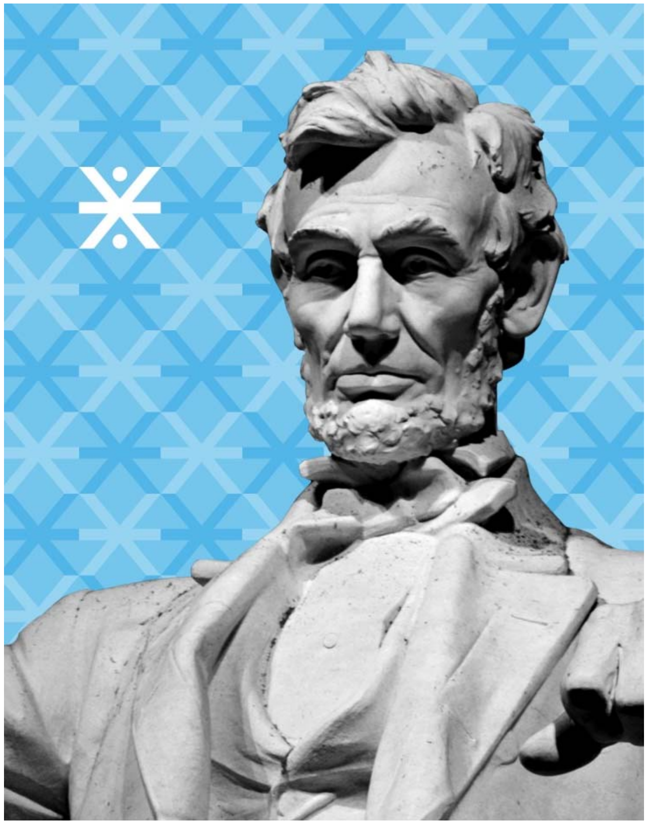
# Improving Administrative Data Quality for Research and Analysis

China Layne, Ph.D.

Association of Public Data Users (APDU) Webinar

The information provided in this webinar are the views of the author alone and do not reflect the views of the Association of Public Data Users (APDU).

June 21, 2017



summit

complexity simplified.

# Purpose

**This webinar is a primer on how to improve the quality of administrative data for research and analysis.**

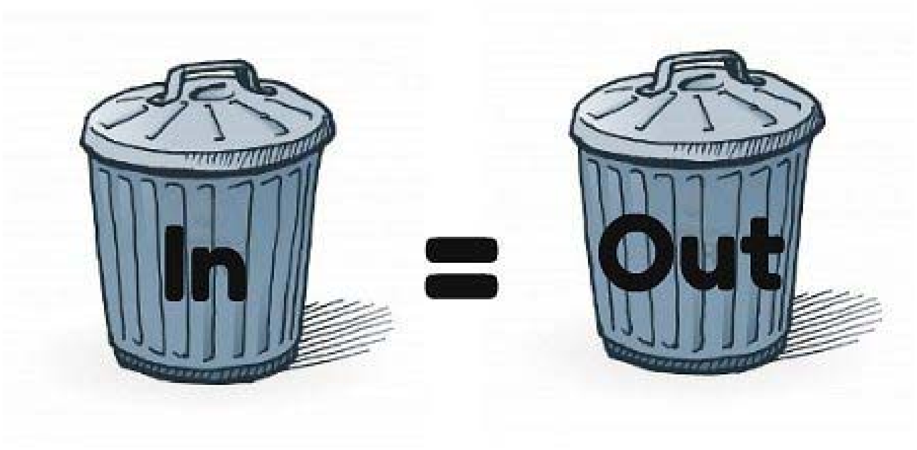
- Webinar is useful for organizations that are new to using administrative data for research
- I will review major issues of data quality for research and discuss concrete strategies for reviewing and cleaning administrative data
- Participants will have the tools to transform their administrative data into a "research-ready" dataset

# Agenda

**The webinar covers four main topics.**

- Importance of administrative data quality for research
- Major issues of data quality
- Strategies for reviewing the quality of the data
- Strategies for cleaning the data

