Acknowledgements

• Supported by the Office of Research on Research Integrity (National Library of Medicine)

• National Science Foundation

• And also, NICHD, Library of Congress
Research Questions

Does data sharing lead to greater research productivity?

What types of factors (PIs, Grants, Institutions) are associated with data sharing & research productivity?

Net of structural characteristics of grants awards, does data sharing lead to greater research productivity?

Is there one metric of research productivity that shows the value of data sharing?
NSF Data Sharing Policy

National Science Foundation Important Notice 106 (April 17, 1989) states: "[NSF] expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections, and other supporting materials created or gathered in the course of the research. It also encourages awardees to share software and inventions or otherwise act to make such items or products derived from them widely useful and usable."
NIH Data Sharing Policy

The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. Starting with the October 1, 2003 receipt date, investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible.
Research Productivity & Data

• Traditionally conceptualized as one’s publications and citation of that original work

• Data sharing as a measure of research productivity
  ▪ Data life cycle - Moving from the original purpose of the data to uses beyond that which the data were intended to be used for.

• Primary and secondary publications
# Prior Results - LEADS Data

<table>
<thead>
<tr>
<th>NSF &amp; NIH Funded Data Collections: Where are they today? N=1,544</th>
</tr>
</thead>
<tbody>
<tr>
<td># Records Reviewed</td>
</tr>
<tr>
<td>--------------------</td>
</tr>
<tr>
<td>Recent NSF (1976+)</td>
</tr>
<tr>
<td>Historic NSF (Pre-1976)</td>
</tr>
<tr>
<td>NIH (1972+)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Are Archived</th>
<th>Has Copy of Data</th>
<th>Data Are Lost</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.2%</td>
<td>58.7%</td>
<td>25.7%</td>
</tr>
</tbody>
</table>
Sources of Information

**NSF**

**NIH**
Sample Inclusion Criteria-LEADS

- Social science and/or behavioral science
- Original or primary data collection proposed, including assembling a database from existing (archival) sources
PI Survey

- Conducted in 2009, web survey
- Awards beginning 1985-1991
- 1,217 responses (24.9% response rate)
- 86.6% PIs report having collected social science data
Publication Measures

- Self-reported
- Total publications (.92 correlation with staff measure)
- Secondary Publications
  - Number of publications without any research team member
- Primary Publications
  - Number of publications that include PI as one of the authors
Data Sharing Status

• Formal data sharing through a data archive or repository (11.5%)

• Informal data sharing (website, upon request) (44.6%)

• Not shared (43.9%)
PI Characteristics

- Gender (48.1% female)
- Race (86.8% white)
- Age (mean 43.4)
- Rank (Senior 54.3%; Junior 25.7%; Non-Fac 20%)
- Discipline (Psy/hlth 62.5%, core soc science 25.5%, other 12%)
- Number of Federal Grants 6.2
Institutional Characteristics

• Carnegie classification
  ▪ research university (78.7%)
  ▪ non research (6.5%)
  ▪ PRO (12.4%)
  ▪ Other (2.5%)

• U.S. Region
  ▪ NE (36%)
  ▪ Midwest (23.7%)
  ▪ South (21.6%)
  ▪ West (18.7%)
Grant Characteristics

- NSF (27.3%) versus NIH
- 3.1 years duration
Bivariate Results

Who shares data (in an archive)

Women
Senior faculty
Core social science (esp. more than psy/health)
Northeast (esp. more than south)
Research Universities and PROs
NSF (22.4% compared to 7.4% of NIH)
Median Publication Counts

- Formal
- Informal
- Not Shared
Analytic Models

- Negative binomial regression models of publication counts
  - Offset by time between initial award and survey year
### Data Sharing Status

<table>
<thead>
<tr>
<th>Status</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archived</td>
<td>1.094 (0.123)</td>
<td>0.884 (0.128)</td>
</tr>
<tr>
<td>Shared Informally-Not Archived</td>
<td>1.020 (0.080)</td>
<td>0.837 (0.079)</td>
</tr>
<tr>
<td>Not Shared</td>
<td>ref</td>
<td>ref</td>
</tr>
</tbody>
</table>

* p<.1; ** p<.05; *** p<.01

### PI Characteristics

- **Age at award**
  - Model 1: 0.025 (0.004) ***

- **Discipline - Health and Psychology**
  - Model 1: -0.254 (0.102) **

- **Discipline - Other (v.s Core Soc Sci)**
  - Model 1: -0.190 (0.130)

### Institutional Characteristics

- **Carnegie-Non Res University**
  - Model 1: -0.685 (0.157) ***

- **Carnegie-Other**
  - Model 1: 1.169 (0.246) ***

- **Carnegie-PRO (vs. Res Univ)**
  - Model 1: 0.230 (0.113)

### Grant Award Characteristics

- **NIH (vs. NSF)**
  - Model 1: 0.075 (0.093)

- **Duration of Award, Years**
  - Model 1: 0.163 (0.027) ***

### Intercepts and Dispersion

- **Intercept**
  - Model 1: 1.646 (0.058)
  - Model 2: 0.199 (0.222)

- **Dispersion**
  - Model 1: 1.186
  - Model 2: 1.052

Log-likelihood estimates (standard errors in parentheses)
Results - Total Publication Count

- 3 times the # primary publications when data are archived
- 2 3/4 times the publications when data are shared informally
- Not explained by PI, Institutional, or Grant Award Characteristics.

- Older PIs (at time of award) greater research productivity (RP).
- Health/Psychology -> less RP (vs. core social science)
- Non-research university produced data -> less PR (vs. research univ).
- Longer awards -> More RP
Results - Secondary Pub. Count

- 12 times the primary publications when data are archived
- 10 3/4 times the publications when data are shared informally.
- Not explained by PI, Institutional, or Grant Award Characteristics (reduced by 1/2).
- Older PIs (at time of award) greater research productivity (RP).
- Health/Psychology & Others -> less RP (vs. core social science)
- PRO -> more RP (vs. research univ).
Results - Primary Publications

- 2 times the primary publications when data are archived
- 2 times the publications when data are shared informally.
- Not explained by PI, Institutional, or Grant Award Characteristics.
- Older PIs (at time of award) greater research productivity (RP).
- Health/Psychology -> more RP (vs. core social science)
- Non-Res & PRO -> less RP (vs. research univ)
- NIH -> more RP than NSF
- Longer awards more RP
Conclusions

- Data sharing is relatively infrequent across the social and behavioral sciences.
- When data are shared, formally or informally, research productivity is higher.
- This is true even after PI, Award, and Institutional characteristics are accounted for.
- When data are archived, the return on the investment is the highest.
- Secondary publications is the most sensitive to data sharing. But, all outcome measures show advantages to data sharing.
Limitations

- Causality - unclear if data sharing actually leads to more primary publications…

- Size of award is not controlled for
Future work

- Evaluation of ARRA funding at NSF and NIH
  - Natural experiment - ARRA award period compared to prior years
  - Publication Outcomes
  - Other Non-publication Outcomes