status - for Fiscal Year 2004-2012, 2004-07 - 2012-07." [6]

Figure 2: Figure to demonstrate numerators and denominators in a table.

Year	# of Medi-Cal Members in Fee For Service (in thousands)	# of Medi-Cal Members in Managed Care (in thousands)
2012	2,775	4,853
2011	3,067	4,527

Column Headings

Table Cell

Row Headings

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

[4] DHCS Health Administrative Manual.
[5] See http://dhcsintranet/technology/ISO/Pages/DataReleases.aspx for more information on record level data release processes.
[6] Report Date: July 2013
http://www.dhcs.ca.gov/dataandstats/statistics/Documents/1_6_Annual_Historic_Trend.pdf

The Cells in the table are the boxes with values in them, as opposed to the row and column headings which are in blue. In the example above, "2,775" is the value in a table cell and it also represents a numerator. The sum of the row for year 2012 (2,775 + 4,853 = 7,628) represents the denominator.

Assessment of risk for data release takes into account assessment of de-identification, as well as legal, programmatic, and policy risks. The statistical analysis described in Steps 1 through 4 of the Reporting Assessment Decision Tree focuses on methods used for de-identification, whereas Step 5 focuses on the legal, programmatic, and policy risks.

The two methods to achieve de-identification in accordance with the HIPAA Privacy Rule are: 1) Expert Determination; and 2) Safe Harbor. The statistical steps described in the Reporting Assessment Decision Tree support the Expert Determination method. The Safe Harbor method is achieved by removing the 18 identifiers specified in Section 164.514(b)(2). Please see Section 8 of this document, Legal Framework, for a detailed description of the Safe Harbor Method.

Reporting Assessment Decision Tree:
Assesses risk for data release of aggregate data through a stepwise process

Steps 1 and 2

The first and second steps in the Reporting Assessment Decision Tree represent the Combination Numerator – Denominator Condition. This 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.

Numerator – number of events with the characteristics of the given row and column
Denominator – the population from which the events arise

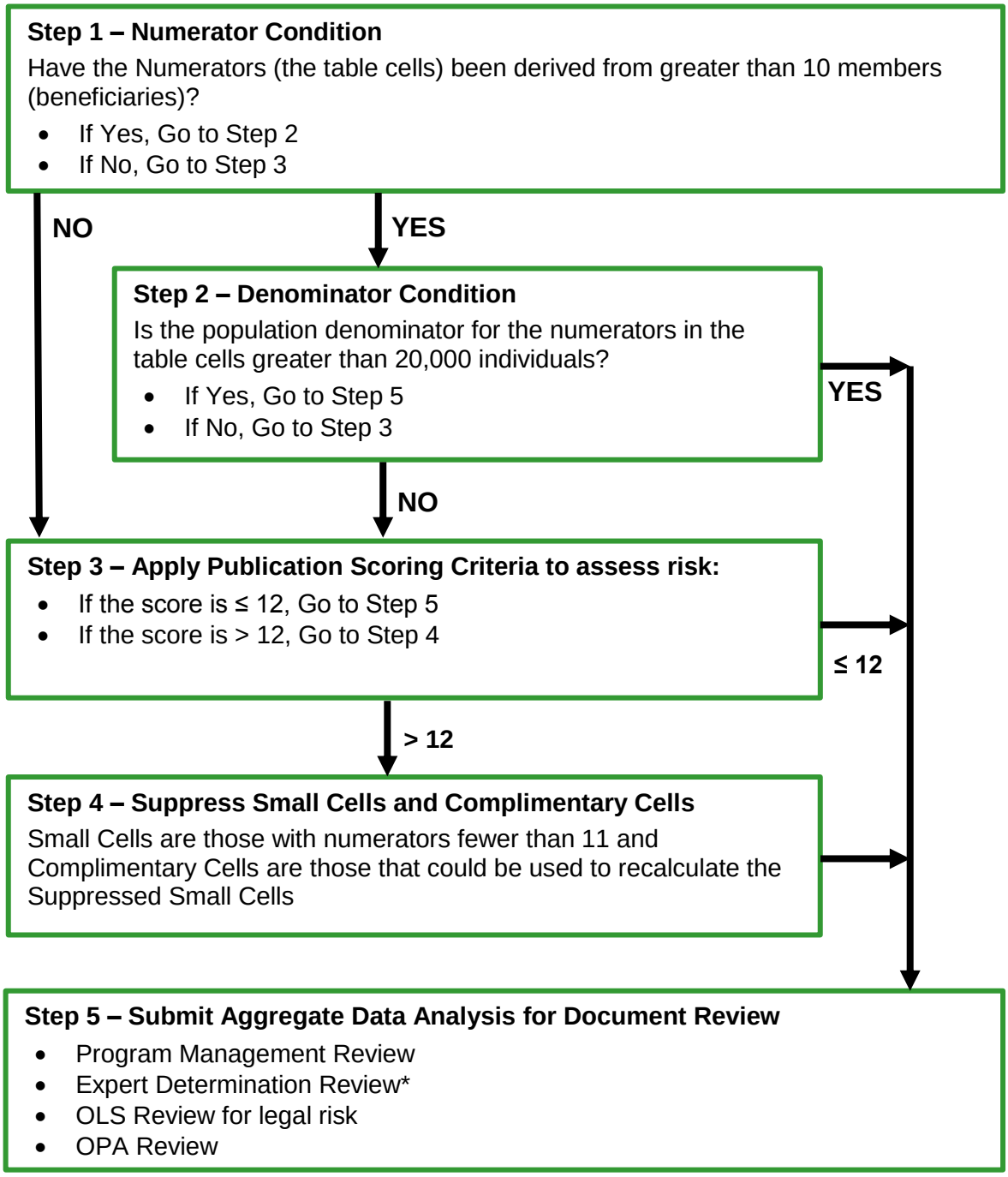
The Numerator Condition sets a lower limit for the cell size of cells displayed in the table. DHCS has set this limit as any value representing aggregated records which are derived from 10 or fewer members (beneficiaries). Most often this will translate to a minimum cell size of 11 for the numerator.

