J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
NORTH AMERICA

USING ADMINISTRATIVE DATA FOR RANDOMIZED EVALUATIONS

Laura Feeney, Jason Bauman, Julia Chabrier,
Geeti Mehra, Michelle Woodford

povertyactionlab.org/na

Summary: Administrative data are information collected, used, and stored primarily for administrative (i.e., operational), rather than research, purposes. These data can be an excellent source of information for use in research and impact evaluation. This guide provides practical guidance on how to obtain and use nonpublic administrative data for a randomized evaluation. While many of the concepts in this guide are relevant across countries and contexts, certain sections are only applicable to research conducted in the United States.

Please send comments, questions, or feedback to admindata@povertyactionlab.org.

Disclaimer: This document is intended for informational purposes only. Any information related to the law contained herein is intended to convey a general understanding and not to provide specific legal advice. Use of this information does not create an attorney-client relationship between you and MIT. Any information provided in this document should not be used as a substitute for competent legal advice from a licensed professional attorney applied to your circumstances.

INTRODUCTION

This guide provides practical guidance on how to obtain and use nonpublic administrative data for a randomized evaluation. Administrative data are information collected, used, and stored primarily for administrative (i.e., operational), rather than research, purposes. Government departments and other organizations collect administrative data for the purposes of registration, transaction, and record keeping, usually during the delivery of a service.[1] Examples of administrative data include credit card transactions, sales records, electronic medical records, insurance claims, educational records, arrest records, and mortality records. This guide focuses on nonpublic (i.e., proprietary or confidential) administrative data that may be used in an individual-level randomized evaluation.

Context of this guide. Many of the concepts in this guide are applicable across countries and contexts. However, sections pertaining to compliance (particularly HIPAA and specific ethics requirements) are directly applicable only in the United States. Other jurisdictions with similar regulatory contexts may have similar legislation (e.g., the European Union has legislation on data protection), resulting in the general applicability of concepts across countries.

This guide focuses on the following topics:

- Standardized processes for accessing administrative data

- The ethical and legal framework surrounding the use of administrative data for randomized evaluations

- Common challenges in using administrative data

---

[1] For further detail, see the Administrative Data Liaison Service's website.

TABLE OF CONTENTS

WHY USE ADMINISTRATIVE DATA?

There are a number of advantages to using administrative data for research:[2]

Cost and ease. Using administrative data may be less expensive and logistically easier than collecting new data. Unlike primary data collection, administrative data collection does not require development and validation of a survey instrument, contracting a survey firm or enumerators, or tracking subjects for follow-up[3].

Reduced participant burden. Subjects are not required to provide information to researchers that has already been shared in other contexts.

Near-universal coverage. Many existing administrative databases provide a near-census of the individuals relevant to a given study. Often, both treatment and control subjects are present equally in these data, as are subjects who may be less likely to respond to follow up surveys for reasons related to their treatment status.

Accuracy. Administrative data may be more accurate than surveys in measuring characteristics that are complex or difficult for subjects to remember (e.g., income, consumption).[4]

> *"An example of the value of administrative data over survey data can be seen in the Oregon Health Insurance Experiment's study of the impact of covering uninsured low-income adults with Medicaid on emergency room use. This randomized evaluation found no statistically significant impact on emergency room use when measured in survey data, but a statistically significant 40 percent increase in emergency room use in administrative data (Taubman, Allen, Wright, Baicker, & Finkelstein 2014). Part of this difference was due to greater accuracy in the administrative data than the survey reports; limiting to the same time periods and the same set of individuals, estimated effects were larger in the administrative data and more precise."* (Amy Finkelstein and Sarah Taubman, "Using Randomized Evaluations to Improve the Efficiency of US Healthcare Delivery." [2015].)[2]

Minimized bias. Using administrative data that are captured passively, rather than actively reported by individuals or program staff, minimizes the risk of social desirability or enumerator bias. See *Reporting Bias* for more details.