# Why is administrative data quality important?



http://www.vivianpartnership.co.uk

**There are several *immediate effects* of poor quality data.**

- Difficulty using the data for analysis
- Fewer types of analyses can be conducted
- Inaccurate or biased analysis results
- Inaccurate records of program performance

# Why is administrative data quality important?

Mooc.udp.cl

**Poor quality data can also have *long-term effects.***

- Data have less overall utility
- Fewer analysts use the data
- Organizations may harm their own reputations by:
  - Incorrectly assessing program success or failure
  - Releasing inaccurate research
  - Developing policies and programs based on inaccurate conclusions

# What are the main issues of data quality?

**There are seven main issues for administrative data quality.**

## 1. Unit of analysis

- Not able to tell to which entity or activity a record corresponds (e.g. participant, intervention, program site)

## 2. Duplicate records

- Multiple records in the data represent one case at the level of the unit of analysis
- Data includes old or out-of-date records for a case

3. Missing values

4. Invalid values

5. Incorrect value formats

6. Outlier values

7. Value inconsistencies

# What are the main issues of data quality?

1. Unit of analysis

2. Duplicate records

**3. Missing values**

- Specific variables missing many values or specific cases missing many variables
- Values missing not at random
- Not able to distinguish between missing and not applicable values

**4. Invalid values**

- Values that don’t fit the logical or format requirements of the variable (e.g. negative values for wage and salary income or respondent age in values with decimals)

5. Inconsistent value formats

6. Outlier values

7. Value inconsistencies

# What are the main issues of data quality?

1. Unit of analysis

2. Duplicate records

3. Missing values

4. Invalid values

**5. Inconsistent value formats**

- Variables with known coding schemes (e.g. industry, occupation, geographies) coded in other formats or variables coded differently across time or across related datasets

**6. Outlier values**

- Values that are abnormally high or low compared to the variable’s other values (e.g. working 90 hours in a week)

7. Value inconsistencies

# What are the main issues of data quality?

1. Unit of analysis

2. Duplicate records

3. Missing values

4. Invalid values

5. Inconsistent value formats

6. Outlier values

**7. Value inconsistencies**

- Values between variables with known relationships that are inconsistent or incorrect (e.g. wage income for unemployed persons)

# How can we review the data?

**Thorough, methodical, and careful data review will uncover errors.**

## 1. Unit of Analysis

- Data documentation (e.g. codebook and data dictionary) should identify the unit of analysis

## 2. Duplicate records

- Many statistical packages have commands for identifying duplicate records (e.g. *duplicates* in Stata)
- You'll need to know the intended unit of analysis (and corresponding variable) to identify duplicate records

# How can we review the data?

## 3. Missing values

- Many statistical packages have commands for identifying missing values (e.g. *missings* in Stata)
- Check number of missing values for every variable
- Check number of missing values across variables for each case
- Check if there is a pattern to missing values by characteristics of unit of analysis (may indicate bias)

## 4. Invalid values

- Documentation provides the accepted value range and formats
- Can use summary stats, histograms, and scatterplots to check
- Can check invalid values as a search for values that fall outside of the specified range

# How can we review the data?

## 5. Inconsistent value formats

- External documentation will help identify correct coding for standardized variables (e.g. industry, geographic areas)
- Check value format consistency across merged data as well

## 6. Outlier values

- Can use summary statistics, histograms, scatterplots, or box plots to identify outlier values

