

California Department of Public Health
Office of AIDS

CDPH/OA Internal Guidelines for Working with Small Cell Sizes

This document provides guidance to CDPH/OA researchers for working with small cell sizes to ensure client privacy and data confidentiality.

September 2014



Purpose

The California Department of Public Health, Office of AIDS (OA) collects and analyzes HIV prevention, care, treatment, and surveillance data provided by local health jurisdictions (LHJs) and community based organizations (CBOs) throughout the state of California. These data are used both internally (for quality assurance, assessment of performance, etc.) and externally (for assessment of performance, reports to policymakers and stakeholders, federal and state data submissions, etc.).

As a source for program data, OA must balance providing data to the public, stakeholders, and policymakers while simultaneously protecting client privacy and data confidentiality. In order to achieve this balance, OA must assess risks associated with small cell sizes. Because this document is intended to serve as internal guidance for working with small cell sizes to prevent disclosure of potentially identifying information, unrelated small cell size topics (e.g., rates) will not be discussed.

Table of Contents

A. General Guidelines and Examples.....	3
1. Small or Sensitive Populations	3
2. Aggregation (Collapsing Variable Values).....	3
3. Primary & Secondary Cell Suppression (Withholding Data from Cells).....	5
4. The Golden Rule	6
5. Dealing with a Periodic Publication.....	9
6. Creating Multiple Tables from the Same Data Set.....	10
B. Statewide-Level Guidelines and Examples.....	12
1. Unstratified	12
2. Stratified by One Variable.....	12
3. Stratified by More Than One Variable	16
C. County-Level Guidelines and Examples.....	16
1. No Further Breakdown than County	16
2. Stratified.....	17

All data presented in this document are fictional and for illustration purposes only.

A. General Guidelines and Examples

1. Small or Sensitive Populations

A.1.a. When dealing with data concerning small and/or sensitive populations (e.g., race/ethnicity of HIV positive transgender youth in Sacramento County), determine if the data are potentially identifying. Typically cells containing zero cases are not potentially identifying, but cells containing 1–4 cases are potentially identifying (see Example 1).

Example 1: Potentially Identifying Data Due to Small or Sensitive Populations

Table: Number of Transgender Youth by Race/Ethnicity, CY 2012 – California (Sacramento County)

	Black/ African American	White	Hispanic/ Latino	American Indian/ Alaska Native	Other*
Transgender (0-12 years)	4	8	0	3	7
Transgender (13-18 years)	5	10	9	5	9

* Other race/ethnicity is defined as Asian and Native Hawaiian/Pacific Islander

The cells highlighted in red above may be identifying, especially considering the small number of transgender youths in Sacramento County. Cells with counts of 1-4 can be collapsed (see Solutions 1a and 1b) or suppressed (see Solution 1c) to prevent inadvertent disclosure of potentially identifying information. Include a legend below your table explaining why data were collapsed or suppressed.

2. Aggregation (Collapsing Variable Values)

A.2.a. When possible, collapse instead of suppress.

In the example above, there are two possible methods of collapsing the cells: (1) collapsing the Black/African American and American Indian/Alaska Native columns into the “Other” race column (see Solution 1a below) or (2) collapsing the 0-12 age group with the 13-18 age group (see Solution 1b below).

Solution 1a: Aggregating Potentially Identifying Data Due to Small or Sensitive Populations by Collapsing the Race/Ethnicity Columns

Table: Number of Transgender Youth by Race/Ethnicity, CY 2012 – California (Sacramento County)

	White	Hispanic/Latino	Other*
Transgender (0-12)	8	0	14
Transgender (13-18)	10	9	19

** Black/African American, American Indian/Alaska Native, and Other Race/Ethnicity have been collapsed to protect confidentiality*

Solution 1b: Aggregating Potentially Identifying Data Due to Small or Sensitive Populations by Collapsing the Age Group Rows

Table: Number of Transgender Youth by Race/Ethnicity, CY 2012 – California (Sacramento County)

	Black/African American	White	Hispanic/Latino	American Indian/Alaska Native	Other*
Transgender (0-18)	9	18	9	8	16

** The 0-12 and 13-18 age groups have been collapsed to protect confidentiality*

3. Primary & Secondary Cell Suppression (Withholding Data from Cells)

A.3.a. If collapsing cells is not a viable option, suppression can be used to prevent inadvertent disclosure of potentially identifying information (see Solution 1c below).