The Denominator Condition sets a minimum value for the denominator. DHCS has identified the lower limit for the denominator for Step 2 to be a value of greater than 20,000.

Since this is a Numerator-Denominator Combination 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 condition in Step 1 or Step 2 is not met, then the expert must proceed to Step 3.

Figure 3: Reporting Assessment Decision Tree

Assesses risk for data release of aggregate data through a stepwise process. Aggregate data may be derived from record level data with identifiers, record level data without identifiers or previously aggregated data.



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

Step 3

Step 3 uses the Publication Scoring Criteria in Figure 4 to assess risk for data release of aggregate data. 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 that are routinely used to publish data in the department.

A variable is a symbol representing an unknown numerical or categorical value in an equation or table.[7] 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 4 quantifies with a score two identification risks: size of potential population and variable specificity. The Publication Scoring Criteria is being used to assess the need to suppress small cells as a result of a small numerator, small denominator, or both. That is why the Publication Scoring Criteria takes into account both numerator, such as Events, and denominator, such as Geography, variables.

[7]Note: in the Business Object Tool, a "variable" is called an "object." In 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, the term "features" is used to describe "variables."

Figure 4: Publication Scoring Criteria

Variable	Characteristics	Score
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	+3
	Detailed Race	+5
Hispanic Ethnicity	yes or no	+2
	Detailed ethnicity	+3
Language Spoken	English, Spanish, Other Language	+2
Events	1000+ events in a specified population	+2
	100-999 events	+3
	11-99 events	+5
	<11 events	+8
Geography	State or geography with population >2,000,000	-5
	Population 560,001 - 2,000,000	-3
	Population 20,001 - 560,000	0
	Population ≤ 20,000	+5
Data Year	5 years aggregated	-5
	2-4 years aggregated	0
	1 year (e.g., 2001)	+3
	Bi-Annual	+4
	Quarterly	+5
	Monthly	+7

This method requires a score less than or equal to 12 for the data table to be released without additional manipulation of the data. Any score over 12 will require cell suppression for cells derived from less than 11 individuals. However, if that cannot be done, data can be published if both the small cells and complimentary cells are suppressed in Step 4.

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.

Special notes for use of the Publication Scoring Criteria
"Events" refers to the number of individuals or occurrences for which data has been collected and is being reported. The number of events should be considered for the particular geography or stratification in question. For example, in reporting the number of typhoid cases by county, the events would be the number of typhoid cases in each county. This would most likely result in cell suppression for the small counties unless other aggregation, such as combining multiple years of data, were to occur.

Step 4

Cell suppression includes two parts: suppression of the cell that has a value less than 11 and suppression of other complimentary cells that may allow for calculation of the cell being suppressed. Complimentary cells may have larger cell sizes that support calculation of the exact number of cases in cells of less than 11 by subtraction or other mathematical means.