Long-term availability. Administrative data may be collected systematically and regularly over time, allowing researchers to observe outcomes for study participants across long spans of time. These long-term outcomes are often the most interesting from both a research and a policy perspective, and may allow researchers to identify impacts that are not present in the short-term. An example of this can be found in Ludwig et al. (2013):

> The Moving to Opportunity (MTO) project tested the impact of offering housing vouchers to families living in high-poverty neighborhoods. Using administrative data from tax returns, researchers found that children

---

[2] For more detail on those listed here, see Finkelstein and Taubman (2015).

[3] This is not to say that administrative data are necessarily cheap; It is not uncommon to see data priced around $10,000 per file.

[4] In Meyer and Mittag (2015), the researchers compared administrative data from SNAP (food stamps), Temporary Assistance for Needy Families (TANF), General Assistance, and subsidized housing from New York State to the Current Population Survey (CPS), which is conducted by the US Census Bureau. They find that income is significantly underreported in the CPS when compared with administrative records, and argue that the choice of data source sharply alters measurements of well-being and the effects of transfer programs.

who were under age 13 when their families moved to lower-poverty neighborhoods had increased rates of college attendance, higher incomes, and lived in lower-poverty neighborhoods later in life. The higher adult incomes yield significantly higher tax payments, which could result in government savings in the long term. These long-term effects are useful in assessing the impact that housing vouchers may have in lifting families out of poverty, but they were not visible in short-term data.

Cost data. Some administrative data sources are the authoritative data source of cost data, enabling research on public finances or cost-effectiveness analysis. For example, Medicare claims record the exact cost to the public of the health utilization covered by Medicare.

Administrative data also have limitations, challenges and risks, as described in more detail in "*Potential Bias When Using Administrative Data*." Both the Oregon and MTO studies took advantage of both administrative data and survey data, allowing the researchers to study a broad range of outcomes and individuals[5].

POTENTIAL BIAS WHEN USING ADMINISTRATIVE DATA

Administrative data may be more accurate and less susceptible to certain biases than survey data (Finkelstein and Taubman 2015; Meyer and Mittag 2015), but administrative data are not immune to issues of bias and inaccuracy. Bias is of particular concern when using administrative data if being assigned to the treatment group affects the likelihood that an individual appears in, or could be linked to, administrative data.

Differential Coverage: Observability in Administrative Data

Differential coverage, i.e., the difference in the type or proportion of missing data between treatment and control subjects, can be a problem even when using administrative data. This problem presents itself when members of the treatment and control group are differentially likely to appear in administrative records, or when researchers are differentially likely to be able to link individuals to their administrative records. When this happens, researchers cannot be sure that the treatment and control groups are still statistically equivalent, and impact estimates based on these data may be biased. In these situations, special care must be taken in considering how to treat missing or unmatched outcomes. Attempting to correct for these problems can be problematic, as the types of individuals missing in each group may differ in unobservable ways.

Differential coverage may arise in the following ways when using administrative data:

- **Identifiers obtained after enrollment or baseline are used to link individuals to administrative data.** For example, members of the treatment group may be more willing to share their Social Security number after several interactions with a member of the study team. Control group members, however, have no such additional trust. This results in researchers having a higher likelihood of finding administrative records for treatment group members than control group members.

---

[5] While a study may use both surveys and multiple sources of administrative data, each outcome should be measured using the same source of data and the same method for both the treatment and control group.

- **Using data generated by the same program to which participants are randomly assigned.** Using program process or monitoring data to evaluate the effects of random assignment to that same program (or encouragement to participate in that same program) is usually not appropriate because treatment group members may be more likely to appear in program implementation data than control group members due to their random assignment status. Further, these data are likely not collected in the same way or likely to be as accurate for both the treatment and control group, leading to bias. For example, consider a hypothetical evaluation that assigns some individuals to receive financial counseling at a specific credit union branch. Measuring financial health using data from that same branch would likely capture a much larger proportion of the treatment group--who are now engaged with the branch--than the control group. In this case, researchers may gain a more complete picture of financial health by using data from a national credit-reporting agency.