Solution 1c: Suppressing Potentially Identifying Data Due to Small or Sensitive Populations

Table: Number of Transgender Youth by Race/Ethnicity, CY 2012 – California (Sacramento County)

	Black/African American	White	Hispanic/Latino	American Indian/Alaska Native	Other
Transgender (0-12)	*	8	0	*	7
Transgender (13-18)	5	10	9	5	9

** Data have been suppressed to protect confidentiality due to small cell sizes*

A cell count threshold defines a value or value ranges. Only cells that did not meet the cell count threshold of 1-4 cases were suppressed (primary suppression) because the example above does not include a Total row and/or column and arithmetic cannot be used to derive potentially identifying information. Extra caution should be taken when potentially identifying cells are located in a table with a row and/or column total, as it may be possible for someone to subtract a cell count from the total and thus derive the cell count of suppressed cells.

A second-round of suppression (secondary suppression) is necessary when arithmetic can be used to derive the value of suppressed cells; when possible, suppress cells with the lowest number of cases. See the Golden Rule in section A.4.a for additional information on secondary suppression.

A.3.b. “Other” categories (e.g., Other race/ethnicity) are not potentially identifying unless crossed-tabbed with a small or sensitive group (e.g., transgender individuals). Suppression is optional when a cell representing an “Other” category contains 1-4 cases and is not cross-tabbed with a small or sensitive group (see Example 2).

Example 2: “Other” Category Less than Threshold – No Small or Sensitive Populations

Table: Number of Individuals Tested for HIV at OA-Funded Mobile HIV Testing Events by Gender and Race/Ethnicity, May 2012 – California

	Black/African American	White	Hispanic/Latino	American Indian/Alaska Native	Other
Male	25	39	20	16	18
Female	14	17	13	10	3
Transgender	0	0	0	0	0

Even though the cell highlighted in yellow has a cell count of 1-4, suppression is not required because the combination of “Female” gender and “Other” race/ethnicity are not considered to be potentially identifying characteristics.

4. The Golden Rule

A.4.a. When secondary suppression is required to prevent arithmetic from being used to derive suppressed cell counts, follow The Golden Rule.

The Golden Rule

1. Each column or row with an unknown (aka suppressed cell) must have a minimum of two unknowns.
2. The total number of unknowns must exceed the number of equations the unknowns reside in; each column or row with an unknown counts as one equation.
3. You may need to suppress cells that are greater than 5 to fulfill rules 1 and 2 above.

The table in Example 3 requires secondary suppression to prevent inadvertent disclosure of potentially identifying information.

Example 3: Potentially Identifying Data Due to Small or Sensitive Populations – Original Table Requiring Application of the Golden Rule (No Suppression)

Table: Number of New HIV Cases by Age Group and Race/Ethnicity, CY 2012 – California (San Diego County)

Age Group	Asian	Black/African American	Hispanic/Latino(a)	White	American Indian/AN	Total
0-12	3	4	5	25	3	40
13-19	7	29	8	40	4	88
20-29	23	20	25	46	15	129
30+	20	45	50	81	10	206
Total	53	98	88	192	32	463

Examples 3a and 3b, below, illustrate the incorrect and correct application of the Golden Rule.

Example 3a: Potentially Identifying Data Due to Small or Sensitive Populations – Incorrect Use of the Golden Rule (Primary Suppression Only)

Table: Number of New HIV Cases by Age Group and Race/Ethnicity, CY 2012 – California (San Diego County)

Age Group	Asian	Black/African American	Hispanic/Latino(a)	White	American Indian/AN	Total
0-12	*	*	5	25	*	40
13-19	7	29	8	40	*	88
20-29	23	20	25	46	15	129
30+	20	45	50	81	10	206
Total	53	98	88	192	32	463

* Cells have been suppressed to protect confidentiality.

