The Research Data Life Cycle and the Probability of Secondary Use in Re-Analysis

Amy Pienta (apienta@umich.edu)
James McNally (jmcnally@umich.edu)
Myron Gutmann (gutmann@umich.edu)

Inter-university Consortium for Political and Social Research
Population Studies Center
University of Michigan

Paper submitted for the Poster Session of the 2006 Meetings of the Population Association of America.

Abstract

The life course of a funded research project represents an important, yet largely understudied phenomenon. Like any demographic event, a research project follows a measurable trajectory; passing through a series of transitional phases leading to completion. This poster will present analysis on the outcomes that primary data collected through funded research can follow. Specifically, we analyze the likelihood that primary research data will be released for secondary analysis and identify factors that increase the risk that data will remain unavailable for re-analysis. We will present specific solutions to common barriers faced that inhibit the transition from primary to secondary data. The benefits of secondary data analysis are undeniable, with a better understanding of what incentives encourage the release of research data into the public domain we can better assist researchers in developing strategies that allow them to add their data to the growing collection of secondary data resources.

Introduction

The life course of a funded research project represents an important, yet largely understudied phenomenon. Like any demographic event, a research project follows a measurable trajectory; passing through a series of transitional phases leading to completion. Tracked and evaluated during its formative life course, the research process employs progress reports, field testing, and preliminary findings to offer insight into the success and potential value of the specific project to the scientific community. Often however, the information flow regarding a study terminates with the end of the funding cycle. With the completion of the primary research stage, many valuable studies can enter a "dormant state" where final reports and publications are completed and the research team moves on to a new research project. This *primary life course* of a research project is a well-established routine and the work emerging from primary research has added immeasurably to our understanding of a wide variety of demographic, health, economic and social behaviors.

Our research suggests that the primary life course of a research project represents only the first phase in a multi-state process in the *total life course* of the research data that serves as the foundation of any project. We argue, and existing practices fully support

this assumption, that there is a tremendous need to better develop our ability to identify and support the transition of research data from their use in primary research into the next stage of it life course, that of secondary research.

This transition is increasingly normative in many areas of social science research. Large scale federally funded surveys increasingly enter the public domain as secondary data and there is a long tradition among federal agencies such as NCHS and the Census Bureau to provide many of their surveys to the research community in a form that protects respondent confidentiality. These initiatives have set the stage for the explosion of secondary data now available, but it has also obscured the fact that the vast majority of primary research data lay dormant at the end of the *primary life course* phase and the opportunities these data collections could represent as secondary resources are not realized.

We do not suggest that this is a result of an unwillingness to allow research data to enter the public domain, although the culture of data sharing varies across disciplines. We argue instead that the growth of primary investigation and the emergence of research tools that facilitate the creation of primary data have resulted in an exponential growth of data that are difficult to organize. The development of a research strategy to evaluate this problem forms the core of the project described in this abstract. How do we systematically review funded research, identify potential sources of primary research data, and ultimately encourage the transition of this data from the *primary life course* into the *secondary life course*? What steps are required to encourage a culture that sees data as having ongoing value rather than as a terminal product whose value ends with the originating project?

Researchers at the Inter-university Consortium for Political and Social research (ICPSR) have been collaboratively addressing this issue. We have focused our initial work on research projects funded by NIH, which is a logical and straightforward starting point, especially because of the broad role of NICHD and NIA in funding demographic research. In our project, we investigate the extent to which NIH-funded social and behavioral science data have been shared.

Data and Methods

NIH uses a search web interface system call CRISP (Computer Retrieval of Information on Scientific Projects). CRISP is a searchable database of federally funded biomedical research projects conducted at universities, hospitals, and other research institutions. Maintained by the Office of Extramural Research at the National Institutes of Health, the database includes projects funded by the National Institutes of Health (NIH). CRISP also provides information on other federal agencies including SAMHSA, HRSA, FDA, CDCP, AHRQ, and (OASH) but for the current project we restricted our analysis to NIH based centers and institutes.[1] We used this source for three reasons: (1) NIH supports a variety of social and behavior research programs; (2) NIH is committed to data sharing; (3) NIH's database of awards is accessible through the internet.

