Best practices in using secondary analysis as a method

Katharine Green, PhD(c), CNM
University of Massachusetts Amherst, USA
July, 2015
Secondary data analysis: why use?

- Uses prior data to obtain new information or interpretation.
- Frequently used, electronic medical records
- Cost effective, Efficient use of time
- Good for large studies, longitudinal data
- Good for comparing data across databases
- “Real life” (example: pharmaceutics)
“What are best practices with this method in health care?”

- Literature review
  - CINAHL, GoogleScholar, Ovid
- Exploring Attributes of the method and considerations in health care
Best practices: design

• Fit of data:
  o Note initial strengths/ limitations and justify use, congruity of type of database and analysis packages

• Size of sample

• Measurement/ analysis

• Did database measure variable of interest?

• Does current conceptual framework fit into the theory that underpinned first study?
Data management

• Similar to any other study: data cleaning
  o Consistency, accuracy, evaluation of outliers
• Coding errors
• Missing data or improperly defined data
• Pilot any forms
• Inclusion and exclusion criteria in primary use and in secondary use
  o Do data points and methods from initial study match current study needs?
Statistical analysis

• Similar to doing any initial study:
  o Descriptive, predictive, and non-linear statistics
  o Standard error should be smaller as unlikely to be randomized to decrease risk of falsely rejecting null hypothesis
Think carefully before proceeding….

• For qualitative use
  o Usually unable to see or hear initial interviews
  o Interpretation issues
  o Sensitivity of data
• For meta analysis
  o Frequently use of outcomes, not initial data
  o Risk of bias
• Usually cannot establish causality: this method is almost always observational
• Who owns the data?
Other cautions

- Self reported data, missing data, unasked questions
  - Missing data may be bias - not just missing
- Changes in practice, diagnoses, treatment
- Societal, economic and political changes
- Time lapse
- Multiple researchers using database
- Was informed consent obtained for current work?
Data linkage

• If link pharmacy, ancillary health services, and electronic medical records during providers’ or hospital visits, may have fairly complete picture of people who did not consent.

• Takes surprisingly few data points to identify someone!
More on data linkage

• Takes only 4 data points from a cell phone provider to identify a person
• 90% adults can be identified through a social media file (Yves-Alexandre et al., 2013)
• Large companies forming aggregate data bases (Tucker, 2013)
• Takes only 5 medical record data points to link a newspaper to a person. (Sweeney, 2013)
HIPAA (Health Insurance Portability and Accountability Act)

Names.

- All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census:
  - The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people.
  - The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.
HIPAA

- All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; 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.
- Telephone numbers.
- Facsimile numbers.
- Electronic mail addresses.
HIPAA

• Social security numbers.
• Medical record numbers.
• Health plan beneficiary numbers.
• Account numbers.
• Certificate/license numbers.
• Vehicle identifiers and serial numbers, including license plate numbers.
• Device identifiers and serial numbers.
• Web universal resource locators (URLs).
HIPAA

- Internet protocol (IP) address numbers.
- Biometric identifiers, including fingerprints and voiceprints.
- Full-face photographic images and any comparable images.
- Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification.
“What are best practices with this method in health care?”

• Aggregated (combined) data
• De-identified data
• Revise consent forms to remove illusion of anonymity unless it is actually possible
• Minimizing individual data if an anonymous study: is it possible?
• Rethinking issues of privacy: is it generational?
References


Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel (2013). Unique in the crowd: The privacy bounds of human mobility." Scientific Reports. DOI: 10.1038/srep01376