Step 5

The final step in preparing aggregate data for publishing is to submit the aggregate data analysis to the document review process. The document review process for documents that include data analysis includes the entities listed below. The review by these entities occurs in the order below with the final review performed by the Office of Public Affairs (OPA). If any entity requests significant changes then the document will return to Program Management Review and move through the process again. The documentation associated with requested changes are to remain with the review documents.

- Program Management Review
- Expert Determination Review
- Office of Legal Services (OLS) Review
- OPA Review

Recognizing that some data analyses may be published as independent tables while other analyses will be part of a larger report, the final review of all data analyses must follow the DHCS policies for document review with the additions noted above. For each entity performing a review, a written summary that includes a recommendation to release or not will be provided. 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. Figure 5 provides a draft cover sheet to accompany documents for review.

Program Management Review:
The program management team is expected to review the table both for accuracy and to assess the publication in accordance with the DHCS HAM Chapter 3 – Communications.

Expert Determination Review:
This will be performed by individuals who have been qualified as 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 apply [i] "the generally accepted statistical and scientific principles and methods, in order to determine that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and [ii] document[] the methods and results of the analysis that justify such determination…"

DHCS intends that these Guidelines provide the starting point for expert determination review; however, the facts of each case chosen for expert determination review must be analyzed on an individual, case-by-case basis by the expert. 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 will include the Expert Determination Checklist in Figure 6. The Expert Determination Checklist 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.

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

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

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

Technical Assistance

The IMD will provide training and on-going technical assistance to programs in the implementation of these guidelines. To request assistance, please contact CMIO@dhcs.ca.gov.

Figure 5: Cover Sheet for Release of Data Products for Public Reporting

Name of Report: _____

Deadline for Publication of Data Analysis: _____

Check all that apply regarding the Data Analysis:

☐ Public Records Act; ☐ Legislative Request/Report; ☐ Media Request; ☐ Program Generated; ☐ Administration Request; ☐ External Request

Reviewer Summary	**Recommendation for Data Release**
Program Management Review *Summary:*	Yes / No Print Name: Signature:
Expert Determination Review *Summary:*	Yes / No Print Name: Signature:
Office of Legal Services Review *Summary:*	Yes / No Print Name: Signature:
Office of Public Affairs Review *Summary:*	Yes / No Print Name: Signature:

Attachments:

- ☐ Data analysis being reviewed
- ☐ Request for the analysis
- ☐ Expert Determination Checklist

Figure 6: Expert Determination Checklist for Release of Data Products for Public Reporting

Name of Report: _____

Document how the conditions of each step are met or not met	**Met / Not Met**
Step 1 – Numerator Condition *Summary:*	
Step 2 – Denominator Condition *Summary:*	
Step 3 – Results of Publication Scoring Criteria *Summary:*	
Step 4 – Suppression Performed *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"*	

The Checklist is to be completed each time the Expert Review is performed and must remain with the analyses throughout the document review process.

5) Examples: Reporting Assessment Decision Tree

Example 1

This example uses a data table currently published by the California Department of Public Health (CDPH) Office of AIDS: Blue font = numerator; **Red & Bolded font** = denominator; and **Green & Bolded font** = numerator and denominator.

Table 1. Living ǂ HIV and AIDS cases by gender, race/ethnicity, current age and exposure category, 2009

