CDSS

KIM JOHNSON
DIRECTOR

STATE OF CALIFORNIA—HEALTH AND HUMAN SERVICES AGENCY
DEPARTMENT OF SOCIAL SERVICES
744 P Street • Sacramento, CA 95814 • www.cdss.ca.gov

GAVIN NEWSOM
GOVERNOR

California Department of Social Services
Data De-Identification Reference Guide

I. Introduction

The California Department of Social Services (CDSS) is committed to providing useful data and promoting the transparency of state government through the public release of data. Prior to public release, all data must be assessed to determine whether any personal characteristics contained in the data pose the risk of identifying individuals. To protect the privacy of individuals served by the Department, the modified version of the Data De-Identification Guidelines (DDG)[1] developed by the California Health and Human Services Agency is being used.

Given that CDSS is not a covered entity under the Health Insurance Portability and Accountability Act (HIPAA), the de-identification guidelines omit procedures mandated for HIPAA covered entities such as the expert determination process[2] and Safe Harbor[3]. This document describes the procedures that must be followed in preparing data for public release.

The CDSS procedures focus on the assessment of aggregate or summary data for purposes of de-identification and public release. Aggregate data is 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.

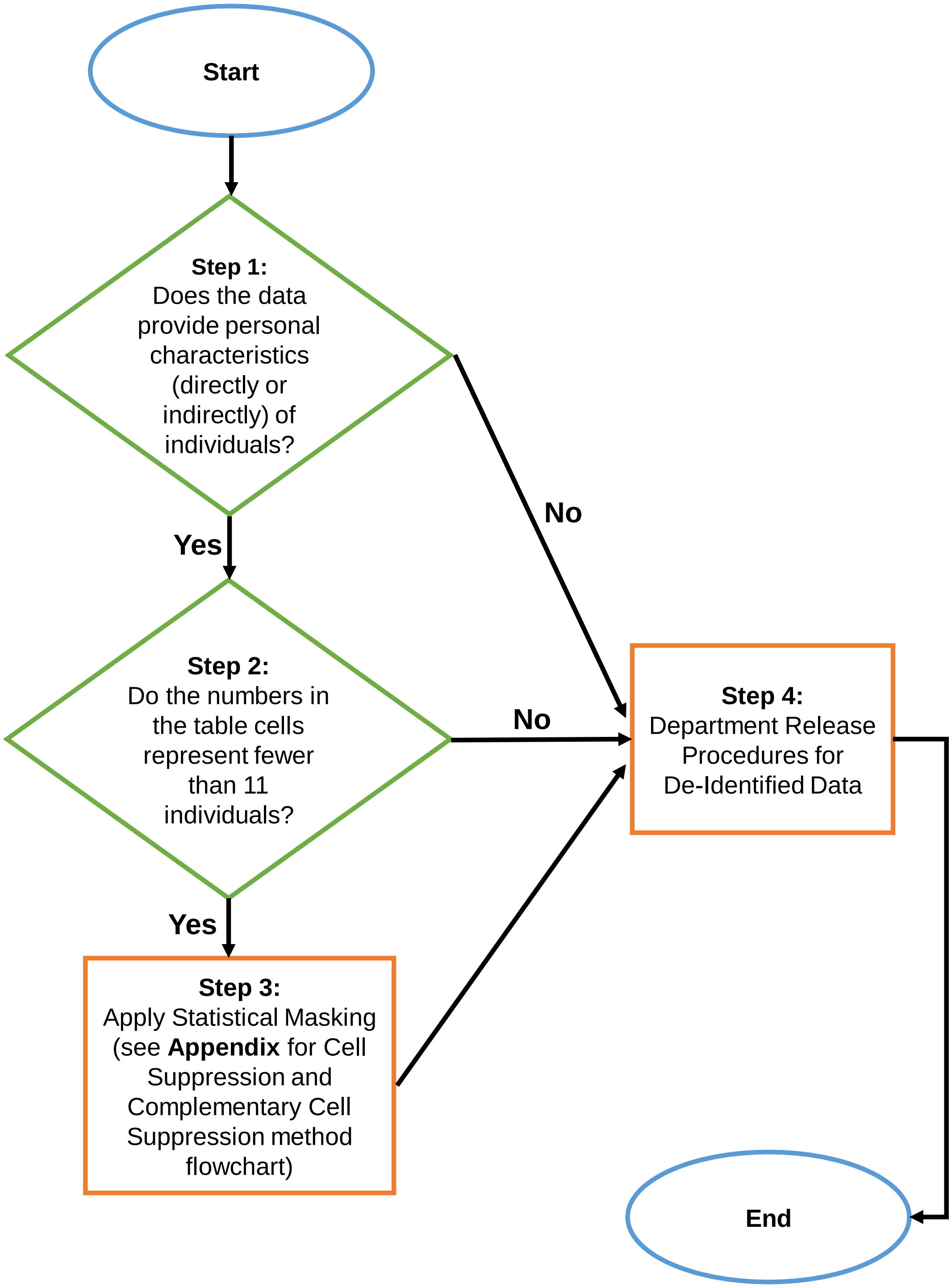
In contrast, record level data refers to a specific person or entity. Even after personal identifiers are removed, record level data inherently has higher risk than aggregate or summary data to identify an individual. Although the procedures should assist in reviewing record level data, a further case-by-case assessment must be made to ensure it is de-identified and does not include personal information that directly identifies an individual.

