Secondary Data Analysis: Non-Anthropological Datasets

ELIZABETH A. HOLDSWORTH

FOR THE DIGITAL ANTHROPOLOGICAL METHODS LECTURE SERIES

WASHINGTON STATE UNIVERSITY

WEDNESDAY, APRIL 13, 2022

Overview

- Overview of secondary data analysis in anthropology, examples from my research
  - Break for questions/discussion
- Fitting research question and dataset together
  - Break for questions/discussion
- Finding datasets
  - Break for questions/discussion
- Data management and analysis
  - Break for questions/discussion
- Summary of pros/cons and considerations

Relevant articles

Rosinger and Ice (2019, AJHB): Secondary data analysis to answer questions in human biology

In evolutionary anthropology:

- Mattison and Sear (2016, Human Nature): Modernizing Evolutionary Anthropology – introduction to special issue
- Stulp et al. (2016, Human Nature): The Reproductive Ecology of Industrial Societies, Parts I and II

Borre and Wilson (2017 in Food Health): Using secondary data in nutritional anthropology research; enhancing ethnographic and formative research

Thorne (1994 in Critical Issues in Qualitative Research Methods): Secondary analysis in qualitative research: issues and implications

Heaton 2004: Reworking Qualitative Data

Cook et al. (2018 in Advances in Archaeological Practices): Teaching open science published data and digital literacy in archaeology classrooms

Types of datasets

"Non-anthropological" – not collected for anthropological studies, but can be used to answer anthropological questions	Previously collected anthropological data
• Ex: health and behavior studies, geographic studies	• Ex: Reanalysis of Boas's skull database (Gravlee, Bernard, & Leonard 2003)

Benefits of a secondary analysis of a dataset

- Larger sample, sometimes across more time than could be feasibly managed for a dissertation project (ex: intergenerational longitudinal studies, global studies)
- May include costly biomarker data that you otherwise would not be able to get funding for
- May be representative of a population

Challenges particular to secondary data analysis

Time-intensive foundational research (Study sample, recruiting methods, data dictionaries, codebooks, data transformation, sample weighting)

Advanced statistical methods to account for the specifics of the sample

Have to be a bit creative in identifying how data can be used to answer your particular question within an anthropological theoretical framework

Why use a secondary dataset?

Compare a nationally-representative dataset to a study population

- Raichlen et al., 2017: compared cardiovascular markers between NHANES and Hadza

Comparison across populations

- Hruschka et al., 2014: Data from DHS across 47 countries to generate a *basal BMI*

Conduct a meta-analysis

Conduct an ecological study

Revisit an older dataset with new analytic methods

- Gravlee, Bernard, & Leonard 2003's re-analysis of Boas's skull data

Previous research projects

NHANES – National Health and Nutrition Examination Survey

- Acculturation and depression among Hispanic and Asian Americans (Human Biology Association)
- Have multiple tutorials: https://wwwn.cdc.gov/nchs/nhanes/tutorials/default.aspx
- Documentation is extensive: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx

Previous research projects

MIDUS – Midlife in the United States

- Effects of early life stress on physiological responses to chronic and acute stress in midlife in the United States (Human Biology Association)
- Found through ICPSR
- Similar methodology and questions to Midlife in Japan (MIDJA) which facilitates cross-country comparisons

My dissertation project as an example

Unequal distribution of stress → differential childhood growth

- Identifying social-emotional environment in infancy and epigenetic regulation of HPA-axis as key mechanisms

Conducted mixed-methods field research with a small sample of mothers to describe how stress was unequally patterned across a population and how it was reflected in mother's hair cortisol concentration

Secondary data analysis of the Avon Longitudinal Study of Parents and Children (ALSPAC) to have adequate sample size for epigenetic and growth analyses

How did I decide on ALSPAC?

Huge study that I was already familiar with through epidemiological and growth studies. Needed a dataset with longitudinal measures of child growth.

Checked study for inclusion of epigenetic data – yes!

Checked publications for any previous research conducted on these questions in this population (ALSPAC has a nice database for this) – no research previously conducted on these questions!