Example 3a, above, contains **5 equations** and **4 unknowns**. The Golden Rule has NOT been fulfilled because the number of unknowns doesn't exceed the number of equations and each row and column with an unknown doesn't contain at least 2 unknowns. The number

of Asian and Black/African Americans in the 0-12 age group can be derived via subtraction from the race totals. These issues have been corrected in Example 3b, below.

Example 3b: Potentially Identifying Data Due to Small or Sensitive Populations – Correct Use of the Golden Rule (Primary and Secondary Suppression)

Table: Number of New HIV Cases by Age Group and Race/Ethnicity, CY 2012 – California (San Diego County)

Age Group	Asian	Black/African American	Hispanic/Latino(a)	White	American Indian/AN	Total
0-12	*	*	5	25	*	40
13-19	*	*	8	40	*	88
20-29	23	20	25	46	15	129
30+	20	45	50	81	10	206
Total	53	98	88	192	32	463

** Cells have been suppressed to protect confidentiality*

This example contains **5 equations** and **6 unknowns** and each row and column with an unknown has at least 2 unknowns; the Golden Rule has been fulfilled

5. Dealing with a Periodic Publication

A.5.a. When dealing with a periodic publication (e.g., quarterly, annual, etc.), ensure potentially identifying information cannot be derived by comparing old reports to new reports. This rule applies when data are cumulative and only the subsequent period has been added to the table (e.g., the CY 2013 report is data from the CY 2012 report plus CY 2013 data and no changes were made to the pre-CY 2013 data). See Example 4 below.

Example 4: Potentially Identifying Data Due to Periodic Publication Containing Cumulative Data

Table 1: Number of Cumulative Living HIV Cases by County and Gender, CY 2012 – California

	Male	Female	Transgender
Alameda	1,000	545	5
...
Modoc	9	0	*

Table 2: Number of Cumulative Living HIV Cases by County and Gender, CY 2013 (Table 1 + CY 2013 Data) – California

	Male	Female	Transgender
Alameda	1,020	555	6
...
Modoc	9	0	*

** This cell has been suppressed to protect confidentiality*

Neither Table 1 nor Table 2 alone is potentially identifying; the Modoc County transgender cell count in Tables 1 and 2 are suppressed because the cell count is 1-4 cases. However, since Table 2 was constructed by adding CY 2013 data to the existing CY 2012 table, it is possible to derive the number of CY 2013 transgender cases in Alameda County (6 total transgender cases in Alameda County in CY 2013 – 5 transgender cases in Alameda County in CY 2012= 1 transgender case added in CY 2013).

To prevent the disclosure of potentially identifying information, data from previous years should be updated in the current cumulative report (e.g., when creating the CY 2013 report,

use updated pre-CY 2013 data as opposed to simply adding CY 2013 data to data from the CY 2012 report). If data from previous years cannot be updated in the current cumulative report, discuss privacy/confidentiality risks with management.

6. Creating Multiple Tables from the Same Data Set

A.6.a. When creating multiple tables from the same dataset, make sure arithmetic cannot be used to derive or infer potentially identifying information (see Example and Solution 5 below).

Example 5: Potentially Identifying Data Due to Small or Sensitive Populations and Multiple Tables

Table 1: Number of Cumulative HIV Cases by County and Mortality Status, CY 2012 – California

	Total Cases	Living Cases	Deceased
Alameda	1,600	1,550	50
...
Modoc	10	10	0

Table 2: Number of Cumulative HIV Cases by County and Gender, CY 2012 – California

	Male	Female	Transgender
Alameda	1,000	545	5
...
Modoc	9	0	*

** This cell has been suppressed to protect confidentiality*

Neither Table 1 nor Table 2 alone is potentially identifying; Table 1 does not contain cell counts of 1-4 cases and the Transgender cell count in Table 2 is suppressed because the cell count is of 1-4 cases.