II. Data Assessment for Public Release Procedure

Prior to the public release of any CDSS data, the following four steps must be taken to ensure that any personal characteristics in the data cannot be used to identify an individual. While the examples below primarily focus on numbers presented in data tables, these steps also apply to any numbers or percent values present in written reports that represent fewer than 11 individuals.

[1] https://chhsdata.github.io/dataplaybook/resource_library/#datade-id
[2] Data De-Identification Guidelines (DDG), California Health and Human Services, Version 1.0 (2016), p. 41
[3] Data De-Identification Guidelines (DDG), California Health and Human Services, Version 1.0 (2016), p. 8

Flowchart of Steps for Data De-Identification


Start
Step 1: Does the data provide personal characteristics (directly or indirectly) of individuals?
No
Yes
Step 2: Do the numbers in the table cells represent fewer than 11 individuals?
No
Step 4: Department Release Procedures for De-Identified Data
Yes
Step 3: Apply Statistical Masking (see **Appendix** for Cell Suppression and Complementary Cell Suppression method flowchart)
End

Step 1 – Personal Characteristics of Individuals

Does the data provide personal characteristics (directly or indirectly) that can be tied back to an individual?

Examples include but are not limited to: age, gender, race, ethnicity, language spoken, location of residence (including county), location of services received or accessed (including county), education status, financial status, physical description, sexual orientation, gender identity, medical history, and employment history.

If 'Yes', go to Step 2 – If 'No', go to Step 4.

Step 2 – Data Values (Table Cell Counts)

Do numbers in products such as tables and reports represent fewer than 11 individuals?

If 'Yes', go to Step 3 – If 'No', go to Step 4.

Step 3 – Statistical Masking

Which of the following statistical masking methods (Reduce Table Dimensions, Combine Categories, or Cell Suppression and Complementary Cell Suppression) will be used to de-identify the data?

Please keep in mind that multiple statistical masking methods may be used on the same set of data or within the same report.

Descriptions and examples of Step 3 are provided on pages 4 through 13. A flow chart of Step 3 is provided in the appendix on page 14.

After Completing Statistical Masking go to Step 4.

Step 4 – Departmental Release Procedures for De-Identified Data

What are the review and release procedures for the organization?

The expectation is that the review of data for de-identification will align with other routine review processes. Products containing de-identified data will undergo the standard data release approval process before public release, including any contractual obligations owed to the originators of the data. Products may include but are not limited to: reports, presentations, publications, tables, Public Records Act responses, media responses, and legislative responses.

Statistical Masking

Statistical Masking Method: Reduce Table Dimensions

A single table containing multiple column and row dimensions may have some cells that represent fewer than 11 individuals. This is seen in the unmasked version of *Example 1* below, which has two dimensions – *Education Level* and *Generation.* Producing multiple tables with fewer table dimensions can be used as a statistical masking method to increase the counts of individuals to at least 11. In the example below, the table dimensions *Education Level* and *Generation* are used to create two distinct tables, which increases the numbers within each table cell to represent at least 11 individuals.

Example 1: Reduced Table Dimensions
Unmasked – Education Level by Generation

Generation	High School Diploma/GED	Associates	Bachelors	Graduate	Total
Millennial (18-34)	134	150	78	8	370
Generation X (35-50)	371	237	50	3	661
Total	505	387	128	11	1031

Masked -Generation

Generation	Total
Millennial (18-34)	370
Generation X (35-50)	661
Total	1031

Masked - Education Level

Education Level	Total
High School Diploma/GED	505
Associates	387
Bachelors	128
Graduate	11
Total	1031

Statistical Masking Method: Combine Categories

Combining multiple categories into a single category may increase the value of a table cell to a number representing at least 11 individuals, as in the example below, in which multiple ethnic groups are combined into a single "Other" category.

