

Best practices in using secondary analysis as a method

Katharine Green, PhD(c), CNM
University of Massachusetts Amherst, USA
July, 2015



University of Massachusetts Amherst, U.S.A.

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 data bases
- “Real life” (example: pharmaceuticals)

“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:
 - 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
 - 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
 - Do data points and methods from initial study match current study needs?

Statistical analysis

- Similar to doing any initial study:
 - Descriptive, predictive, and non-linear statistics
 - 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
 - Usually unable to see or hear initial interviews
 - Interpretation issues
 - Sensitivity of data
- For meta analysis
 - Frequently use of outcomes, not initial data
 - 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

- Alvarez, J., Canduela, J. and Raeside, R. (2012), Knowledge creation and the use of secondary data. *Journal of Clinical Nursing*, 21: 2699–2710.
- Aponte, J. (2010). Key Elements of Large Survey Data Sets. *Nursing Economic\$, 28(1)*, 27-36.
- Benningfield, M., Dietrich, M., Jones, H., Kaltenbach, K., Heil, S., Stine, S., ... & Martin, P. (2012), Opioid dependence during pregnancy: relationships of anxiety and depression symptoms to treatment outcomes. *Addiction*, 107: 74–82.
- Berger, A. M., & Berger, C. R. (2004). Data mining as a tool for research and knowledge development in nursing. *Computer Information in Nursing*, 22, 123-131.
- Clarke, S., & Cossette, S. (2000). Secondary analysis: theoretical, methodological, and practical considerations. *Canadian Journal Of Nursing Research*, 32(3), 109-129.
- Doolan D. & Froelicher E. (2009) Using an existing data set to answer new research questions: a methodological review. *Research Theory and Nursing Practice* 23, 203–215.
- Finlayson, M., Egan, M., & Black, C. (1999). Secondary analysis of survey data: a research method with potential for occupational therapists. *Canadian Journal Of Occupational Therapy*, 66(2), 83-91.
- Gooding, B. (1988). Secondary analysis: A method for learning research activities. *Journal of Nursing Education* 27(5), 229-30.
- Higgins JPT, Green S (editors) (2011). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Retrieved from www.cochrane-handbook.org.
- Health Insurance Portability and Accountability Act of 1996, 42 U.S.C. § 1320d-9 (2010)
- Kneipp, S., & Yarandi, H. (2002). Complex sampling designs and statistical issues in secondary analysis. *Western Journal Of Nursing Research*, 24(5), 552-566.
- Long-Sutehall, T., Sque, M., & Addington-Hall, J. (2011). Secondary analysis of qualitative data: a valuable method for exploring sensitive issues with an elusive population?. *Journal Of Research In Nursing*, 16(4), 335-344.
- Magee, T., Lee, S., Giuliano, K., & Munro, B. (2006). Generating new knowledge from existing data: the use of large data sets for nursing research. *Nursing Research*, 55(2S), S50-6..
- Safron, C., Bloomrosen, M., Hammond, W., Labkoff, S., Markel-Fox, S., Tang, P., Detmer, D. (2007). Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*. 14, (1), 1–9.
- Saylor, J., Friedman, E., Lee, H. (2012). Navigating complex sample analysis using national survey data. *Nursing Research*, 61(3): 231-237.
- Smith A, Ayanian J, Covinsky K, Landon B, McCarthy E, Wee C, Steinman M. (2011). Conducting high-value secondary dataset analysis: an introductory guide and resources. *J Gen Intern Med*. 2011 Aug; 26(8): 920-9.
- Sweeney, L (2013). Matching known patients to health records in Washington state data. Harvard University. [dataprivacylab.org projects/wa](http://dataprivacylab.org/projects/wa) . Retrieved from <http://dataprivacylab.org/projects/wa/1089-1.pdf>
- Talbert, S., Sole M. (2013). Too much information: research issues associated with large databases. *Clinical Nurse Specialist*, 27(2): 73-80.
- Tucker, P. (2013). Business report: Has big data made anonymity impossible? MIT Technology Review. May 7. Retrieved from: <http://www.technologyreview.com/news/514351/has-big-data-made-anonymity-impossible/>
- Windle P. (2010) Secondary data analysis: is it useful and valid? *Journal of Perianaesthesia Nursing*, 25(5), 322–324.
- Yiannakoulias, N. (2011). Understanding identifiability in secondary health data. *Canadian Journal Of Public Health*, 102(4), 291-293.
- 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