Working with Administrative Data

Laura Feeney
Lecture Overview

- What are administrative data?
- Randomized evaluations that use administrative data
- Why are administrative data useful?
- Acquiring administrative data

**Information gathering**
- Locate data source
- Understand data universe and content
- Understand the legal environment

**Implementation**
- Develop data flow
- Negotiate data use agreement
- Get IRB approval

- Challenges when working with administrative data
What are administrative data?

Information collected, used, and stored primarily for administrative (i.e. operational), rather than research, purposes

- Medical records
- Educational records
- Arrest records
- Banking records
- Personnel records
Evaluations utilizing administrative data

Social Incentives and Tax Compliance in Bangladesh
• Business tax records

Market Structuring of Sludge Management for the Benefit of Vulnerable Households in Dakar (Senegal)
• Sludge treatment center records

Oregon Health Insurance Experiment (United States)
• Medicaid enrollment and claims data, credit reports, criminal charges
Why are administrative data useful?

The outcomes and metrics required for a study may already be tracked by a government or organization:

- Available retrospectively
- Enable long-term follow-up
- Reduce logistical burden
- Include near census of relevant population
- Often cheaper than surveys
Why are administrative data useful?

May minimize bias and error

- **Recall bias**
  - Data recorded at the time of occurrence—no memory needed (e.g., banking records)

- **Social desirability bias**
  - Non-self-reported data (e.g., arrest records)

- **Differential attrition and non-response bias**
  - Near census of relevant population
Lecture Overview

- What are administrative data?
- Randomized evaluations that use administrative data
- Why are administrative data useful?
- Acquiring administrative data

- Challenges when working with administrative data

**Information gathering**
- Locate data source
- Understand data universe and content
- Understand the legal environment

**Implementation**
- Develop data flow
- Negotiate data use agreement
- Get IRB approval
Acquiring Administrative Data

Information gathering
- Locate data source
- Understand data universe and content
- Understand the legal environment

Implementation
- Develop data flow
- Negotiate data use agreement
- Get IRB approval
That’s a lot of steps...

More guidance is available: povertyactionlab.org/admin data
Acquiring Administrative Data

Information gathering

Locate data source
Understand data universe and content
Understand the legal environment

Implementation

Develop data flow
Negotiate data use agreement
Get IRB approval
Locate a Data Source

• J-PAL’s Catalog of US data sets
• Health
  – Vital statistics office
  – Health facilities (e.g., hospitals, clinics)
• Finance
  – Banks, credit unions
  – Credit rating agencies
• Education
  – Schools
  – Department of education
• Statistics department
Acquiring Administrative Data

Information gathering

- Locate data source
- Understand data universe and content
- Understand the legal environment

Implementation

- Develop data flow
- Negotiate data use agreement
- Get IRB approval
Understand the data universe

Data Universe

- Which individuals are captured in the data
- Reasons for inclusion or exclusion
- Example: Medicaid claims data
Understand the data content

Data Content
• Variables of interest
  – Data dictionaries
Understand the data content

More likely...

- Request variable list from the provider
- Iterate with a data expert on which variables would be useful or necessary
Understand the data content

1. Make a list of variables that seem interesting
2. Request clarity on variable definitions
3. Request suggestions of other related variables
4. Ensure there are adequate identifiers to link your evaluation data with the administrative data
Acquiring Administrative Data

**Information gathering**
- Locate data source
- Understand data universe and content
- Understand the legal environment

**Implementation**
- Develop data flow
- Negotiate data use agreement
- Get IRB approval
Understand the legal environment

More Identified + More Sensitive = More Hurdles

- Legal restrictions
- Data use agreements
- IRB approval
- Data security
- Trade secrets
- Trust
Understand the legal environment

- National, local, or institutional regulations
- Do your own due diligence
- Though federal laws may permit the sharing of data for research, they do not necessarily require the sharing of data
Personally Identifiable Information (PII)

Definition of PII is broad and is not limited to:

- Name
- Identification numbers
- Address
- Photos or biometric characteristics
Understand the legal environment

Many combinations of variables can identify individuals
• E.g., anonymous Netflix data combined with IMDb data can identify individuals (Arvind and Shmatikov 2008)
Understand the legal environment

Regulations that limit access to identified data

- **US: Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule**
  - Obligations depend on level of identification and your contract with the data provider
    - Data security requirements
    - Fines for data leaks
    - Individual Authorizations or waivers (similar to informed consent)

- **US: Family Educational Rights and Privacy Act (FERPA)**
  - Protects students’ educational records
  - Requires student consent for release of records

- **Other regulations depending on country context**
Levels of identified data

Is it always necessary to know exactly who’s who?