When compared, however, the number of transgender cases in Modoc County can be derived by subtracting the number of males and females in Modoc County from the total number of cases in the county (10 total – 9 males – 0 females = 1 transgender). To prevent the disclosure of potentially identifying information, another cell should be suppressed in the row representing Modoc County in Table 2 (see Solution 5 below).

Solution 5: Potentially Identifying Data Due to Small or Sensitive Populations and Multiple Tables – Suppress More than One Cell

Table 1: Number of Cumulative HIV Cases by County and Mortality Status, CY 2012 – California

	Total Cases	Living Cases	Deceased
Alameda	1,600	1,550	50
...
Modoc	10	10	0

Table 2: Number of Cumulative HIV Cases by County and Gender, CY 2012 – California

	Male	Female	Transgender
Alameda	1,000	545	5
...
Modoc	9	*	*

** Cells have been suppressed to protect confidentiality*

In the example above, the cell representing females in Modoc County was suppressed because it prevents arithmetic from being used to determine the gender of the remaining case, while allowing the majority of Modoc’s data to be displayed.

B. Statewide-Level Guidelines and Examples

1. Unstratified

B.1.a. Suppression rules are not needed for statewide data with no further breakdown (see Example 6).

Example 6: Statewide Data with No Further Breakdown

Table: Number HIV Test Events and Newly-Identified Confirmed HIV-Positive Test Events Conducted with Category B Funds, CY 2012 – California

Total Category B Test Events January 2012	Newly-Identified Confirmed HIV-Positive Test Events
800	4

2. Stratified by One Variable

B.2.a. Suppression rules do not apply to statewide data broken down by one of the following demographics (see Example 7):

- Gender
- Race/Ethnicity
- Age Group
- Risk Group

Example 7: Statewide Data Stratified by One Variable (Gender)

Table: Number OA-Funded HIV Test Events and Newly-Identified Confirmed HIV-Positive Test Events Conducted by Gender, January 1st - June 30th 2012 – California

	Total Tests Events	Newly-Identified Confirmed HIV-Positive Test Events
Male	12,300	105
Female	5,300	11
Transgender	400	4

Even though the cell highlighted in green has a cell count of 1-4 cases, the information is likely not a threat to confidentiality because the population denominator reflects the statewide transgender population and there is no further breakdown by demographics.

B.2.b. If broken down by County/LHJ only, suppression rules do not apply to cumulative and living cases so long as they are not new cases only (see Example 8). If the number of new cases can be derived by comparing multiple reports, refer to rule A.6.a and discuss privacy/confidentiality risks with management.

Example 8: Statewide Data Stratified by One Variable (County)

Table: Newly-Identified Confirmed HIV-Positive Living Cases by County, CY 2013 – California

	Number of Living Cases (previous + new)
Alameda	40
Contra Costa	30
Modoc	4
...	...

Even though the cell highlighted in green has a cell count of less than 5 cases, the information is not likely a threat to confidentiality because the population denominator reflects the county population and there is no further breakdown by demographics. If however, the number of new cases can be derived using other reports/publications, action may be necessary to protect confidentiality; refer to rule A.6.a and discuss privacy/confidentiality risks with management.

B.2.c. If broken down by zip-code, suppress cell sizes that are 1-4 cases (see Example and Solution 9). If OA staff are able to establish the US Census population denominator for the zip code(s) in question and the population denominator minus the cell numerator is greater than 100, staff may be able to leave the cell unsuppressed; however, staff must first discuss the risks for individual identification with their supervisor and obtain supervisor approval.

Example 9: Statewide Data Broken Down by Zip Code

Table: Newly-Identified Confirmed HIV-Positive Test Events by Zip Code, CY 2013 – California

	Newly-Identified Confirmed HIV-Positive Test Events
95901	4
95691	15
95833	8
...	...

The cell representing the 95901 zip code (highlighted in red) requires suppression because the cell size is 1-4 cases and populations within zip codes can be small, presenting a higher risk for individual identification.

Solution 9: Statewide Data Broken Down by Zip Code

Table: Newly-Identified Confirmed HIV-Positive Test Events by Zip Code, CY 2013 – California

	Newly-Identified Confirmed HIV-Positive Test Events
95901	*
95691	15
95833	8
...	...

