Creating a fully de-identified public use dataset (PUDS): Policy and practice

Steven C. Macdonald PhD, MPH
Washington State Dept. of Health
Office of Epidemiology

Washington State Department of Health

CDC – Good professional practice for data release

- Public health and scientific advancement are best served when data are released to, or shared … in an open, timely, and appropriate way
- The interests of the public—which include timely releases of data for further analysis—transcends whatever claim scientists may believe they have to ownership of data acquired or generated using federal funds.
- Such data are, in fact, owned by the federal government and thus belong to the citizens of the United States.

CDC *Policy on Releasing and Sharing Data*

- Data that CDC collects or holds and that can be legally released to the public should be released through a public-use data set within a year after the data are evaluated for quality and shared with any partners in data collection.
- To ensure that issues of confidentiality, proprietary use, and informed consent are addressed correctly, CIOs [CDC operating units] may choose to develop specific data release plans for each data set.
- Each plan should include the following: A procedure to ensure that confidential information is not disclosed…

WA-DOH policy on data release

- To maintain the public's trust and achieve its mission, the department must act as a responsible custodian of the information it holds.
- It must protect the privacy of individuals and … provide appropriate access to confidential data/information in limited situations authorized in law.
- In most cases, when data elements or information that may identify an individual are removed, the records can be disclosed.

Washington State Department of Health

WA-DOH policy on disclosure

- Executive Order 00-03 directs all state agencies …to eliminate the use of personal identifiers from documents and information that may be subject to disclosure to ensure that the confidentiality of Washington's citizens is protected.
- Personal information may include, but is not limited to, name, address, telephone number, social security number, credit card and other individual financial identification numbers, and medical record number.

Washington State Department of Health

WA-DOH policy on confidentiality

- … confidential data/information in any form where the individual may be identified is not to be disclosed, except as allowed by law.
- In most cases, records can be disclosed when data/information that identifies or may reasonably lead to the identification of an individual are removed.

Washington State Department of Health

Confidentiality Protection

1. Limit disclosure of potential identifiers
2. Aggregate data values
3. Limit the number of records or the number of fields
4. Use numerator rules for data aggregation or suppression
5. Use denominator rules for data aggregation or suppression
6. Refrain from using techniques that distort data

CDC NCIPC
Minimal Data Quality Standards

- Completeness of case ascertainment
- Completeness, validity, and consistency of data
- Description of original data element and created analysis variables (metadata)
- Assurance of confidentiality and nondisclosure

Washington State Department of Health

Approaches

Field	Greater aggregation (coarse granularity)	Less aggregation (fine granularity)
Age	six “Twenty-year” categories – 0-19, 20-44, 45-64, 65-84, 85+ – <1, 1-14, 15-24, 25-44, 45-64, 65+)	• 11 “Ten-year” categories (<1, 1-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+) • 20 “Five-year” categories

Washington State Department of Health

Approaches

Field	Greater aggregation (coarse granularity)	Less aggregation (fine granularity)
Location of residence	• PHEPR (9) • Zip-code – 3 digits (14) • PUMS (17)	• County (39) • Zip-code – 5 digits (583) • Census tract

Approaches

Field	Greater aggregation (coarse granularity)	Less aggregation (finer granularity)
Date of event	• Multi-year • Year • Season	• Month • Day

Washington State Department of Health

WA-DOH *Guidelines for Working With Small Numbers*

- Examine numerator size for each cell.
  - If the count of cases or events in a cell is less than three, the data analyst needs to consider whether a breach of confidentiality is likely.
  - A count of no events in the cell is clearly no threat to confidentiality, but a count of one or two events may be.

Washington State Department of Health

WA-DOH *Guidelines for Working With Small Numbers*

- Examine denominator size for each cell.
  - tabular data based on denominators greater than 300 persons per cell present minimal risk for individual identification
  - caution should be exercised by the analyst if the population size is between 100 and 300
  - extreme caution is warranted when the population is less than 100