Example 2: Combine Categories
Unmasked – Services Accessed by Ethnicity

Service Type	Black	White	Latino	Asian	Native American	Other	Total
Substance Abuse Service	40	208	88	4	3	37	380
Mental Health Service	28	237	79	11	8	19	382
Total	68	445	167	15	11	56	762

Masked – Ethnicity

Service Type	Black	White	Latino	Other	Total
Substance Abuse Service	40	208	88	44	380
Mental Health Service	28	237	79	38	382
Total	68	445	167	82	762

In *Example 2 Asian*, *Native American*, and *Other*, will be merged into a single *Other* category. Creating a single *Other* category will ensure that no cells represent fewer than 11 individuals. The advantage of combining categories is the ability to present two data elements, such as ethnicity and service type, in a single table.

The ethnicity categories *Asian* and *Native American* were selected to be combined because they contain the smallest values. After combining the *Asian* and *Native American* categories, however, the number of individuals accessing *Substance Abuse* services still represents fewer than 11 individuals. Since the category *Other* provides less granular information than *Black*, *White*, or *Latino*, which designate specific ethnic groups, *Other* will be combined with *Asian* and *Native American* to ensure that the new *Other* category contains numbers that represent at least 11 individuals.

Note: Footnotes should be used to indicate which categories have been combined.

Statistical Masking Method: Cell Suppression and Complementary Cell Suppression

If reducing table dimensions or combining categories is not practical, then it may be necessary to suppress all cells that represent fewer than 11 individuals. Complementary cells, which are cells that could be used to calculate and re-identify suppressed cells (i.e., cells representing fewer than 11 individuals), will also need to be suppressed. This masking method might be selected if the report requires a greater level of detail, such as a county-based report.

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

- "Values are not visible to protect the confidentiality of the individuals summarized in the data."[4]

How to Suppress Small Cells and Perform Complementary Cell Suppression:

1: Suppress Small Cells

Small Cells: Cells that represents fewer than 11 individuals.

- Mask all numbers (cell values) that are less than 11 (i.e., derived from fewer than 11 individuals) with an asterisk (*), when possible.
    - Note: Values of zero (0) are typically shown since a non-event cannot be identified[5].
- If a complementary cell must be suppressed and has a value of 11 or greater, a double asterisk (**) should be used, whenever possible, to differentiate it from cells suppressed with a value of less than 11.

Example 3: Small Cell Suppression Unmasked – Application Approvals by Family Type

Applications	Single Parent	Two Parent
Approved	56	15
Denied	5	0
Pending	12	6

Masked Application Approvals by Family Type

Applications	Single Parent	Two Parent
Approved	56	15
Denied	*	0
Pending	12	*

[4] Data De-Identification Guidelines (DDG), California Health and Human Services, Version 1.0 (2016), p. 39

[5] Data De-Identification Guidelines (DDG), California Health and Human Services, Version 1.0 (2016), p. 15

2: Complementary Cell Suppression

Complementary Cell: A number representing more than 11 individuals that can be used to calculate and re-identify a small cell or small cells.

Complementary Cell Suppression: When a number representing 11 or more individuals is suppressed using one of the methods listed below to prevent the re-identification of other suppressed cells.

- Numbers 11 and higher may need to be suppressed (i.e., complementary cell suppression) if any numbers less than 11 can be re-identified through the addition or subtraction of any unsuppressed numbers. When numbers 11 and higher are suppressed, a double asterisk should be used whenever possible.
- In each column or row containing a suppressed number, at least one other number must be suppressed through complementary cell suppression[6].
- In cross tables (tables containing both column and row totals), if a number is suppressed then both the column and row must be checked to determine if complementary cell suppression is necessary.

***Methods of Complementary Suppression*[7]:**
One of the following methods should be selected for use. The composition of the data or specific reporting needs of the organization may be used to determine which of the following four options will be used.

- Next Smallest Number:
    - Suppress the next smallest unsuppressed number. This method retains larger numbers that represent more individuals (see *Example 4 – Option A* on page 8).
      Note: There is no maximum value of the "Next Smallest Number."

- Suppress/Roll-up Total:
    - Suppress the number containing the row or column sum. This method allows for automation of the suppression process (i.e., through the use of Excel macros and formulas), which reduces human error (see *Example 4 – Option B* on page 9).