[1] From the CRISP Webpage http://crisp.cit.nih.gov/

Piloting the LEADS Procedures and Decision Criteria: Our initial sample universe was restricted to new projects awarded funding in 2000 or 2001. The CRISP database contains more than 160,000 records for these two years. To make screening more efficient, we relied on front-end restrictions to the sample frame development. Many existing research projects apply for extensions or supplemental funds. Investigators move to new institutions and bring the grant with them. All of these administrative events trigger new entries in the CRISP system. Since we are interested only in new awards, we systematically excluded administrative entries. This brought the total number of awards in the sample frame to approximately 24,000.

These 24,000 abstracts were next screened using two key criteria: (1) does the study constitute ***behavioral or social*** research and (2) does the study ***collect new quantitative*** data? The screening team used information provided through the CRISP database to make these decisions. Most entries include an abstract (ranging from 100 to 500 words) which (in most cases) provides adequate detail to make a determination. Some records do not include abstracts, but provide sufficient detail in the study title to make exclusion judgments. For each of the records reviewed, the screeners make one of three judgments: (1) the study meets criteria, (2) the study does not meet criteria or (3) it is unclear whether this study meets criteria. Decisions were made based on the abstract, study title, and thesaurus terms available from CRISP. The abstract was the best source of information. The title while less useful, provided additional clues about the purpose of the study. Several hundred records are missing abstracts. In these cases, the title and thesaurus terms were used to apply decision rules.

NIH funds a wide variety of research projects. In order to be considered for this study, the abstract is required to describe a **social/behavioral** research project. We define behavior as any human activity that may be observed unobtrusively. Though biological, neurological, and other genetic processes may influence behavior, we did not consider these processes unobtrusively observable. Unobtrusively does not rule out experimental, or instrument based research [i.e., a study focusing on cognitive flexibility in children which asks the child to perform tasks in a laboratory environment may still be considered behavioral research]. The research needed to be concerned with actual human beings, not computer simulations or animal models. Also, studies concerned with group behavior (e.g. disease rates across categories of people) have been included.

We also **defined original** to be newly collected data. This could include a new survey or experiment; a data file containing information extracted from medical records (which were not previously prepared for statistical analysis), or any other gathering of new information in a numerical form. Re-analysis of existing data, sometimes called "secondary analysis" did not meet these criteria. Finally, **quantitative data is defined as** any information that may be analyzed statistically. Surveys, record abstractions, experiments coded numerically, etc are quantitative. *Qualitative* data (e.g., open ended narrative interviewing intended for non-numerical analysis), traditional participant observation, ethnography, oral history, and narrative or visual analyses do not meet our criteria.

From this pilot screening project, we were able to determine which NIH institutes funded the largest proportion of social and behavioral research. Table 1 reflects the NIH funded activity areas we reviewed as part of this research project. The three largest centers support social and behavioral science research (NICHD, NIA, and NIMH) were given intensive scrutiny, but we also examined the activity of ten additional institutes because of their work in the social and behavioral science. The institutes in Table 1 are listed in order from highest to lowest percentage of abstracts likely to be screened in as containing social/behavioral science and primary data collection.

Table 1: NIH funding centers reviewed with CRISP

National Institute on Child Health and Human Development (NICHD)
National Institute on Aging (NIA)
National Institute on Mental Health (NIMH)
National Institute of Nursing Research
Agency for Healthcare Research and Quality
National Institute on Alcohol Abuse and Alcoholism
National Institute on Drug Abuse
Clinical Center
National Institute on Deafness and other communication disorders
Fogarty International Center
National Cancer Institute
National Heart, Lung, and Blood Institute (NHLBI)
National Institute of Diabetes and Digestive and Kidney Diseases

For the 2001-2002, all NIH institutes were screened according to the criteria described above. Next, we screened NIA, NICHD, and NIMH for all available years from the NIH database (1979-2005). Currently, we are in the process of screening all remaining institutes listed in Table 1 (1979-2005). In total, 141,918 records have been screened using the decision criteria outlined above.

Preliminary Results

Of the 141,918 NIH-funded studies screened to date, 6,052 meet our decision criteria of describing plans for collecting new quantitative data that were social/behavioral. An additional 3,673 studies were flagged for further review.

SBC and DBS Screening

Next, we determine what proportion of research data collected as part of a primary investigation never enters the public domain. Using a subset of LEADS, we selected 500 demographic studies funded by DBS (1984-2005) and it's predecessor (1979-1984) for further review. Of these, more than 189 proposed an original quantitative data collection.