There are methods for matching study data with administrative data that don’t require the release of direct identifiers.
Acquiring Administrative Data

**Information gathering**
- Locate data source
- Understand data universe and content
- Understand the legal environment

**Implementation**
- Develop data flow
- Negotiate data use agreement
- Get IRB approval
Data Flow

Matching administrative data with program/evaluation data

1. Which identifiers will be used for matching
2. What software and algorithm will be used
3. Which identifiers the research team will have access to
4. Which team will perform the match
Identifiers for matching

• **Understand what identifiers the data agency collects**
  - Collect those same identifiers from your study sample at baseline

• **Use numeric identifiers instead of string variables**
  - Identification number and DOB instead of names and addresses

• **Participants may not be willing to provide sensitive identifiers**
  - E.g., identification numbers. Emphasize privacy & confidentiality during study enrollment
Matching process

Exact/Deterministic

- Minor discrepancies are not well accounted for
  - E.g., typos in name, reversed day and month in DOB
- Some records are not identified as matches even though they may be (false negatives)

<table>
<thead>
<tr>
<th>FINDER FILE</th>
<th>ADMINISTRATIVE DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECORD</td>
<td>NAME</td>
</tr>
<tr>
<td>1</td>
<td>Jane Doe</td>
</tr>
<tr>
<td>2</td>
<td>Jonathan Smith</td>
</tr>
</tbody>
</table>
Matching process

Fuzzy/Probabilistic

- Accounts for the likelihood that identifiers may not align exactly to those in a data system
  - E.g., SSN and last name match, DOB is off by a month—count as a match

<table>
<thead>
<tr>
<th>FINDER FILE</th>
<th>ADMINISTRATIVE DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECORD</td>
<td>NAME</td>
</tr>
<tr>
<td>1</td>
<td>Jane Doe</td>
</tr>
<tr>
<td>2</td>
<td>Jonathan Smith</td>
</tr>
</tbody>
</table>
Separate identifiers from outcomes

Identified Finder File

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Name</th>
<th>DOB</th>
<th>SSN</th>
</tr>
</thead>
</table>

Administrative Data File

<table>
<thead>
<tr>
<th>Name</th>
<th>DOB</th>
<th>SSN</th>
<th>Outcome 1</th>
<th>Outcome 2</th>
</tr>
</thead>
</table>

De-identified Analysis File

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Treatment Status</th>
<th>Outcome 1</th>
<th>Outcome 2</th>
</tr>
</thead>
</table>

Data Flow: Option 0

**Data Agency**

Sends identified data and outcomes

**Identified Admin Data**

**Researcher**

Conducts match, strips identifiers off of data set to create de-identified analysis file

Stores finder file & de-identified analysis file separately
Data Flow: Option 1

Data agency allows researcher to conduct matching on-site

- Bring encrypted finder file
- Conduct match
- Leave with de-identified analysis file
Data Flow: Option 2

Data agency provides secure computer

- Researcher conducts match and analysis on this monitored device
Data Flow: Option 3

Data Agency
- Sends list of variable names
- Conducts match
- Runs analysis code
- Sends analytic results

Researcher
- Writes and tests analysis code
- *Never* has access to outcomes
Data Flow: Option 4

Researcher agrees **never** to match
finder file & de-
identified analysis file

Conducts match
and strips identifiers
off of data set (but
keeps study ID)
Data Flow: Option 5

**Research Partner**
Assigns study IDs, **never** has access to administrative data

**Data Agency**
Conducts match and strips identifiers off of data set

**Researcher**
**Never** has access to identifiers
Acquiring Administrative Data

Information gathering
- Locate data source
- Understand data universe and content
- Understand the legal environment

Implementation
- Develop data flow
- Negotiate data use agreement
- Get IRB approval
Data Use Agreements

Documents the terms under which a data provider shares data with a researcher’s home institution for use by the researcher.

**Elements**
- Project description
- Users and analysts
- Data security procedures
- Data to be shared
- Timeframe
- Data destruction
- Publication review
Data Use Agreements

DUAs should be signed by an official institutional representative rather than an individual PI or staff member

- Ensures proper legal review
- No over-promising
- Prevents unapproved institutional liability
- Ensures no individual liability (to the extent possible)
Data Use Agreements

Templates

• Best to start from one of the contracting parties’ preferred templates
  – Most universities, as well as some implementing partners and data providers, have their own templates
  – [MIT sample DUA](#)
When two legal teams are involved, it can quickly become adversarial.