- 'Least Interesting' Category:

  Suppress the 'least interesting' category. This is often a category such as 'other' or 'I don't know' (see *Example 4 – Option C* on page 10).

[6] Data De-Identification Guidelines (DDG), California Health and Human Services, Version 1.0 (2016), p. 19
[7] Data De-Identification Guidelines (DDG), California Health and Human Services, Version 1.0 (2016), p. 20

- Similar Group:
    - Suppress 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 (see *Example 4 – Option D* on page 10).

Example 4: Complementary Cell Suppression
Unmasked – Barriers to Housing

Ethnicity	Poor Credit	Past Evictions	Criminal Record (Self)	Criminal Record (Family Member)	Other	Total
Black	1,561	1,178	1	12	13	2,765
White	3,732	1,465	9	16	22	5,244
Latino	4,028	1,227	13	15	15	5,298
Other	4,929	1,510	11	19	17	6,486

1: Suppress Small Cells

Ethnicity	Poor Credit	Past Evictions	Criminal Record (Self)	Criminal Record (Family Member)	Other	Total
Black	1,561	1,178	*	12	13	2,765
White	3,732	1,465	*	16	22	5,244
Latino	4,028	1,227	13	15	15	5,298
Other	4,929	1,510	11	19	17	6,486

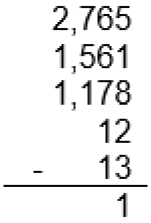
2: Complementary Cell Suppression

Option A – Next Smallest Number

Ethnicity	Poor Credit	Past Evictions	Criminal Record (Self)	Criminal Record (Family Member)	Other	Total
Black	1,561	1,178	*	**	13	2,765
White	3,732	1,465	*	**	22	5,244
Latino	4,028	1,227	13	15	15	5,298
Other	4,929	1,510	11	19	17	6,486

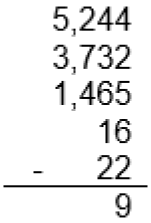
In the previous table, the next smallest number was suppressed where it would be possible to re-identify a suppressed small cell (see highlighted cells). The examples below demonstrate how cells can be re-identified by subtracting all unsuppressed cells from the total, if complementary cell suppression does not occur.

Black: 1 – Total *Black* individuals with *Criminal Record (Self)* as a barrier to housing



```
    2,765
    1,561
    1,178
       12
 -     13
---------
        1
```

White: 9 – Total *White* individuals with *Criminal Record (Self)* as a barrier to housing



```
    5,244
    3,732
    1,465
       16
 -     22
---------
        9
```

2: Complementary Cell Suppression

Option B – Suppress/Roll-up Total

Ethnicity	Poor Credit	Past Evictions	Criminal Record (Self)	Criminal Record (Family Member)	Other	Total
Black	1,561	1,178	*	12	13	**
White	3,732	1,465	*	16	22	**
Latino	4,028	1,227	13	15	15	5,298
Other	4,929	1,510	11	19	17	6,486

Instead of suppressing the next smallest number, the total column can be suppressed or rolled up to prevent the re-identification of small cells. Because the total column is suppressed, one cannot use the total to calculate the suppressed numbers in the *Criminal Record (Self)* cells.

2: Complementary Cell Suppression

Option C – 'Least Interesting' Category

Ethnicity	Poor Credit	Past Evictions	Criminal Record (Self)	Criminal Record (Family Member)	Other	Total
Black	1,561	1,178	*	12	**	2,765
White	3,732	1,465	*	16	**	5,244
Latino	4,028	1,227	13	15	15	5,298
Other	4,929	1,510	11	19	17	6,486

The table above demonstrates complementary cell suppression in which the 'least interesting' category is suppressed. *Other* is the 'least interesting' category, since it is not a specific barrier to housing like the other categories

2: Complementary Cell Suppression

Option D – Similar Group

Ethnicity	Poor Credit	Past Evictions	Criminal Record (Self)	Criminal Record (Family Member)	Other	Total
Black	1,561	1,178	*	**	13	2,765
White	3,732	1,465	*	**	22	5,244
Latino	4,028	1,227	13	15	15	5,298
Other	4,929	1,510	11	19	17	6,486

Complementary cell suppression in the table above is accomplished by suppressing the group with characteristics most similar to the category requiring small cell suppression. Since *Criminal Record (Self)* is suppressed due to small cell size, the category which represents a group with similar characteristics, *Criminal Record (Family Member)*, is selected for complementary cell suppression. Coincidentally, the tables produced by suppressing the next smallest number (Option A) and a similar group (Option D) are the same in this example. This may not always be the case with all data.