	HIV (non-AIDS)*			AIDS**			HIV/AIDS***		
	N	%	Rate ǂ	N	%	Rate ǂ	N	%	Rate ǂ
GENDER									
Male	31,052	86.1%	160.8	59,161	87.5%	306.4	90,213	87.0%	467.3
Female	4,595	12.7%	23.7	7,680	11.4%	39.6	12,275	11.8%	63.3
Transgender	406	1.1%	-	799	1.2%	-	1,205	1.2%	-
RACE/ETHNICITY									
White	17,236	47.8%	104.9	30,287	44.8%	184.3	47,523	45.8%	289.2
Black	6,454	17.9%	283.2	12,487	18.5%	547.9	18,941	18.3%	831.1
Latino	10,394	28.8%	73.3	21,586	31.9%	152.2	31,980	30.8%	225.5
Asian/PI	1,337	3.7%	28.2	2,316	3.4%	48.8	3,653	3.5%	77
AI/AN	165	0.5%	70.1	280	0.4%	118.9	445	0.4%	189
Multirace	467	1.3%	57.5	664	1.0%	81.8	1,131	1.1%	139.3
CURRENT AGE									
< 13	130	0.4%	1.8	50	0.1%	0.7	180	0.2%	2.6
13-19	298	0.8%	7.1	223	0.3%	5.3	521	0.5%	12.4
20-29	4,862	13.5%	89.4	2,772	4.1%	51	7,634	7.4%	140.4
30-39	8,826	24.5%	169.3	10,607	15.7%	203.4	19,433	18.7%	372.7
40-49	12,614	35.0%	220.2	27,794	41.1%	485.3	40,408	39.0%	705.5
50+	9,323	25.9%	84.2	26,194	38.7%	236.7	35,517	34.3%	320.9
EXPOSURE CATEGORY									
MSM	24,765	68.7%	-	43,195	63.9%	-	67,960	65.5%	-
IDU	2,085	5.8%	-	5,977	8.8%	-	8,062	7.8%	-
MSM/IDU	2,378	6.6%	-	6,126	9.1%	-	8,504	8.2%	-
Heterosexual	3,189	8.8%	-	6,475	9.6%	-	9,664	9.3%	-
Blood/other	72	0.2%	-	384	0.6%	-	456	0.4%	-
NIR/NRR	3,293	9.1%	-	5,168	7.6%	-	8,461	8.2%	-
Perinatal	271	0.8%	-	315	0.5%	-	586	0.6%	-
TOTAL	**36,053**	**100.0%**	**93.2**	**67,640**	**100.0%**	**174.8**	**103,693**	**100.0%**	**268.0**

Case reported by Feb 22, 2012 and diagnosed by Dec 31, 2009

ǂ Includes persons currently living with HIV infection (regardless of diagnosis state) whose last known residence is in California

*Cases which remained HIV cases on Feb 22, 2012

***The sum of HIV cases and AIDS cases

ǂ Rates per 100,000 persons are based on 2009 population estimates from the California Department of Finance

Dash (-) indicates the rate could not be calculated due to unknown population denominators

There were 19 cases with unknown race/ethnicity

Source: California HIV/AIDS Epidemiological Profile, 2009 Update

In this case, there are no numerators less than 11 and the denominators are greater than 20,000. This is expected given the data is published for the entire state and has not been stratified based on sub-state geographic categories. Since Step 1 and Step 2 conditions are both met, proceed to Step 5: Review.

Example 2

This example uses a data table currently published by the CDPH Office of AIDS: Blue font = numerator; and **Red & Bolded font** = denominator.

Step 1 and Step 2 – In this case there are numerators less than 11 and denominators less than 20,001. Since the condition for Step 1 is not met, proceed to Step 3.

Table 3. Living HIV infection cases by race/ethnicity and current age group

	Age Group (Yrs)													
	0-12		13-19		20-29		30-39		40-49		50-59		60+	
RACE/ETHNICITY	N	%	N	%	N	%	N	%	N	%	N	%	N	%
White (N=47,523)	22	0.1	85	0.2	1,857	3.9	6,215	13.1	19,105	40.2	14,507	30.5	5,732	12.1
Black (N=18,941)	55	0.3	197	1.0	1,806	9.5	3,323	17.5	7,010	37.0	5,024	26.5	1,526	8.1
Latino (N=31,980)	88	0.3	220	0.7	3,475	10.9	8,540	26.7	12,365	38.7	5,558	17.4	1,734	5.4
Asian/PI (N=3,653)	7	0.2	10	0.3	331	9.1	1,014	27.8	1,310	35.9	724	19.8	257	7.0
AI/AN (N=445)	0	0.0	1	0.2	26	5.8	93	20.9	198	44.5	102	22.9	25	5.6
Multirace (N=1,131)	8	0.7	7	0.6	139	12.3	248	21.9	413	36.5	253	22.4	63	5.6

N=19, Unknown race/ethnicity

Step 3: Publication Scoring Criteria – This table is for the State of California, giving the Geography a score of -5, and for one year of data, giving the Data Year a score of +3. The table includes Detailed Race categories which result in a score of +5; Hispanic/Ethnicity categories (Latino Yes or No) which result in a score of +2; and Age categories (6-10 year age groups) which result in a score of +3. Additionally, the number of events in a given table cell is less than 11. This is scored with the "Event" variable and results in a score of +8. The overall score (-5 +3 +5 +2 +3 +8 = 16) is 16. Since the overall score is greater than 12, this table would not be released to the public without suppression of the table cells that are less than 11 or without further evaluation. This table has been released publicly by CDPH since the risk of re-identifying the one American Indian/American Native between the ages of 13-19 living with HIV in California is very small based on this table.