## 7. Value inconsistencies

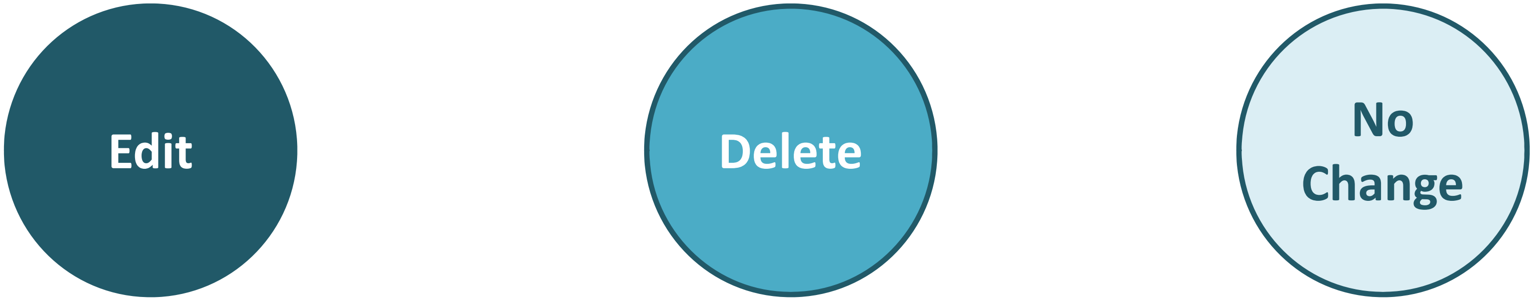
- Documentation can identify expected variable relationships
- Scatterplots and correlations can show inconsistent values
- Be sure to check reciprocal relationships
- Check value inconsistencies across merged data as well

# How can we review the data?

- **Should review these issues for every variable**
- Data documentation is invaluable for knowing how the data *should* look
- Automate as much of the review process as possible
- Widespread errors may indicate systematic collection problems
- Data review is like solving a puzzle!

# How can we clean the data?

**There are three primary ways to resolve data quality issues:**



Edit

Delete

No Change

- Which method you use will depend on the extent and severity of the issue and the availability of confirmation information

## 1. Unit of Analysis ● ○

- Be sure that any aggregate variables are written out to every record for the respective unit of analysis
- If there are no problems with aggregate variables, no cleaning is needed

# How can we clean the data?

## 2. Duplicate cases ●

- Remove duplicate records based on the unit of analysis
- Keep only the most recent or updated record for each case
- Will require identifying the variable(s) that determine record duplication (e.g. date), if documentation doesn’t provide it

## 3. Missing values ● ● ○

- Use imputation to fill in missing values
- Methods include mean imputation, regression-based imputation, and multiple imputation
- **Leave outcome variables un-imputed**
- May need to delete an entire case if too many variables have missing values

# How can we clean the data?

## 4. Invalid values ● ●

- Edit with caution
- Only edit cases when information is available to confirm the correct value
- Documentation can help determine the correct value
- If information is not available, delete known erroneous values

## 5. Outlier values ● ○

- Top- or bottom-code extreme values (e.g. recoding annual earnings over $250,000 to $250,000)
- May also leave outlier values, but flag these values using another variable

# How can we clean the data?

## 6. Inconsistent value formats ● ○

- Seek out standard coding schemes for variables
- Recode variables to match these standards if possible
- Recode variables for merged data as well
- If variables can’t be recoded, leave as is

## 7. Value inconsistencies ● ○

- Edit with caution
- Only edit cases when information is available
- Documentation can help determine the correct value
- Edit value inconsistencies for merged data as well
- If information is not available, leave the values and use with caution

# How can we clean the data?

- Be consistent in the editing process
- Document the editing process, assumptions, and results
- Name and label recoded, imputed, and edited variables separately from original variables
- Keep programs for common recoding activities
- Identify unit of analysis and remove duplicate records before merging

# About Summit Consulting, LLC

## Check out our panel at the 2017 APDU Annual Conference: “Data Integration to Improve Program Effectiveness”

**China Layne, Ph.D.**
Manager, Data Analytics and Research
202.407.8300
China.Layne@summitllc.us

**Anthony Curcio**
Principal, Federal Credit Practice
202.407.8303
Anthony.Curcio@summitllc.us

## “Are your Administrative Data Ready for Public Use?”

https://www.summitllc.us/white-paper-2016-are-your-administrative-data-ready-for-public-release

*Summit is a specialized analytics advisory firm that guides Federal agencies, financial institutions, and litigators as they decode their most complex analytical challenges. Summit’s staff of economists, econometricians, and research scientists use quantitative techniques to assist our clients as they model risk, evaluate program performance, and predict future performance. Our distinct capabilities include program evaluation, applied statistics and economics, mortgage finance, financial services, Federal Credit modeling and forecasting, and litigation analytics.*

*Summit: complexity simplified*

*www.summitllc.us*