Important Data Issues to Consider:

Complementary Cell Suppression for Cells with Identical Data

In a row containing two suppressed cells, the numbers contained in the suppressed cells typically cannot be re-identified by subtracting the unsuppressed cells from the total. When each suppressed cell equals one, however, it is possible to re-identify the number in each suppressed cell if the total cell is not suppressed. In *Example 5*, the next smallest number is suppressed during complementary cell suppression to prevent the cells containing the number of infants in a Group Home or Guardian placement from being re-identified.

Example 5: Complementary Cell Suppression

Unmasked – Infant Placement Types

Infants (Under 1)	Foster Care	Group Home	Guardian	Other	Total
Count	1,178	1	1	18	1,198

1: Suppress Small Cells

Infants (Under 1)	Foster Care	Group Home	Guardian	Other	Total
Count	1,178	*	*	18	1,198

2: Complementary Cell Suppression (Next Smallest Number)

Infants (Under 1)	Foster Care	Group Home	Guardian	Other	Total
Count	1,178	*	*	**	1,198

Even when a row or column has at least two numbers suppressed, it may still be possible to re-identify a suppressed number. In this case, all numbers that can be used to re-identify a suppressed number must be masked.

Complementary Cell Suppression with Multiple Cells Containing Zeros

In *Example 6*, the column containing the count for each family size group has one suppressed row, the *1 to 2 children* category. All other rows contain zeros. Because the column total reflects the value of the suppressed cell, the column total is also suppressed. This suppression results in a table in which the only visible values are zeroes. It is important to remember, however, that zeroes are important pieces of data that can convey meaningful information.

Example 6: Complementary Cell Suppression

Unmasked – Family Size (Small County)

Family Size	Count
1 to 2 children	1
3 to 4 children	0
5 to 6 children	0
6+ children	0
Total	1

Suppress Small Cells

Family Size	Count
1 to 2 children	*
3 to 4 children	0
5 to 6 children	0
6+ children	0
Total	*

Cell Suppression of Numbers in Text

Cell suppression guidelines also apply to any numbers that are included in the body of a report or other written document.

Numbers representing fewer than 11 individuals are unsuppressed in the report text below, which goes against the de-identification guidelines.

> **Sample Sentence:**
> "Out of 35 children, 6 were in care for 12 to 23 months and 5 were in care for 24 to 35 months"

This can be corrected by combining the two time-in-care categories into either a single category:

> **Sentence Revised for Suppression:**
> **"**Out of 35 children, 11 were in care for 12 to 35 months."

Or by using asterisks or text to suppress numbers representing fewer than 11 individuals:

> **Sentence Revised for Suppression:**
> "Out of 35 children, fewer than 11 were in care for 12 to 23 months and fewer than 11 were in care for 24 to 35 months"

Cell Suppression of Percentages in Text

Cell suppression guidelines apply to any reports or tables containing percentages. Percent values of small cells or complementary cells which could be used to re-identify masked numbers also need to be suppressed.

Because it is possible to determine the number of children in each group by multiplying the percent values against the total number of children [6 children were in care for 12 to 23 months (35 × 0.17 = 5.95) and 5 children were in care for 24 to 35 months (35 × 0.14 = 4.9)] the percent values must also be suppressed.

> **Sample Sentence:**
> “Out of 35 children, 17% were in care for 12 to 23 months and 14% were in care for 24 to 35 months”

This can be corrected by combining the two time in care categories into a single category:

> **Sentence Revised for Suppression:**
> “Out of 35 children, 31% were in care for 12 to 35 months.”

Or by using asterisks or text to suppress numbers representing fewer than 11 individuals:

> **Sentence Revised for Suppression:**
> “Out of 35 children, *% were in care for 12 to 23 months and *% were in care for 24 to 35 months.”