Washington State Department of Health

CDC EPHTN
draft Data Re-release Plan

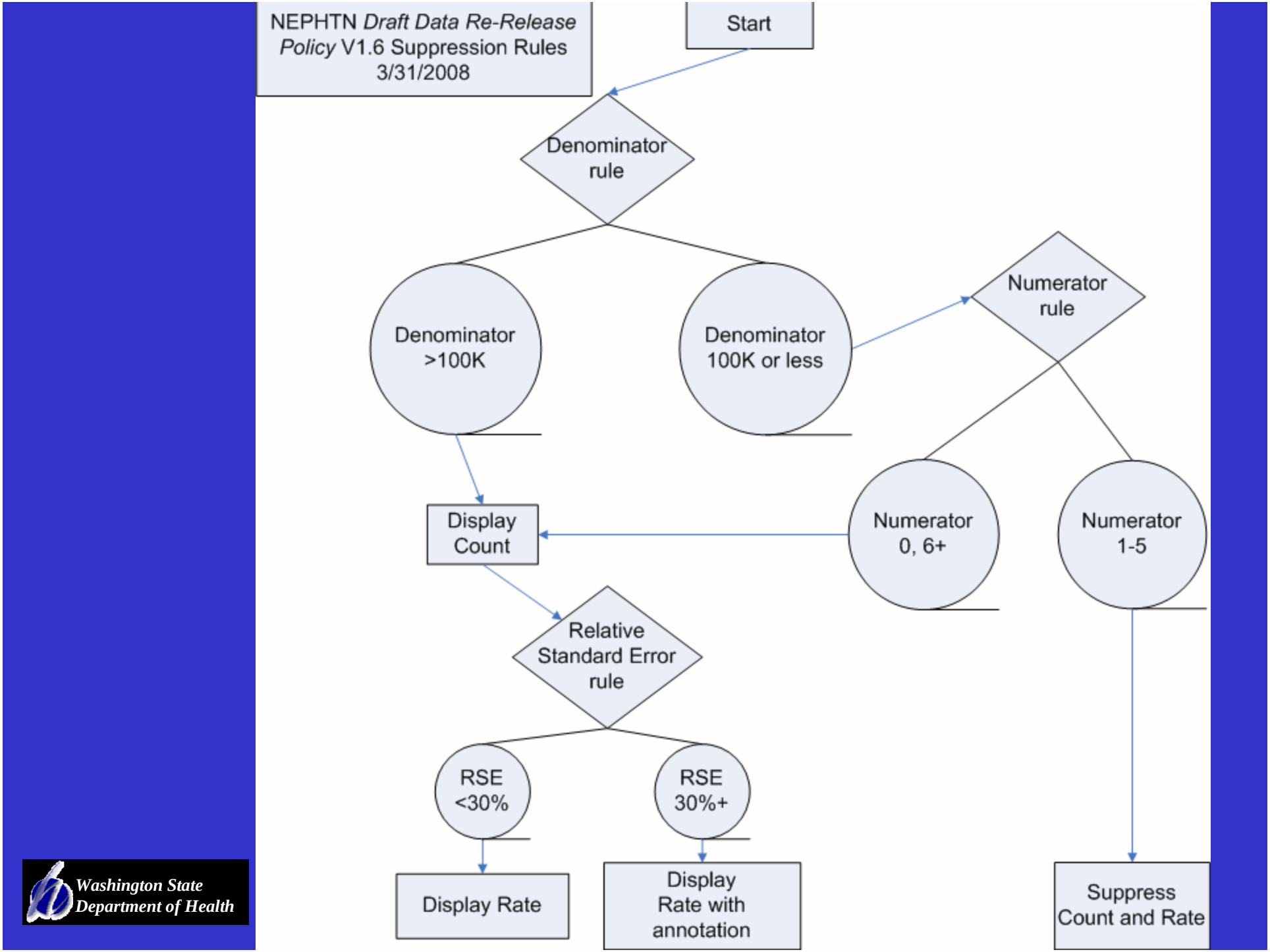
- Combined denominator and numerator
- Allows differential display
  - Display rate in tables
  - Display rate with annotation (due to statistical stability concerns)
  - Suppress count and rate (due to disclosure risk concerns)

Washington State Department of Health

CDC EPHTN Suppression rules

- Suppress count and rate in tabular display
  - If denominator 100,000 or less and
  - non-zero numerator <6
- Display rate with annotation
  - Relative Standard Error (RSE) 30%+
- Display rate
  - RSE <30%

Washington State Department of Health



Washington State Department of Health

EPHTN Confidentiality protection & Statistical stability

- Aggregation
  - with dynamic variable restriction
- Suppression
  - primary
  - complementary
- Smoothing
  - Empirical Bayes method

Washington State Department of Health

Disclosure risk assessment

- Statistical approaches
  - Poisson log-linear models
  - Bayesian hierarchical models
  - Bayesian model-averaging predictive probabilities
  - Subtraction-attribution probability (SAP)

