Post-Doctoral Researchers' Use of Preexisting Data in Cancer Epidemiology Research

Betsy Rolland
Human Centered Design & Engineering;
University of Washington Public Health Sciences Division;
Fred Hutchinson Cancer Research Center
1100 Fairview Ave N
M4-B402
Seattle, WA 98109
brolland@fhcrc.org

Charlotte P. Lee
Human Centered Design & Engineering;
University of Washington
423 Sieg Hall
College of Engineering
Seattle, WA 98195
cplee@uw.edu

ABSTRACT

While calls to reuse research data are increasingly loud and urgent, little research has been done on how those data are subsequently used, especially in the field of cancer epidemiology. We interviewed eleven post-doctoral researchers working in the field of cancer epidemiology at the Fred Hutchinson Cancer Research Center in Seattle, WA. Despite the availability of high-quality written study documentation, post-docs still had unanswered questions about the data that required interacting with PIs and study staff of the original project. These questions fell into three categories: (1) study design; (2) a variable's origins or coding procedures; and (3) variables with unfamiliar measures. Without answers to these questions, post-docs were unable to complete their analyses in a scientifically responsible way.

Author Keywords
Data reuse; Data sharing; Data intensive science; Cancer epidemiology; Qualitative research methods

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors; Design

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.

INTRODUCTION

Cancer epidemiologists have a long history of sharing data with colleagues, including their post-doctoral researchers. Reciprocal sharing between and among investigators who know and trust one another generally results from relationships developed through a shared history or network. Such relationships guarantee a reasonable understanding of how, when, and why the original research has been done. They also assure personal access to the original investigators in case of questions about the background of the research and a grasp of the original motivations and hypotheses. That connection between the original investigator and the dataset recipient is being lost as journals and funding agencies demand that researchers make data available in vast repositories as a condition of publication or funding.

Yet these raw data tell only part of the story, leaving out crucial contextual details essential to interpretation and potentially leading to the mistaken assumption, "that the necessary syntheses of raw data can be performed automatically." [1] Without such contextual details, researchers are more likely to misinterpret what a given data point means or utilize data incorrectly, leading to bad science.

As science becomes more data intensive, larger and more complicated data sets are being made available to researchers with no connection to the original study personnel. This raises important questions of how usable a data set is without that connection and whether there are types of data that are simply impossible to use without interaction with the original data collectors. Such interactions are costly and may deter scientists from sharing simply to avoid this burden; however, having no interaction where one is required has the potential to lead to serious errors.

So, if we take as our goal helping people share data effectively, we need to understand how shared data can be interpreted appropriately in order to be used in a way that is scientifically valid. We also need to understand if this can be done without access to the original study staff. To answer this question, we interviewed post-doctoral researchers at the Fred Hutchinson Cancer Research Center about their use of preexisting data.

BACKGROUND

Post-docs in the field of cancer epidemiology are generally brought in to do new analyses on a set of data that have already been collected. As such, they represent a group that is just one step removed from the data collection. We think that their experience can help us start to understand the challenges facing someone as they use data they did not themselves collect. We started with the following research question: How do cancer epidemiology post-doctoral researchers determine how to use a variable from an existing dataset appropriately for their own analyses?

As discussed, scientists are increasingly being pushed to share their data. In a recent comment piece in the Lancet, the leaders of two major funding agencies, the Wellcome Trust and the Hewlett Packard Foundation, asserted that not sharing data is harming public health research. [5] They propose an agenda for discussion and immediate steps to take to solve this issue, one of which is developing standards around data collection in public health research. This call to sharing was also signed by the leaders of other major foundations and funding agencies.

There seems to be general agreement that data sharing can have tremendous benefits, if done well. These include the ability to ask larger questions than any individual study might be able to ask on its own, simply due to a larger and more complex data set; the ability to take advantage of existing datasets rather than spending time collecting new data; and potentially large cost savings because there is no need to assemble new projects.

Of course, the field of CSCW has also documented significant challenges of data sharing. These include difficulties in the interpretation of data and appropriately conveying context [3]; issues of trust in the quality of data [6]; and a reluctance to give away intellectual property [3]. Baker and Yarmey (2008) define context as "the properties of the broader physical environment in space and time and is recorded in the accompanying metadata. The context (and thus the metadata) also includes the technical and social environments comprised of instruments, people, traditions and organizational entities associated with obtaining the measurement as well as the later processing, storage, use, transport and reuse of the resulting data" [2]. This definition makes clear that context is complex and difficult to document.