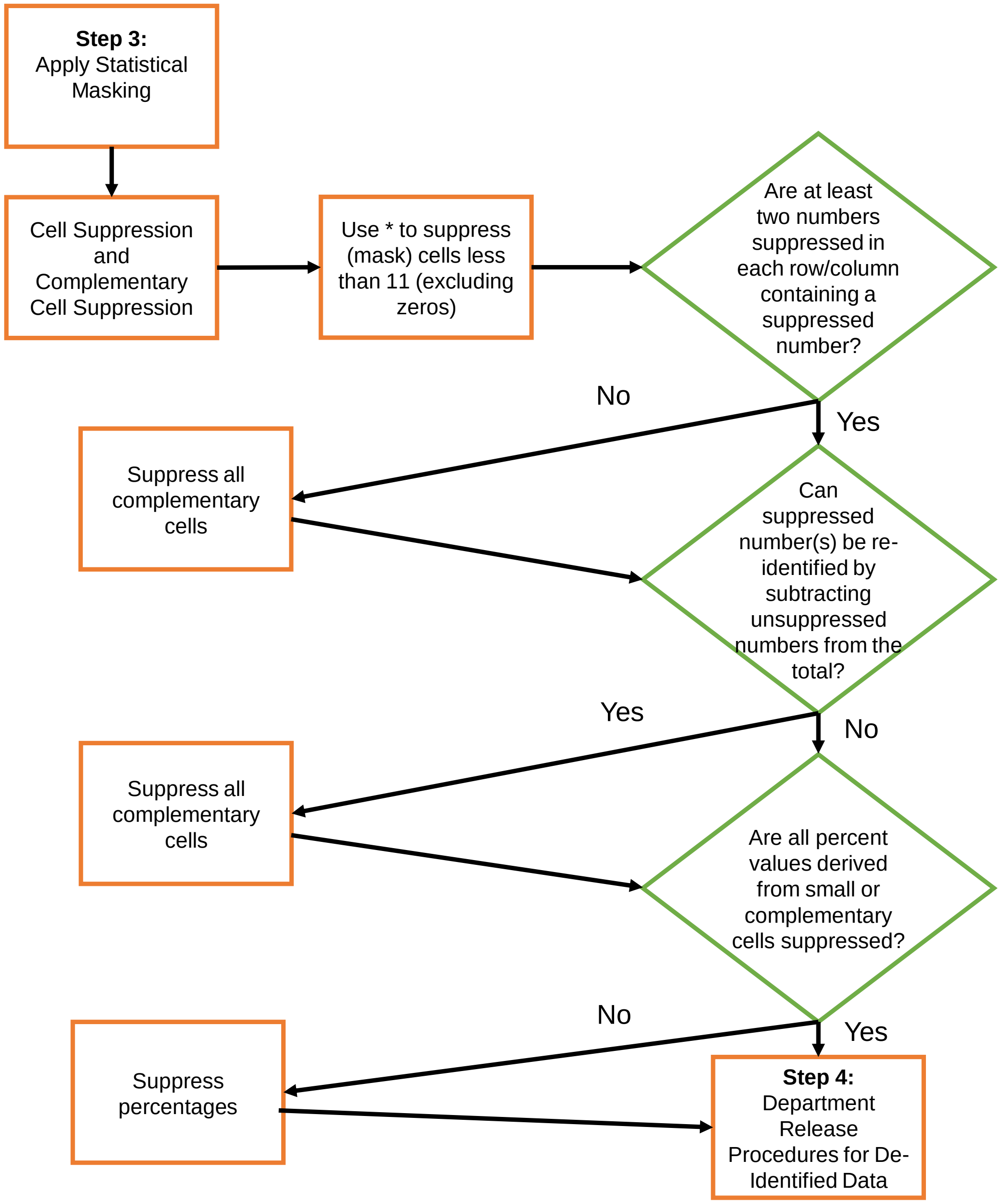
Small Numbers Not Representing Individuals

Small numbers do not need to be suppressed if they do not represent individuals. In *Example 7*, the number 9 does not need to be suppressed because it shows the difference between the cases carried forward from last month and the total from last month’s report (i.e., 6,454 - 6,445 = 9), not 9 individuals.

Example 7: Small Numbers Not Representing Individuals

Caseloads	Two Parent Families
Cases carried forward from last month	6,454
Total from last month’s report	6,445
Adjustment (difference between rows)	9

Appendix: Cell Suppression and Complementary Cell Suppression Flowchart


Step 3:
Apply Statistical Masking
Cell Suppression and Complementary Cell Suppression
Use * to suppress (mask) cells less than 11 (excluding zeros)
Are at least two numbers suppressed in each row/column containing a suppressed number?
No
Yes
Suppress all complementary cells
Can suppressed number(s) be re-identified by subtracting unsuppressed numbers from the total?
Yes
No
Suppress all complementary cells
Are all percent values derived from small or complementary cells suppressed?
No
Yes
Suppress percentages
Step 4:
Department Release Procedures for De-Identified Data