USING ADMINISTRATIVE DATA FOR RANDOMIZED EVALUATIONS

Laura Feeney, Jason Bauman, Julia Chabrier, Geetika Mehra, and Michelle Woodford

povertyactionlab.org/admindata

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
NORTH AMERICA

ADMINISTRATIVE DATA

Administrative data are information collected, used, and stored primarily for administrative (i.e., operational) purposes. These data can be an excellent source of information for use in research and impact evaluation.

A **randomized evaluation** is a type of impact evaluation that uses random assignment to allocate resources, run programs, or apply policies as part of the study design. In particular, randomized evaluations measure program effectiveness by comparing outcomes between those randomly assigned to a "treatment group," who received the program, and those randomly assigned to a "control group," who did not receive the program.

ADVANTAGES OF ADMINISTRATIVE DATA

Project Management

- Longitudinal availability
- Cheaper and/or easier than conducting surveys
- Large sample size

Measurement & Analysis

Reduces threats of…

- Recall bias
- Social desirability bias
- Non-response bias
- Differential attrition

POTENTIAL SOURCES OF BIAS

Misreporting

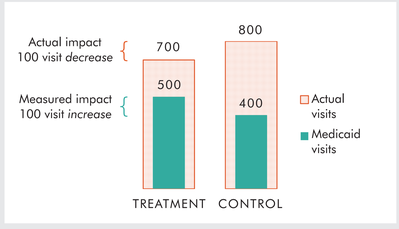
- Individual or organizational incentives to under- or over-report

Differential Coverage

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

MEASUREMENT BIAS UNDER DIFFERENTIAL COVERAGE OF DATA

Illustrative Example: Measuring hospital visits through Medicaid claims



Researchers are studying the effects of a home health-care program on hospital visits. The home health program also helps participants enroll in social services such as Medicaid. Due to the enrollment assistance, individuals in the treatment group are more likely to appear in Medicaid records than individuals in the control group. Measuring program impact on hospital visits through Medicaid claims may lead to biased results.

BARRIERS TO ACCESS

Individual-level administrative data are a powerful resource for researchers, but regulations designed to protect individual or institutional privacy restrict access to identified data sets. *The more identified or sensitive the data set, the harder it is for researchers to gain access.*

Whether a data set is identified depends on the amount of **Personally Identifiable Information (PII)** included in the data set. PII is any piece or combination of information that can be used to identify a particular individual with a reasonable amount of certainty, including, but not limited to an **individual's name, identification numbers, address, photos**, or **biometric characteristics.**[1]

TYPES OF DATA SETS

*Access to an identifiable data set generally requires that the researcher navigate **IRB approval, data use agreements**, and other **legal restrictions** to gain access.*

In sectors from health to education, definitions of PII are purposefully broad to prevent de-identified data from becoming identified. Despite efforts to de-identify data sets, in many cases, de-identified data combined with additional information, can lead to an identified data set.

For example, researchers re-identified individuals from a de-identified Netflix data set. They combined the Netflix data, containing movie ratings of individual subscribers, with individuals' identified, publicly-available movie ratings from the Internet Movie Database (IMDb) to identify individuals (Narayanan and Shmatikov 2008).

1 This definition is consistent with several regulations in place including:

- The Federal Policy for the Protection of Human Subjects or the 'Common Rule'
- The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule
- The Family Educational Rights and Privacy Act (FERPA)

DATA FLOW

Using administrative data to measure the impact of a program or policy typically requires matching individuals in a study or program to their administrative records. Given the strict legal environment surrounding access to identified data, a data flow strategy that limits the researcher's direct contact with identified data can simplify the data access process and reduce additional restrictions imposed by data providers. The following five data flow strategies may be used to match study data with individual-level administrative data.