Faniel and Jacobsen (2010) document many of the challenges inherent in data reuse by earthquake engineering (EE) researchers. [4] Their study identified three factors EE researchers consider when thinking about the reusability of the data of colleagues: (1) relevance; (2) their ability to understand the data; and (3) trustworthiness. They note that the reusability of data is rarely discussed in the recommendations and reports from funding agencies, even as those agencies press for a greater level of sharing. Faniel and Jacobsen further emphasize that is not simply the availability of data that will spur reuse and, ultimately, scientific innovation, but, rather, the availability of data is easily reused.

So, funders and publishers are demanding not just sharing but deposit of data into repositories. Yet they rarely address the question of how usable those data are once they've been deposited and give little support to researchers in either sharing their data or using the data of others.

CANCER EPIDEMIOLOGY

Epidemiology is the study of disease risk at the population level. Some examples of populations studied by epidemiologists may include post-menopausal women, prostate-cancer survivors, or children. Studies focus on exposures such as tobacco use, family medical history, diet, or proximity to contaminants. Cancer-epidemiology datasets possess a number of characteristics which make them particularly interesting for CSCW researchers to study. First, epidemiologic data are generally collected by questionnaire. Questions must be straightforward enough for a non-scientist to administer them and for a participant to answer them. While epidemiology questions are not standardized, there are some generally accepted practices which make it easier for researchers to understand one another's data sets. For example, anyone collecting lifestyle data will want to collect smoking data with both frequency and duration in order to produce the derived variable "pack-years". Additionally, epidemiologists are asking similar questions over different populations. Studying the effect of smoking on breast cancer in post-menopausal women is not that different, at a high level, than studying the effect of smoking on prostate cancer in Hispanic men.

Most importantly, cancer epidemiologists have a history of sharing data within their trusted social networks. This is

partially attributable to the similarity of data collected but also speaks to a core belief within the cancer-epidemiology community that a data set that isn't being used is worthless. As funding for new projects and new data collection gets tighter and computing power increases, it is quite likely that this sharing will increase, yet little study has been done on how cancer epidemiologists share and use preexisting data sets.

RESEARCH SITE AND METHODS

This research took place at the Fred Hutchinson Cancer Research Center (FHCRC), an NCI-designated cancer center in Seattle, WA. This class of research institute is categorized by a high level of scientific excellence and patient-centered research. FHCRC has approximately 3,000 employees and is organized into five divisions. One of these divisions is the Public Health Sciences Division, home to most of the organization's cancer epidemiologists.

Fred Hutchinson Cancer Research Center (Photo from FHCRC.org December 2011)

Post-doctoral researchers at FHCRC are generally brought in to do new analyses on data that have already been collected, either earlier in the project or for a completely different project. They work with a specific mentor, most often the PI who originally collected the data under analysis. In some cases, post-docs were written into grants and then hired; in other cases, they were hired under training grants and then found a project within FHCRC.

Over the course of six weeks, we conducted semi-structured interviews with a diverse group of post-doctoral researchers in cancer epidemiology at FHCRC, including four men and seven women. Interviews lasted between 30 minutes and one hour and were conducted at FHCRC. Participating post-docs hailed from several different institutions and fields, including medicine and public health (MD/MPHs), behavioral epidemiology and molecular epidemiology. They worked with different mentors and were at different points in their post-docs, ranging from 3 months to 2 years. After the interviews, we analyzed the transcripts using a grounded theory approach.

FINDINGS: WRITTEN INFORMATION SOURCES WERE NOT ENOUGH

Below we present findings based on a preliminary analysis of our data. Participants were given a variety of written information sources when they began their data analysis projects. Interestingly, the types of sources given to them were remarkably consistent across our interviews, possibly representing a standard set of information sources available to post-docs at FHCRC or even within the field of cancer-epidemiology. These written information sources included:

- Questionnaires and Codebooks
- Websites with variable lists
- Grants (previous and current)
- Published manuscripts from this dataset
- Comments in analysis code

However, all participants reported having unanswered questions about their data sets, questions the written documentation could not answer. In order to complete their analyses, they need to interact with the original study personnel to gain additional contextual details missing from the documentation. These contextual details were generally those that had not been important in the analysis done on the data set for the original study but were crucial for the new analyses being performed by the post-docs. Personnel consulted ranged from the original PIs, study and data managers, to other post-docs. The person they chose to interact with on a given question depended on the information they were seeking, as well as their relationship with the personnel. For example, some post-docs were reluctant to bother a senior PI mentor with a question about data collection methods, and took those to data or study managers instead. Others took all questions to their PI because they were the only one with the necessary knowledge.