Steps to use the data set

1) Submit a research proposal

- Project abstract, aims/objectives, methods, variables to be requested (exposures, outcomes and confounders), impacts of research, reasons for using ALSPAC (8 pages approx., including references). Advisor was PI, myself and another committee member Co-PIs

2) Sign terms of agreement

- NOTE terms: I am required to publish gold open access, as well as send back any variables I derive, get approval from study admin before publishing anything, and take additional security steps to protect the data

3) IRB approval for secondary data analysis

4) List of variables I wanted from dataset

5) Approval from my NSF DDRIG Program Officer to use funds to purchase dataset

Benefits to this approach

Growth throughout childhood – I could not have done this myself due to time (unless I wanted my PhD to last 10 years!)

Large sample (over 1000 children had measures of growth) – I could not do this alone, without a large staff

Epigenome analysis – I could not obtain funds to cover these costs (NSF DDRIG is capped at $20k)

Developed both fieldwork skills and data analytic skills

Able to continue research on this dataset for many years to come

- And also share with students, as long as the research questions reasonably fall under the overall project goals

Set up my research program with a good foundation to look at these relationships in different populations and environments

What I wish I had done differently

Asked my “data buddy” to check how many participants had overlapping data

- ~1000 had epigenetic data, ~1000 had growth data, but only ~120 had both

Requested more variables

- I have less sociodemographic data than I would like

Questions on my experience?

Matching research question to dataset

Matching your research question to a dataset

What study design is relevant to your question?

What sample do you need?

- Should data be collected on individual, household, locality, or national level?

What variables (and relevant covariates do you need)?

Should you use multiple datasets?

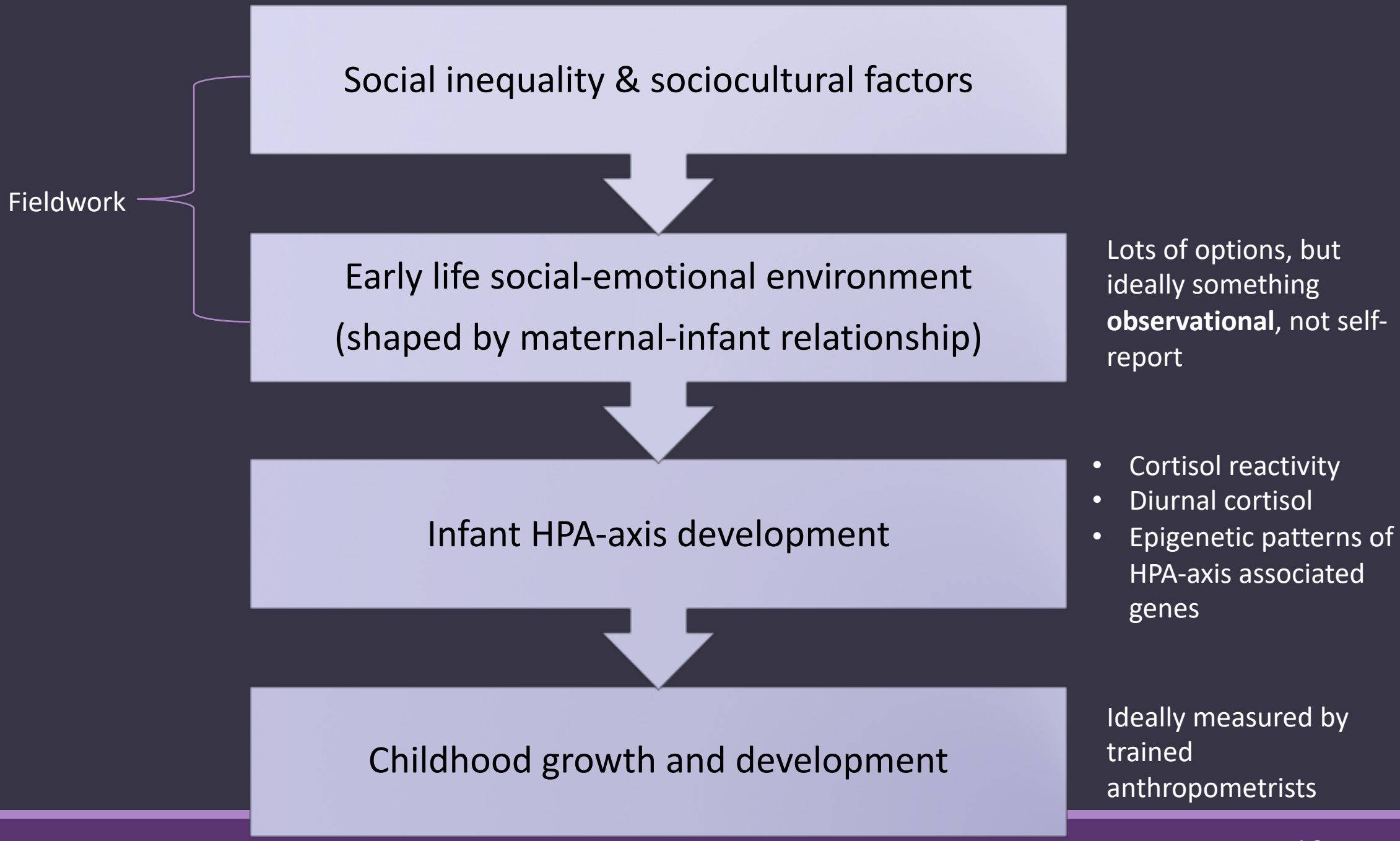
Matching research question to dataset is an iterative process

