FRED HUTCHINSON
CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research

Betsy Rolland, MLIS PhD, Fred Hutchinson Cancer Research Center
Charlotte P. Lee, PhD, University of Washington

HCDE Human Centered Design & Engineering
University of Washington

Small Data Reuse is Hard

From *The New York Times* to *Nature*, Big Data is a hot topic, while Small Data rarely gets mentioned. In many biomedical research fields, such as cancer epidemiology, small datasets collected for individual studies, when combined or used for new analyses, represent a potential for new discoveries similar to that attributed to Big Data. However, reuse of these small data sets is fraught with challenges.

Small datasets can be difficult to find, as they are rarely deposited in repositories, but, rather, live on investigators' hard drives or lab servers. Documentation is often informal, spotty or nonexistent except for minimal details documented in the methods section of published papers. These issues create enormous potential for misinterpretation and misuse of the data.

Funding agencies and open science activists are increasingly calling for sharing of research data, but we still don't have a firm understanding of how the eventual recipients will use the data being shared. As a first step toward that understanding, this research set out to answer the question, **How do cancer epidemiology postdoctoral researchers determine how to use a variable from an existing dataset appropriately for their own analyses?**

Research Site & Methods

Eleven post-doctoral researchers in cancer epidemiology at the Fred Hutchinson Cancer Research Center were interviewed for this study using a semi-structured interview protocol. Questions focused on how post-docs work to understand data, that they did not themselves collect, for use in new analyses. The interviews were transcribed, then analyzed using a grounded theory approach.

Rolland B and Lee CP. (2013). "Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research." Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW). San Antonio, Texas, USA, ACM: 435-444.

Contact and Funding

Betsy Rolland, MLIS PhD

Project Manager
Asia Cohort Consortium Coordinating Center
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North, M4-B402
Seattle, WA 98109-1024
brolland@fhcrc.org

This research is supported by the Fred Hutchinson Cancer Research Center and the NIH (R03CA150036).

A Typology of Information Needs: Nine Questions Post-Docs Asked When Reusing Data

Post-docs needed answers to a series of questions as they progressed through their project, beginning from the proposal stage through submission of the manuscript for publication.

Question Type	Examples
1. What datasets are available to use in my research?	What datasets does my advisor have access to around my research area?
2. Will this dataset help me answer my research question?	Does this dataset have the population, study design and outcome data I need?
3. What else has been done with this dataset?	What manuscripts have been previously published from this study?
4. Where do I find the information I need to understand this study?	Where can I find copies of the questionnaire and codebook? Which copy of the dataset should I use?
5. How was this dataset constructed?	On what basis did the original study include or exclude participants? Why was this question included on the questionnaire?
6. How were these data constructed?	How did the original study staff code my variables of interest? Why are so many participants missing data on the variable "alcohol use"?
7. What do my variables of interest mean?	How did the original study define the variables "heavy smoker" and "frequent aspirin user"?
8. Am I using the data and the dataset correctly?	Does my interpretation of the meaning of this variable match that of previous data users? Did I select the correct data file?
9. What have I done with this dataset?	How did I define my categorical variables in my analysis and how should I represent that in my manuscript?

Most Frequent & Difficult to Answer: Question #6 - How were these data constructed?

Data are highly social, reflecting the values and practices of those who engage with the research and the data[1]. Data are also constructed as a result of the plethora of small, often seemingly insignificant, decisions and actions taken during a research project. However, knowledge of such decisions or actions often lives exclusively in the heads of those who took them, not through negligence but through pragmatism.

It is simply impossible for researchers to document each decision formally while still moving their work forward. These decisions and actions may or may not be apparent in the data as irregularities or missing data. In either case, they can have a profound impact on the outcome of the data analyses if not understood, as they reflect the myriad of assumptions and interpretations, some explicit and some tacit, made by researchers and study staff along the way. Questions post-docs asked about the construction of variables in their datasets focused not just on the meaning of the variables, but on the decisions and actions that led to the current state of the dataset as presented to the post-doc.

> *I had received a variable for exposure to [medication] use that was months of [medication] use, and I assumed from the labeling of the variable in the dataset … that an individual was coded as a 1 if they had used at least one month of [medication], and so I said that our study excluded women who didn't have less than a month of [medication]. But apparently, women with one day of [medication] were counted as a one because it was really a zero to one month. And that took forever to figure out … because then [the data manager] had to go back to the original code in which she had created the variable and reinterpret the code to break down exactly what had happened, and it was like all this looping. So things like that were frustrating. We had a lot of setbacks where you're like, "So what does this variable really mean?" and every time you ask, "What does this variable really mean?" it's never straightforward (Ginger, 55).*

> *It's also just sometimes there's nuances you want to understand. You know, someone's colonoscopy history, there's some questions about that, and you want to understand it, trying to rule out the colonoscopy that was used to diagnose the colorectal cancer. So you want to know if they were screened ever. And actually, that's kind of hard in this dataset because of the way the questions were asked, which is unfortunate. But so you're using like a time between diagnosis and when they say they had a colonoscopy to infer maybe what that was about. … And you have to go back to the data dictionary, for sure, but I find myself going back to the questionnaire to see what I think the question really was asking, you know. Because the data dictionary… it's just descriptive, it's kind of what they thought they were asking… it's like oh, the value can be one or two. One is yes and two is no. And, you know, it'll say had a colonoscopy [Laughter]. But when you look at the question, it's have you ever had a colonoscopy, you know, more than two years ago or something? So there's a difference. So there can be differences (Stewart, 162).*

[1] Birnholtz, J.P., & Bietz, M.J. (2003). Data at work: supporting sharing in science and engineering. In Proc. ACM SIGGROUP, ACM Press (2003), 339-348.

Research and Policy Implications

Bottom Line:
In this study, users of data someone else collected required interaction with the original study personnel in order to understand the data well enough to use it successfully. Data sharing is challenging and, if done poorly, can lead to data that are misinterpreted and misused, creating enormous potential for erroneous conclusions.

More research is needed to determine:
- What is "enough" or "thorough" documentation?
- How can researchers produce targeted documentation that supports data reuse, rather than just more documentation?
- How can projects easily document and store the study history without taking too much time away from the science?

Policy:
- What types of incentives from funders will encourage documentation and data curation?
- How much of a project's funding should be allocated for use in documentation and data curation to ensure funding agency's investment in the data?