Example 3

DHCS routinely seeks to provide information publicly by county in DHCS business reports. This example will specifically address considerations that must be considered. California counties vary in population from approximately 1,100 people in Alpine County to almost 10 million people in Los Angeles County. Many California counties are larger than the state with the smallest population, which is Wyoming, with approximately 560,000 people. Counties in bold below have populations less than 20,001.

County	Estimated Population
Alameda	1510272
Alpine*	**1175**
Amador	38091
Butte	220000
Calaveras	45578
Colusa	21419
Contra Costa	1049025
Del Norte	28610
El Dorado	181058
Fresno	930450
Glenn	28122
Humboldt	134623
Imperial	174528
Inyo*	**18546**
Kern	839631
Kings	152982
Lake	64665
Lassen	34835
Los Angeles	9818609
Madera	150865
Marin	252409
Mariposa*	**18251**
Mendocino	87841
Merced	255793
Modoc*	**9686**
Mono*	**14202**
Monterey	415057
Napa	136484
Nevada	98764
Orange	3010232
Placer	348432

County	Estimated Population
Plumas	20007
Riverside	2189640
Sacramento	1418789
San Benito	55269
San Bernardino	2035209
San Diego	3095313
San Francisco	805236
San Joaquin	685306
San Luis Obispo	269637
San Mateo	718452
Santa Barbara	423895
Santa Clara	1781643
Santa Cruz	262382
Shasta	177223
Sierra*	**3240**
Siskiyou	44900
Solano	413344
Sonoma	483878
Stanislaus	514453
Sutter	94737
Tehama	63463
Trinity*	**13786**
Tulare	442179
Tuolumne	55365
Ventura	823318
Yolo	200849
Yuba	72155

The county estimates above are from the California Department of Finance. When applying Step 2, the Denominator Condition, to California county populations, there are seven counties with less than 20,001. This can be addressed by combining bordering counties to create larger populations. This is demonstrated in the examples below.

Combine Alpine with Bordering Counties		
Alpine + Mono =	15377	too small
Alpine + Mono + Inyo =	**29579**	**>20000**
Alpine + El Dorado =	182233	>20000
Alpine + Amador =	**39266**	**>20000**
Alpine + Calaveras =	46753	>20000
Alpine + Tuolumne =	56540	>20000

Combine Inyo with Bordering Counties		
Inyo + Mono=	**28404**	**>20000**
Inyo + Fresno =	944652	>20000
Inyo + Tulare =	456381	>20000
Inyo + Kern =	853833	>20000
Inyo + San Bernardino =	2049411	>20000

Combine Mariposa with Bordering Counties		
Mariposa + Tuolumne =	**73616**	**>20000**
Mariposa + Merced =	274044	>20000
Mariposa + Madera =	169116	>20000

Combine Modoc with Bordering Counties		
Modoc + Siskiyou =	54586	>20000
Modoc + Shasta =	186909	>20000
Modoc + Lassen =	**49037**	**>20000**

Combine Mono with Bordering Counties		
Mono + Alpine =	15377	too small
Mono + Tuolumne =	69567	>20000
Mono + Madera =	165067	>20000
Mono + Fresno =	944652	>20000
Mono + Inyo =	**28404**	**>20000**

Combine Sierra with Bordering Counties		
Sierra + Lassen =	38075	>20000
Sierra + Plumas =	**23247**	**>20000**
Sierra + Yuba =	75395	>20000
Sierra + Nevada =	102004	>20000

Combine Trinity with Bordering Counties		
Trinity + Siskiyou =	**58686**	**>20000**
Trinity + Shasta =	191009	>20000
Trinity + Tehama =	77249	>20000
Trinity + Mendocino =	101627	>20000
Trinity + Humboldt =	148409	>20000

Alternately, for tables for which the program wants to publish data by county, the Publication Scoring Criteria may be used to assess the risk of a given table.