Recommend conceptual mapping of your larger theoretical questions and identifying a couple discrete questions that might be answerable by another dataset

- Identify a few possible different biomarkers or behavioral or material correlates

If one variable is not directly measured, you may be able to approximate the construct by combining variables

- Ex: allostatic load

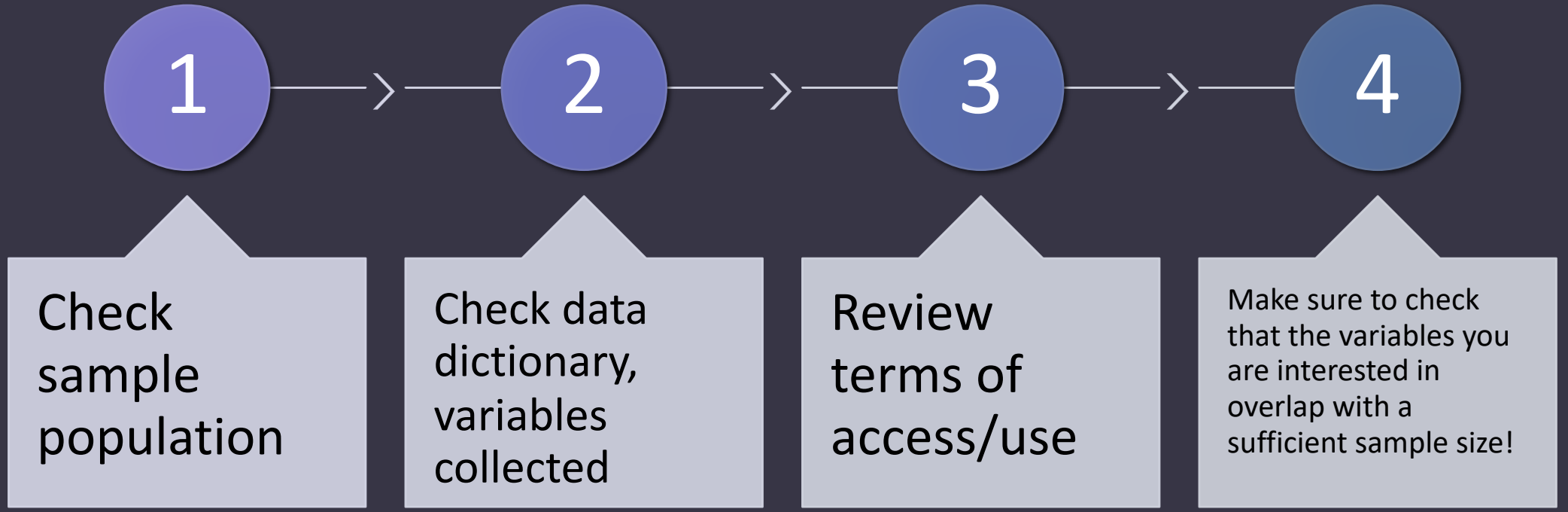

Social inequality & sociocultural factors
Fieldwork
Early life social-emotional environment
(shaped by maternal-infant relationship)
Lots of options, but ideally something **observational**, not self-report
Infant HPA-axis development
• Cortisol reactivity
• Diurnal cortisol
• Epigenetic patterns of HPA-axis associated genes
Childhood growth and development
Ideally measured by trained anthropometrists

Anthropology with non-anthropological datasets

Anthropological methods (cross-cultural perspectives, ethnographic and field methods) are unlikely to be encompassed in non-anthropological datasets

- May have to choose between ideal methods and ideal populations

Confirming the dataset works with your research question



Some creative uses of secondary datasets

Julie Lesnik (2017): combined data from World Bank, World Factbook, and World List of Edible Insects

Louis Alvarado (2010): Testosterone and prostate cancer risk from a reproductive ecological framework, using data from 12 different studies

Oskar Burger et al. (2011): Combined data from World Resource Institute (Earthtrends database) on per capita energy use. Life history traits from WRI, previously published research, and Selected Indicators from the United Nations Population Division. Answered question of how extra-metabolic energy sources may impact human demographics and life history traits.