THREE TYPES OF FILES ARE CENTRAL TO THE DATA FLOW PROCESS

Finder file

Study ID	Name	DOB	SSN

Administrative data file

Name	DOB	SSN	Outcome 1	Outcome 2	Outcome 3

De-identified analysis file

Study ID	Study Group	Outcome 1	Outcome 2	Outcome 3

DATA FLOW STRATEGIES

Option One

Researchers conduct matching on-site at the data agency. Researchers bring a finder file, conduct the match, and leave with a de-identified analysis file.

In the Oregon Health Insurance Experiment, researchers used this strategy to match study data to hospital discharge data (Taubman et al. 2014).

Option Two

Researchers conduct matching and analysis with a secure computer provided by the data agency.

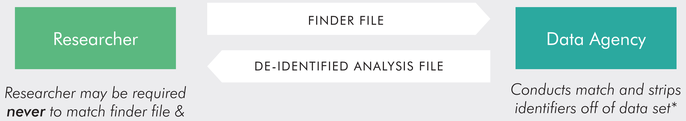
Researchers evaluating a program to reduce inappropriate prescribing of controlled substances used this strategy to match study data to Medicare Part D records including prescription drug fill records (Sacarny et al. 2016).

Option Three

Data agency sends researchers variable names included in the administrative data file of interest to the researcher. Researchers write and test analysis code with these variable names and then send the code to the agency. The data agency runs the code and sends the analytic results—but not the full data set—to researchers.

In the Oregon Health Insurance Experiment, researchers used this strategy to match study data with Social Security Administration data on annual earnings and receipt of Social Security Disability Insurance and Supplemental Security Income (Baicker et al. 2014).

Option Four



Researchers evaluating a nurse home visiting program are using this strategy with the help of the South Carolina Revenue and Fiscal Affairs office to match insurance claims and vital statistics data to individuals in the study (J-PAL The Impact of a Nurse Home Visiting Program 2016).

**Alternatively, the data agency assigns new study IDs, preventing the researcher from matching the finder file and the de-identified analysis file.*

Option Five

Researchers measuring the impact of outreach and application assistance on take-up of Supplemental Nutrition Assistance Program benefits are using this strategy to match application and enrollment data from public benefits programs to health care claims for individuals in the study (J-PAL SNAP Take-Up Evaluation 2016).

ACKNOWLEDGEMENTS

This work was made possible by support from the Alfred P. Sloan Foundation and the Laura and John Arnold Foundation. Kenya Heard provided excellent research assistance.

REFERENCES

Baicker, Katherine, Amy Finkelstein, Jae Song and Sarah Taubman. 2014. "The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment." *American Economic Review* 104(5): 322-28. Accessed October 18, 2016. doi: 10.1257/aer.104.5.322.

J-PAL. "Health Care Hotspotting in the United States." Accessed October 19, 2016. https://www.povertyactionlab.org/evaluation/health-care-hotspotting-united-states.

J-PAL. "SNAP Take-Up Evaluation." Accessed October 19, 2016. https://www.povertyactionlab.org/evaluation/snap-take-evaluation.

J-PAL. "The Impact of a Nurse Home Visiting Program on Maternal and Child Health Outcomes in the United States." Accessed October 19, 2016. https://www.povertyactionlab.org/evaluation/impact-nurse-home-visiting-program-maternal-and-child-health-outcomes-united-states.

Narayanan, Arvind and Vitaly Shmatikov. 2008. "Robust De-anonymization of Large Sparse Datasets." *IEEE*. doi: 10.1109/SP.2008.33

Sacarny, Adam, David Yokum, Amy Finkelstein and Shantanu Agrawal. 2016. "Medicare Letters To Curb Overprescribing Of Controlled Substances Had No Detectable Effect on Providers." *Health Affairs* 35(3): 471-9. Accessed October 18, 2016. doi: 10.1377/hlthaff.2015.1025.

Taubman, Sarah L., Heidi L. Allen, Bill J. Wright, Katherine Baicker and Amy N. Finkelstein. 2014. "Supplementary Materials for Medicaid Increases Emergency-Department Use: Evidence from Oregon's Health Insurance Experiment." *Science Express*. Accessed October 19, 2016. doi: 10.1126/science.1246183.