These written and human information sources were used at all stages of participants' projects. The majority of the projects had websites with variable lists that participants could use in selecting their variables for their project proposals. Grants, previously published manuscripts and codebooks were also used at this stage, as were any of the people listed above. Once the project had been approved and the data received, participants used their information sources to clean the data, often checking their results against published analyses. At the stage of descriptive statistics, participants reported especially heavy usage of questionnaires and codebooks, as well as the data manager. It was at this stage that they made decisions about which participants to include or exclude based on missing data, so ensuring a full understanding of why those data were missing was crucial, as discussed below. As participants completed their analyses, they again went back and compared their findings with accepted scientific findings published in journals and consulted analysis code

previously written by other post-docs. Finally, as they wrote their analysis into manuscripts, participants consulted the comments in their own analysis code to remind themselves of how they had set up their analyses.

The questions participants asked about their data sets ranged from simple to quite complex, requiring a quick email or a lengthy recreation of coding a derived variable. However, it is interesting to note that the topics covered by the questions were not infinitely complex but, rather, fell into three categories.

QUESTION #1: STUDY DESIGN

> "Well, I mean, the biggest one that I've struggled with the most, really, is the tumor stage. Some of that has to do more with administrative reasons that not every site within the [project] … transmits the same information. So for whatever reason … they don't have whether or not the tumor has metastasized. And so I didn't understand that variable in the sense that there's a lot of missing variables that they just don't have that data." – Stewart

Because the post-docs did not participate in the original study collection, they were not aware of the minutiae of the study design, contextual details that had not been documented but that original study staff knew. In the interview above, Stewart had been previously unaware that human subjects rules in the country where the data had been collected prohibited a specific study site from sharing a data point. Without that knowledge, he would have been forced to eliminate those subjects from that study site from his analyses due to the missing data.

Other participants identified issues with collection procedures of biologic samples that had been documented but not as thoroughly as was necessary for their own analyses. Population sampling techniques was another frequently mentioned study design detail missing. Post-docs wanted more details about how the study population had been decided upon and why.

QUESTION #2: A VARIABLE'S ORIGIN OR CODING PROCEDURES

> "…sometimes when a variable was measured was confusing because some were measured at diagnosis, and some were measured at the reference date. ... Some were measured over the interval, and then you had this categorization into current use and recent use and how those were defined ... there were all these different ways to categorize these things, and that was a little confusing." – Ginger

Ginger was investigating the link between a common prescription drug and the risk of a second cancer, so time was an important factor in her analysis, whereas the original analyses of the study had not used this variable. After her analysis turned up surprising findings, Ginger and the data manager retraced the steps taken to code the medication variable. In the end, they discovered that the coding procedure was inappropriate for her analysis, and they recoded the original study data with more precise time information to better suit her needs.

Participants made clear that sometimes the variable's origin or coding procedure was irrelevant, but when it was unclear or directly affected their analysis, they needed deeper information than was available in the study's codebook or in previously published papers.

QUESTION #3: VARIABLES WITH UNFAMILIAR MEASURES

> "And then the other thing that was difficult was using some of the treatment variables. And again, I don't know if it was necessarily - I don't think it's how the data was set up, but it was just me learning how to use that kind of data, especially using things like they have certain sort of scoring measures that they use for chemotherapy dose and radiation dose and things like that and really understanding. Not being a clinician, making sure that I'm using the types of variables appropriately and scoring them right and that sort of thing. So I'd say those were - I wouldn't even say it's necessarily the data part of it. It's what that data means was the challenging part to me, and I think that's probably kind of standard across some of these other data sets that I've used that sort of our clinical epidemiology type data sets is really understanding what's in there." – Abbie

The third category of questions described by our participants was that of variables with unfamiliar measures. In the quote, above, Abbie is describing the difficulties she faced, as a non-clinician, when trying to integrate clinical data into her analyses. Here her primary focus is on making sure she uses the data correctly to ensure the integrity of her results. In order to do so, she must consult with clinicians who specialize in the area of the treatment variables.

DISCUSSION

What we see, then, is that information-sharing interactions occurred at all stages of the analysis process. Questions arose when post-docs were first writing their proposals to use the existing datasets, when they were first poking around in the data, when they were performing their analyses and when they were writing up their results. Participants first tried the written documentation which had been provided to them, then sought out appropriate human

information sources to augment their knowledge. This was not always an easy process, but sometimes required weeks of work to answer the question, either by tracking down an original study member or retracing the steps taken to clean and store the data.