Example 4

As an example, for a table that includes the number of Medi-Cal members per county in a given year, the parts of the Publication Scoring Criteria that would apply are:

Events	1000+ events for geography	+2
	100-999 events	+3
	11-99 events	+5
	< 11 events	+8
Geography	State or geography with population >2,000,000	-5
	Population 560,001 - 2,000,000	-3
	Population 20,001 - 560,000	0
	Population ≤ 20,000	+5
Data Year	5 years aggregated	-5
	2-4 years aggregated	0
	1 year (e.g., 2001)	+3
	Bi-Annual	+4
	Quarterly	+5
	Monthly	+7

The number of Medi-Cal members per county per year would be scored based on the smallest number. Alpine County has approximately 190 Medi-Cal members which results in a score of +3 for the Events variable. There are seven counties with a population less than 20,001, giving the Geography variable a score of +5. It is a report for a given year, so the Data Year variable results in a score of +3. When combined (+3 +5 +3 = +11), the overall score is +11. That means this example can be published without suppression even though the Denominator Condition was not met.

6) Justification of Thresholds Identified

6.1 Establishing Minimum Numerator

DHCS 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 CMS uses a small cell policy of suppressing values derived from fewer than 11 beneficiaries. As stated in a recent 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]." [8] Given that DHCS submits data routinely to CMS regarding the Medi-Cal program and this data is published publicly by CMS through the Medicaid Statistical Information System, DHCS has decided to align with CMS for the Numerator Condition. Of note, CMS only uses a Numerator Condition.

6.2 Establishing Minimum Denominator

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, DHCS 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 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 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

[8] "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.

States.[9,10] This means that over half of U.S. residents could be uniquely described just with these three data elements. In contrast, it has been estimated that the combination of Year of Birth, Sex, and 3-Digit ZIP Code is unique for approximately 0.04% of residents in the United States.[11] 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 a 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 gender 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.

While the above specifics describe the potential uniqueness using the 3-Digit ZIP Code, DHCS does not routinely publish at the ZIP Code level. Typically, programs and stakeholders ask DHCS to publish data by county or region. Except for California's smallest counties, county level aggregation is less specific than the 3-Digit ZIP Code. Therefore, DHCS is applying the population level for the denominator of 20,000 as opposed to identifying a specific type of geo-political boundary.

6.3 Publication Scoring Criteria Methodology

As noted in the decision tree, DHCS does not intend to rely on one method for aggregation but rather a combination of methods. This is reflected in the Publication Scoring Criteria, which 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 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.

[9] 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.
[10] See L. Sweeney. K-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*. 2002; 10(5): 557-570.
[11] See L. Sweeney. Testimony before that National Center for Vital and Health Statistics Workgroup for Secondary Uses of Health information. August 23, 2007.

Sex

Variable	Characteristics	Score
Sex	Male or Female	+1

Sex is either Male or Female. There are only two choices for this variable making it a low risk variable for re-identification. The score of +1 reflects that inclusion of the variable in a table introduces increased specificity; however, that it only has two responses gives it a low risk.

Although the variable “Sex” is often called “Gender”, it should not be confused with the variable “Gender” which may have values such as Lesbian, Gay, Bi-sexual, Transgender. If a program wishes to display “Gender” classification, as opposed to “sex,” then it will need to be reviewed on a case by case basis.

Of note, “Sex” is not listed as one of the 18 identifiers in the HIPAA Safe Harbor method.

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 Hispanic Ethnicity

Variable	Characteristics	Score
Race Group	White, Asian, Black	+3
	Detailed Race	+5
Hispanic Ethnicity	yes or no	+2
	Detailed ethnicity	+3

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 Medi-Cal application prior to 2014 collected "Race/Ethnicity" as one field. It did not distinguish between Race and Ethnicity and it was a voluntary data field that was typically completed between 90 and 95% of the time. Beginning in 2014, the Medi-Cal application asks for Race and Ethnicity as two separate questions, which are still voluntary fields.

The California population in general is approximately:[12]

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

Based on these percentages, Race Group at the level of White, Asian and Black is given a score of +3 because the Asian and Black groups are relatively small. If more

[12] 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

specificity is requested for Race Groups the score is +5 because the other groups are much smaller at less than 1% of the overall population. Similarly, for the Hispanic Ethnicity the score is a +2 for a yes or no answer, whereas more detailed ethnicity results in a higher score of +3.

Due to the voluntary reporting nature of this field, the data is inherently masked due to the absence of data for a portion of records.

For tables in which there are small numbers due to the specificity added by either Detailed Race or Detailed Ethnicity, the Events score may also come into play due to the small numerators that can be associated with the Detailed Race and Detailed Ethnicity variables.