** This cell has been suppressed to protect confidentiality*

If the table above contained a total column, an additional cell would need to be suppressed to prevent inadvertent disclosure of confidential information (see the Golden Rule, section A.4.a).

B.2.d. If broken down by census-tract, suppress cell sizes that are 1-4 cases (see Example and Solution 10). If the US Census population denominator corresponding to the cell in question minus the cell numerator is greater than 100, OA staff may be able to leave the cell unsuppressed; however, staff must first discuss the risks for individual identification with their supervisor and obtain supervisor approval.

Example 10: Statewide Data Broken Down by Census Tract

Table: Newly-Identified Confirmed HIV-Positive Test Events by Census Tract, CY 2013 – California

	Newly-Identified Confirmed HIV-Positive Test Events
5125.08	4
5049.01	15
5001	8
...	...

Since the cell representing the 5125.08 census tract (highlighted in red) has a cell count of 1-4 cases, staff should suppress the cell unless the population within the census tract minus the cell count is > 100 and there is minimal risk for individual identification. Census tract population counts can be found on the US Census Bureau’s website.

Solution 10: Statewide Data Broken Down by Census Tract

Table: Newly-Identified Confirmed HIV-Positive Test Events by Census Tract, CY 2013 – California

	Newly-Identified Confirmed Positives
5125.08	*
5049.01	15
5001	8
...	...

** This cell has been suppressed to protect confidentiality*

3. Stratified by More Than One Variable

B.3.a. If statewide-level data are stratified by one or more demographic variables, suppression is generally not required unless a sensitive population is involved (see Example 11). If a sensitive population is involved, identify potential threats to confidentiality and discuss them with management.

Example 11: Statewide Data Stratified by More than One Variable

Table: CA Living Cases Stratified by Race/Ethnicity and Gender, CY 2013 – California

	Male	Female	Transgender
Black/African American	5,000	1,200	90
White	16,000	1,500	130
Hispanic/Latino	8,500	1,500	150
Other	2,000	300	4

Since the statewide transgender population with an Other race is not likely at risk for individual identification, it is acceptable to publish the cell count highlighted in green.

C. County-Level Guidelines and Examples

1. No Further Breakdown than County

C.1.a. If the table is not broken down by further demographics, suppression is not required (see Example 12).

Example 12: County-Level Data with No Further Breakdown than County

Table: Number of Test Events and Newly-Identified Confirmed HIV-Positive Test Events, CY 2013 – California (Modoc County)

	Total Test Events	Newly-Identified Confirmed HIV-Positive Test Events
Modoc	400	4

Since the table above is only stratified by county, suppression is not required.

2. Stratified

C.2.a. If county-level data are stratified by one or more additional identifying demographic variables (e.g., age, gender, race, risk group, etc.), cells with numerators that are 1-4 cases must be suppressed (see Example and Solution 13).

Example 13: County-Level Data Stratified by One or More Demographic Variables

Table: Number of Test Events and Newly-Identified Confirmed HIV-Positive Test Events by Gender, CY 2013 – California (Alameda County)

		Total Test Events	Newly-Identified Confirmed HIV-Positive Test Events
Alameda	Male	8,000	80
	Female	4,000	21
	Transgender	300	4

Since the table is stratified by county and an additional demographic variable, and the cell highlighted in red is between 1-4 cases, suppression is required.

Solution 13: County-Level Data Stratified by One or More Demographic Variables

Table: Number of Test Events and Newly-Identified Confirmed HIV-Positive Test Events by Gender, CY 2013 – California (Alameda County)

		Total Test Events	Newly-Identified Confirmed HIV-Positive Test Events
Alameda	Male	8,000	80
	Female	4,000	21
	Transgender	300	*

** This cell has been suppressed to protect confidentiality*

Only the cell with a count of 1-4 cases was suppressed because arithmetic cannot be used to derive the suppressed cell count. However, if the total number of newly-identified

positives was included in the table, additional suppression would be required (see the Golden Rule, section A.4.a).