Foremost in participants' minds was always their responsibility to use the data in a scientifically responsible way. The questions they asked were not frivolous, stemming from mere curiosity, but rather, sought important contextual details that had not been documented in the original study, but without which data use was impossible. It is important to stress here that the fact that these details were not documented is not a failure, per se, of the original study. Much of the knowledge about how a study was conducted remains tacit or appears superfluous to the results. What our participants point out is that it is impossible to document contextual details of a study so thoroughly that a future ancillary study will be able to seamlessly use the data. There is simply no way to predict all future uses to which a given dataset may be put.

What, then, is legitimate to expect of researchers in the realm of data sharing and deposit? We believe that these results show that in order for cancer epidemiology data to be reused, the link between the data and the PI may need to be maintained and certain information sharing practices supported. Preparing data to share and answering subsequent questions about those data is labor intensive and expensive. If future users are unable to understand the data sufficiently to use them in a way that is scientifically appropriate and responsible without making further demands on the original study team, those efforts are wasted.

This research represents a first step toward understanding how cancer epidemiologists determine if they are using variables appropriately when they did not themselves collect the data. We have shown that cancer epidemiology post-doctoral researchers are unable to use preexisting datasets without access to both written and human information sources. They require additional contextual details not documented in codebooks or published manuscripts. Analysis of this data is ongoing and will further detail how researchers go about finding the information they need to use variables appropriately within the three classes of questions detailed above.

FUTURE WORK

In the near-term, future work will also explore two additional questions:

- How can PIs, study coordinators and data managers ensure that their data are used appropriately by others?
- How do users of data repositories without access to original study personnel interpret data?

The long-term goal of this research is to produce a qualitative model of how epidemiological data is shared for the purpose of promoting a principled way of designing databases and associated tools and systems to support data sharing. We also want to understand further what is the relationship between the categories of questions and the types of projects on which the researchers are working.

ACKNOWLEDGEMENTS

We gratefully thank our participants for their time and cooperation with this research study. This work was supported by the Fred Hutchinson Cancer Research Center and by the National Institutes for Health (R03CA150036).

REFERENCES

1. Agre, P. and M. Rotenberg (1997). Technology and Privacy: The New Landscape, MIT Press.
2. Baker, K. S. and L. Yarmey (2009). "Data stewardship: Environmental data curation and a web-of-repositories." International Journal of Digital Curation 4(2): 1-12.
3. Birnholtz, J. P. and M. J. Bietz (2003). Data at work: supporting sharing in science and engineering. Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work. Sanibel Island, Florida, USA, ACM: 339-348.
4. Faniel, I. M. and T. E. Jacobsen (2010). "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data." Computer Supported Cooperative Work 19(3-4): 355-375.
5. Walport, M. and P. Brest (2011). "Sharing research data to improve public health." The Lancet 6736(10): 9-11.
6. Zimmerman, A. (2007). "Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse." International Journal on Digital Libraries 7(1-2): 5-16.

BIOGRAPHIES

Betsy Rolland, is a PhD student in Human Centered Design & Engineering (HCDE). Betsy holds a BA in Russian Languages and Literature from Northwestern University and a Master of Library and Information Science from the Information School at the University of Washington. Her coursework at the UW focused on the development of information systems for biomedical research collaborations and included courses in HCI,

usability testing and biomedical and health informatics. She is the Project Manager for the Asia Cohort Consortium Coordinating Center at the Fred Hutchinson Cancer Research Center in Seattle, WA. The Asia Cohort Consortium is a biomedical research collaboration comprised of more than 20 Asia-based cohort studies seeking to build a collective cohort of over a million people. Her research in the PhD program focuses on the coordination and support of collaborative biomedical research. Betsy recently completed a research project, funded by the Special Libraries Association Research Grant, in which she and a colleague studied how librarians are using their traditional library-based skills in non-traditional ways in biomedical research.

Dr. Charlotte Lee has a B.A. in Sociology from the University of California, Berkeley, an M.A. in Sociology from San Jose State University and a Ph.D in Information Studies from the University of California, Los Angeles. Dr. Lee's research is in the fields of Computer Supported Cooperative Work (CSCW), Social Informatics, Design Studies, and Science and Technology Studies. Her work focuses on empirically describing and theorizing the informational practices, artifacts, and collaborative structures of communities of practice working towards a shared goal: collaborative design. Her paper entitled the "Human Infrastructure of Cyberinfrastructure" was nominated for the Best Paper Award at the ACM's Conference on Computer Supported Cooperative Work. Dr. Lee is the principle investigator of three NSF-funded projects studying aspects of collaboration in the development of cyberinfrastructure including a National Science Foundation CAREER Award for "junior faculty who exemplify the role of outstanding teacher-scholars" awarded in 2010. She is also on the Editorial Advisory Board of the Journal of Computer Supported Cooperative Work.