Of note, “Race” and “Ethnicity” are not listed as one of the 18 identifiers in the HIPAA Safe Harbor method.

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.

	Race/Ethnicity							
State/County	**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

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
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
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

Language Spoken

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

Language spoken is captured on the applications for services to support members 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.[13]

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
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

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

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 Ethnicity as a “Yes or No”.

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.

Of note, “Language Spoken” is not listed as one of the 18 identifiers in the HIPAA Safe Harbor method.

Events

Variable	Characteristics	Score
Events	1000+ events for geography	+2
	100-999 events	+3
	11-99 events	+5
	<11 events	+8

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 (+8) was chosen to be consistent with the Numerator Condition. The Publication Scoring Criteria is used when the Numerator Condition, Denominator Condition or both are not met. Therefore, when the Numerator Condition is not met with respect to the Events variable, a high score is given.

Of note, “Events” is not listed as one of the 18 identifiers in the HIPAA Safe Harbor method.

Geography

Variable	Characteristics	Score
Geography	State or geography with population >2,000,000	-5
	Population 560,001 - 2,000,000	-3
	Population 20,001 - 560,000	0
	Population ≤ 20,000	+5

The Geography score represents a score for the denominator. This will often be reflected in the title of the table if a statewide table. Otherwise the geography will be

represented in the rows or columns. 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 cut-off of 20,000 for the score of "+5" was chosen to be consistent with the Denominator Condition described above. The Publication Scoring Criteria is used when the Numerator Condition, Denominator Condition or both are not met. Therefore, when the Denominator Condition is not met with respect to the Geography variable, a high score is given.

Data Year

Variable	Characteristics	Score
Data Year	5 years aggregated	-5
	2-4 years aggregated	0
	1 year (e.g., 2001)	+3
	Bi-Annual	+4
	Quarterly	+5
	Monthly	+7

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 +7.

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.

Other Variables

Variables other than those specified in the Publication Scoring Criteria can be released only after an additional review by the Chief Medical Information Officer (CMIO) on a case by case basis. A guideline that can be considered in performing this review is the following scoring.

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-9 Code). Also consider the availability of the variable to the public when also associated with other information, in particular with potentially identifying information.

7) Development Process

The PAR-GDP was initiated by Karen Johnson, Chief Deputy Director Policy and Program Support of DHCS, who has served as the Executive Sponsor of the PAR-GDP. Project management support of the project has been provided by Dr. Linette Scott, CMIO and Deputy Director for the IMD. The project initially convened a Review Team which included senior management from the IMD, Research and Analytic Studies Division, OLS, and Information Technology Services Division (ITSD). Specifically, the CMIO and Office of HIPAA Compliance within the IMD, the Privacy Team within OLS and the Information Security Office within ITSD were represented on the Review Team. The Review Team met twice to review and finalize the Project Charter with the Executive Sponsor.

The Development Team was convened following the finalization of the Project Charter. The invitation to participate in the Development Team was distributed broadly to the department to ensure broad representation. Many individuals participating in the Development Team also have been members of the Data and Research Committee, which reviews requests from external researchers and public health for DHCS data. The following divisions were represented on the Development Team:

- Audits & Investigations,
- Benefits Division,
- Clinical Assurance and Administrative Support Division,
- Director's Office,
- Eligibility Division,
- Information Management Division,
- Information Technology Services Division,
- Medi-Cal Managed Care Division,
- Mental Health Division,
- Office of Family Planning,

- Office of HIPAA Compliance,
- Office of Legal Services,
- Office of the Medical Director,
- Office of Public Affairs,
- Pharmacy Division,
- Primary Rural and Indian Health Division,
- Research and Analytic Services Division (RASD), and
- Substance Use Disorder (SUD) Prevention, Treatment & Recovery Services Division.

The first four meetings of the Development Team included participation and presentations from external departments, including:

- CDPH – Center for Health Statistics and Informatics (CDPH-CHSI), and
- Office of Statewide Health Planning and Development (OSHPD).