Disclosure risk limitation

- Application of the concept of Uniqueness
- Perturbation
  - Masking
  - Data swapping, blanking, blurring
  - Adding random noise (“Salting”)
  - Controlled random rounding
- Anonymised microdata: R-package *sdcMicro*

Washington State Department of Health

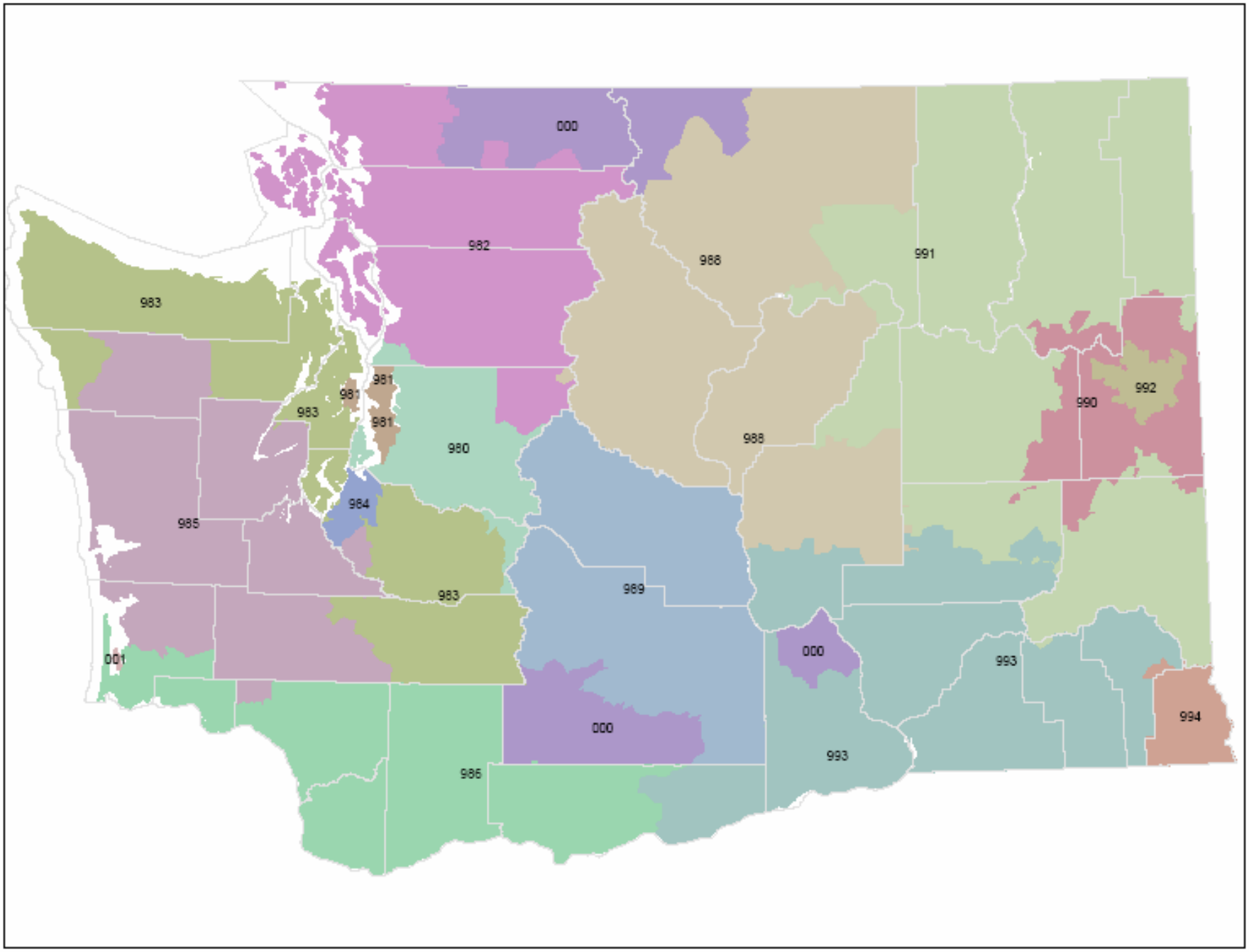
Confidentiality-utility tradeoff

- Complaint: Aggregation of numbers for rural areas and subgroup characteristics results in loss of information to user.
- Answer: PUDS may not support detailed analyses, but rather serve as a preliminary study tool. DSAs are available for users who need finer aggregation.

Discussion & questions

Washington State Department of Health


Washington State
Department of Health


000
982
988
991
983
981
981
983
981
990
992
980
988
984
985
983
969
001
000
993
994
000
993
986