Any questions on refining research question?

Finding Datasets

Journals and grants may require datasets to be shared

Check journal requirements on repositories

- https://www.nature.com/sdata/policies/repositories

Where do I find datasets?

Data repositories:

- ICPSR
- Dryad
- Dataverse
- Figshare
- List of repositories from Scientific Data (Nature)

US Federal Data:

- data.gov
- healthdata.gov
- NIH
- DHS (global data)
  - Spatial Data Repository
- NHGIS

WHO and global data:

- World Resources Institute
- WHO multi-country studies archive

Topics-related sites:

- g2aging.org
- Digital Index of North American Archaeology (DINAA)

UK studies

- Closer discovery

Database search engine:

- Google's Dataset Search

Published articles and other researchers

Some highlighted datasets

- Adolescent brain cognitive development (ABCD)
- Behavioral risk factor surveillance system (BRFSS)
- Cebu longitudinal health and nutrition survey (CLHNS)
- China health and nutrition survey (CHNS)
- Demographic and Health Surveys (DHS)
- National Family Health Survey, India
- National Health and Nutrition Examination Survey (NHANES)
- National Health Interview Survey (NHIS)
- National Longitudinal Survey of Youth 1979
- National social, life, health, and aging project (NSHAP)

Some highlighted datasets

National Survey on Drug Use and Health (NSDUH)

The Russia Longitudinal Monitoring Survey

Tsimane Amazonian Panel Study (TAPS)

UNICEF Multiple Indicator Cluster Surveys (MICS)

WHO Study on global aging and adult health (SAGE)

Finding datasets

Keep in mind that some datasets may have all the variables you need even if the main goal of the study does not seem related

Any questions on finding datasets?

Let’s try searching for datasets!

Demographic and Health Surveys (DHS)

ICPSR

Google’s Dataset Search

NIH

Figshare

Dataverse

Dryad

Data Management and Analysis

Data management

See Rosinger and Ice's article for details on this

Need to consider how to merge data, identify missing variables

May also need to transform your dataset from wide to long, or vice versa

This requires you to know how the data is structured

Save your generated datasets under different names, save your derived variables under different names

Check your n at each step

Data analysis

Know your study design – are there sampling weights? Clustering?

Univariate analyses first! Check for outliers, odd distributions, missing data, and reference the codebook or data manual for clarification on how the data was collected

- Was any variable transformed or revised before inclusion in the dataset? Were there problems in collection?
- Check codebook for information on how missing (or “don’t know”) data was coded

Document your process!

- Save all syntax you write, and annotate and date your decisions/coding
- Document your dataset construction, data cleaning, variable construction, analysis, figures/tables

Questions on data management and analysis?

Summary

Benefits to secondary data analysis

Learn best practices in data management and analysis

Contribute to a scientific need to analyze vast amounts of data collected (Mattison & Sear 2016; Stulp et al., 2016)

Can use grant money to conduct these analyses, particularly as part of a mixed-methods study

Can answer research questions you otherwise could not

Challenges/downsides to secondary data analysis

Have to build other field skills in other ways

May not be able to answer the exact questions you have in mind for your research

Data management and data analysis will be time consuming

You may be constrained by contracts of data use

Steps to conduct secondary data analysis

1) Draft your research questions, hypotheses, and material/biological correlates
2) Review other publications in your field for other secondary data analysis
3) Search for datasets with relevant variables and relevant populations
4) Review dataset documentation, publications, terms of use
5) Revise research questions/hypotheses/correlates as necessary
6) Identify total list of needed variables
7) Obtain dataset (may need to draft research proposal for access)
8) Data management – cleaning data and linking IDs if necessary
9) Data analysis

Contact me in the future!

e.holdsworth@wsu.edu

e.a.holdsworth5@gmail.com

Twitter: @eaholdsworth