The Development Team meetings were scheduled such that the first four meetings provided an overview of statistical methods, examples from other states and examples from other departments and programs in California. Documents and literature reviewed in the development of the presentations for the Development Team meetings are listed in the Reference section of this document. These were well attended with over 30 participants in the room and via webinar with discussion of the pros and cons of different methods. The last three meetings were restricted to DHCS and focused on draft guidelines for DHCS. There continued to be robust conversation from programs with questions and examples about both content and implementation of the Guidelines. The schedule of Development Team meetings is below:

4/3 – Broad overview of the PAR-GDP was given to participants

4/10 – Presentations of aggregate reporting by UCLA California Health Interview Survey, Washington State and Illinois Department of Public Health

4/24 – Presentation of aggregate reporting by OSHPD, CDPH and RASD

4/28 – Presentation on SAS Statistical Aggregation

5/6 – Draft Aggregate Reporting methods presented

5/14 – Update of Draft Aggregate Reporting Guidelines reviewed

5/22 – Update of Draft Aggregate Reporting Guidelines reviewed

The development process was designed to include a literature review, case examples and broad discussion among DHCS programs. Publishing data publicly is always a

balance between the protection of confidentiality and the usability of the data. This was a theme throughout the conversations of the Development Team.

The final Draft Aggregate Reporting Guidelines from the Development Team were then sent to the Review Team for review. A particular area of focus for the Review Team was on the Document Review and Approval Process for the department. This was initially out of scope for the PAR-GDP but there was agreement that the review for documents with data analyses needed more specific reviews. This was addressed by adding detail to Step 5 of the Reporting Assessment Decision Tree. Additionally, clarification was provided indicating that the Guidelines are to be used for DHCS Business Reports only. Assessment of aggregate data for potential release in response to external requests (e.g., but not limited to, Public Records Act (PRA) requests), is subject to the Public Aggregate Reporting - For External Requests (PAR-FER). External requests tend to seek data constructions that, while aggregate, may be accompanied by higher levels of identification risk than in the DHCS business reports due to cumulative identification risk, information already available in the public domain, or information possessed by the requestor. As such, aggregation processes for these external requests are governed by the PAR-FER.

8) Abbreviations and Acronyms

CDC ..... Centers for Disease Control and Prevention
CDPH..... California Department of Public Health
CMIO..... Chief Medical Information Officer
CMS ..... Centers for Medicare and Medicaid Services
DHCS ..... Department of Health Care Services
HAM ..... Health Administrative Manual
HIPAA ..... Health Insurance Portability and Accountability Act
IMD..... Information Management Division
ITSD..... Information Technology Services Division
OLS..... Office of Legal Services
OPA..... Office of Public Affairs
OSHPD ..... Office of Statewide Health Planning and Development
PAR-DBR…. Public Aggregate Reporting - DHCS Business Reports
PAR-FER…..Public Aggregate Reporting - For External Requests
PAR-GDP…. Public Aggregate Reporting – Guidelines Development Project
PHI ..... Protected Health Information
PI..... Personal Information
RASD ..... Research and Analytic Services Division

9) 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).

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

10) Legal Framework

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

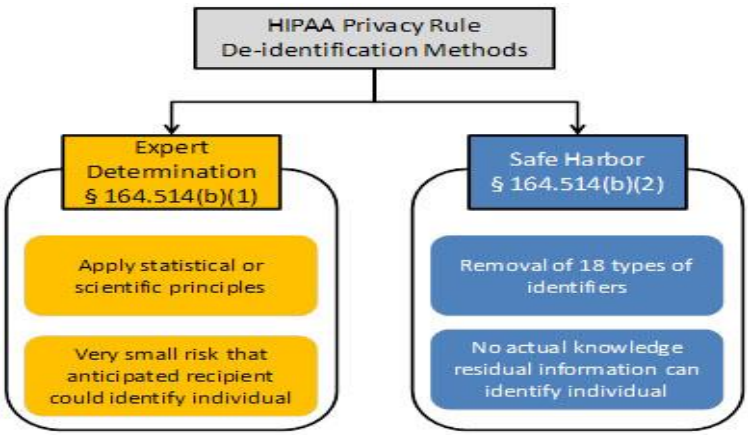
The De-identification Standard

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.

11) 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

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, 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.

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.

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.

12) Approvals

We have reviewed the document, “Public Aggregate Reporting – Guidelines,” and hereby approve it as our official position for DHCS business reports.

Original Signed By _____ Date 8/27/14

Toby Douglas, Director, DHCS

Original Signed By _____ Date 8/26/14

Karen Johnson, Chief Deputy Director Policy and Program Support, DHCS

Original Signed By _____ Date 8/25/14

Douglas Press, Chief Counsel and Deputy Director, Office of Legal Services, DHCS

Original Signed By _____ Date 8/25/14

Linette Scott, MD, MPH, Chief Medical Information Officer and Deputy Director, Information Management Division, DHCS