We reviewed the data holdings of ICPSR, Sociometrics, Roper Institute, Murray Research Data Center and the NICHD-Population Centers looking for archived datasets matching the 189 studies on our list. The vast majority of DBS-funded data collections

were not publicly available through these archives/institutes. ICPSR holds 17 of studies, Sociometrics disseminates 10 additional studies, the Murray Research Data Center distributes an additional three, and Princeton University archives an additional 2. These results are preliminary, but reveal the large gap between that which is collected and that which is shared widely through a public archive.

Analysis of the Research Data Life Course

Building upon the preliminary analysis, the proposed poster presentation will present findings from the systematic analysis of the full LEADS database extraction. The findings offered in this presentation address research questions that are central to a broader understanding of the *total life course* of a funded research project.

We are specifically interested in determining the proportion of new research supported by NIH in recent years that required the collection of primary data in order to fulfill their project goals. This analysis provides valuable information on the availability of primary data pertaining to specific research topics. This also provides a more accurate base to calculate the likelihood that primary data collections will be made available for re-analysis in the form of secondary data. This is an important issue as we currently lack accurate measures of the potential universe of research data could be made available for re-analysis. While we can measure both the incidence and the prevalence of secondary data available in the public domain, the actual denominator of all primary data collected during a specific time period remains unknown. By refining our measure of the denominator of existing primary data (even if to a restricted universe such as NIA funded studies) we obtain a more accurate understanding of the life course of research data and its outcome after the funded research period has concluded. We will also be able more accurately to estimate the risk that certain kinds of data are more likely to become dormant at the end of a funding cycle as opposed to those that more easily bridge the transition of becoming a secondary data set.

These types of research questions are timely, as NIH has established clear guidelines regarding the requirement to make funded research data available to the research community once the *primary life course* of a research project has concluded. There is now a clear expectation on the part of many funding agencies that primary data collection will transition into a *secondary life course* and that data will be made available to the research community for re-analysis.

Despite this expectation, there is still reluctance on the part of many research projects to make this transition, but the factors associated with this reluctance have not been examined analytically. To address this issue, the poster provides findings from our analysis on the differences between primary research data made available for re-analysis from funded research as opposed to the proportion still in a dormant state. This analysis is informative, specifically because it helps to identify factors that predict the rapid transition of certain from the primary to secondary life course as opposed to data that does not. This analysis is concerned with estimating risk profiles for data types or study designs that result in a reduced likelihood for entering the public domain.

This represents an important research area where there is considerable antidotal evidence but little systematic research. We logically assume that specific types of research are less likely to make transition from primary analysis to re-analysis because of issues of study design. Concerns over confidentiality and respondent protection can result in a reluctance to release data for secondary use. Similarly, data with small or unusual samples may be seen as having less application outside of the original study for which they were collected. Often these issues create genuine challenges that conflict with the recognized obligation to share data with interested colleagues. Presently however no structural analysis has been performed on the proportion of studies that face these kinds of challenges and how this influences the risk of non-release. In performing this analysis, we are able to describe factors that negatively affect the transition from the primary life course to the secondary life course of research data.

Based upon this analysis we identify what incentives lead to data sharing and what impediments interfere with this process. In identifying the barriers to data sharing, we also present specific policy suggestions to address these concerns. Clearly, not all data can be distributed in the same manner. Often primary data do face special concerns, particularly in the realm of confidentiality and disclosure risk. There are however, specific solutions to almost all data concerns that allows for re-analysis under distribution rules that safely address the risks associated with the use of primary data as a secondary data resource. Similarly, while some sample designs may not be immediately useful for re-analysis there is still a need to preserve such data for the historic record and the chance it will someday have additional research value. Resolving these issues requires a better understanding of the why some data enters the public domain and why some data remains dormant at the end of the initial research project. The findings from this poster presentation will significantly add to our understanding of these important issues.

Conclusion

This poster presentation will exhibit preliminary results for the pilot work on the lifecourse of funded research currently ongoing at ICPSR. It will present information on the paths that primary data collected as part of funded research can follow. Specifically we will present findings on the factors that influence likelihood that data will be released for secondary analysis and on the risks factors that increase the chances a data set will remain dormant after funding concludes. The presentation will present information on specific solutions to common barriers to research projects making the transition from primary to secondary data. Issues of confidentiality, generalizability and costs associated with the distribution of research data represent genuine challenges, but these challenges can be resolved. The benefits of secondary data analysis are undeniable, with a better understanding of what incentives encourage the release of research data to the public domain we can better assist researchers in developing appropriate strategies to add their data to the growing collection of information in the *secondary life course* of research.