Act as an intermediary to help smooth the negotiation

- Understand the legal context – “needs” vs. “wants”
- Establish a good relationship with the data provider
- Emphasize the importance of the research and relevance to the data provider’s mission (if applicable)
Acquiring Administrative Data

Information gathering

Locate data source
Understand data universe and content
Understand the legal environment

Implementation

Develop data flow
Negotiate data use agreement
Get IRB approval
IRB Approval

Informed Consent

- Consent to your use of their administrative data
- IRBs and/or data providers may require that individuals consent to each specific data set that may be used
- Waiver of informed consent

HIPAA - Individual Authorizations

- Authorization to use health data for research
- Usually obtained in conjunction with informed consent
Get an early start

Plan to get administrative data before the study begins

• Allow the data to inform the analysis plan

Plan appropriately for data lags

• Some data are available on a one-year or more lag
• Allow time for the data provider to extract data and transfer
Lecture Overview

- What are administrative data?
- Randomized evaluations that use administrative data
- Why are administrative data useful?
- Acquiring administrative data

Information gathering

- Locate data source
- Understand data universe and content
- Understand the legal environment

Implementation

- Develop data flow
- Negotiate data use agreement
- Get IRB approval

Challenges when working with administrative data
Mitigating challenges

Records are in an unusable format

- Hand-written records
- PDF file
Mitigating challenges

To address records in an unusable format

Digitize

- Business tax records in Bangladesh
- Sludge treatment center records in Senegal
Mitigating challenges

How can we ensure that the data are accurate?

– Unlike with survey data, the researcher does not have a say in the data collection and processing phase.
Mitigating challenges

To address possibly inaccurate data

- Cross-reference with other sources to ensure accuracy
- Identify the data agency’s quality control protocol
- Choose indicators that are unlikely to be incorrectly reported
  - Select variables that are straightforward and less susceptible to human error
  - Request raw variables
- Communicate with program or implementing partner responsible for collecting data
  - Ask how and why data are collected
Mitigating challenges

**Reporting Bias**

- From an individual
  - E.g., under-reporting income to qualify for a social welfare program

- From an administrative organization
  - E.g., schools over-report attendance to meet requirements
Mitigating challenges

To address reporting bias

• **Identify the context in which the data were collected**
  – Were there incentives to misreport information?

• **Choose variables that are not susceptible to bias**
  – E.g., hospital visit v. value of insurance claim
Mitigating challenges

**Differential Coverage**

- Differential ability to **link** individuals to administrative records
- Treatment and control are differentially likely to **appear** in administrative records

![Bar chart showing actual and measured visits for treatment and control groups.](chart.png)
Mitigating challenges

To address differential coverage bias

• **Collect identifiers for linking during the baseline survey**
  – To ensure that you are equally likely to be able to **link** treatment and control individuals to their records

• **Identify the data universe**
  – Which individuals are included in the data and which are excluded, and why?
  – To ensure the intervention does not affect the likelihood of **appearing** in a data set
Mitigating challenges

Differential Reporting

- Likelihood of reporting outcome is correlated with treatment
  - True value of the outcome may not differ between treatment and control, but due to the intervention, treatment group is more likely to report a certain outcome or appear in administrative records.
Mitigating challenges

To address differential reporting

• **Identify how the intervention may affect the reporting of outcomes**
  
  – Identify the context in which the data were collected
  
  – Determine direction in which estimates are likely to be biased
  
  – E.g., do number of doctor’s visits reflect severity of sickness or stronger connection to the health system?
Lecture Overview

• What are administrative data?
• Randomized evaluations that use administrative data
• Why are administrative data useful?
• Acquiring administrative data

Information gathering

- Locate data source
- Understand data universe and content
- Understand the legal environment

Implementation

- Develop data flow
- Negotiate data use agreement
- Get IRB approval

• Challenges when working with administrative data
Conclusion

**Administrative data**...

- are a valuable complement or substitute for survey data
- are available longitudinally and can be accessed retrospectively
- reduce bias found in survey data but also have bias of their own
- require IRB review and DUAs that are shaped by legal restrictions to protect PII
- require forethought and planning