- **Members of the treatment group are more (or less) likely than the control group to appear in a certain source of administrative data due to their treatment assignment.** For example:

  > Consider the following hypothetical example: Researchers plan to measure the impact of an intensive home health-care program on frequency of hospital admissions using Medicaid claims data. As a part of their normal services, the home health program also helps participants enroll in social services they may be eligible for, including Medicaid. Because of the assistance offered by the home health program, individuals in the treatment group are more likely to appear in Medicaid records than individuals in the control group. As illustrated below, even if the intervention actually reduces hospitalization (i.e. from 800 to 700), researchers analyzing the Medicaid data may see the opposite effect (i.e. an increase from 400 to 500) due to the larger share of the treatment group's hospitalizations that appear in the data.

FIGURE 1: ILLUSTRATION OF BIASED RESULTS WHEN TREATMENT GROUP MEMBERS ARE MORE LIKELY TO APPEAR IN A DATA SET.
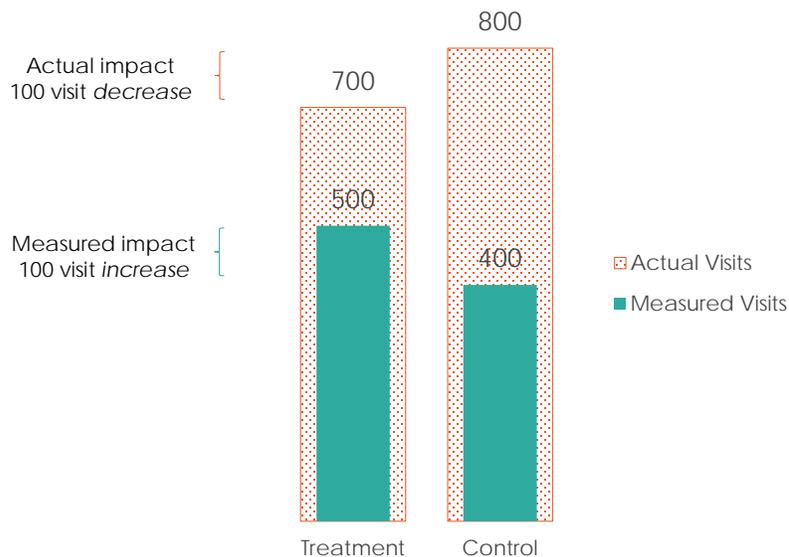
Reporting Bias

Many types of administrative data are collected passively, as sales, orders, or transactions occur. However, certain elements within an administrative data set may be actively recorded by a human, rather than passively captured as transactions occur. For example, mothers' self-reports are the basis for much of the pregnancy history and prenatal care records present on birth certificates. Studies validating birth certificate data have found that these self-reported data have lower validity in comparison with information on birth weight, Apgar score, obstetric age, and methods of delivery, which are recorded based on the observations and measurements of a professional (Lee et. al 2015).

Depending on the original purpose for collecting the data, the organization collecting the data may have differing incentives to verify the accuracy of the information. There may also be incentives for either (or both) the subject and the administrative organization to misreport data. For example, an individual may be incentivized to underreport income in an application for social welfare services. For health-care claims, expensive procedures are likely to be reported in order to receive payment, while some smaller procedures may not be reported if the expected probability of payment is low.

Understanding how and why administrative data are reported, collected and verified is critical to assessing its validity. Further, it is important to understand whether the experiment itself is expected to manipulate incentives relevant to data reporting and validity.

For example, high-stakes testing has been linked to a number of strategies to artificially boost standardized test scores. The administrative records of schools in which substantial score-boosting has occurred may be less accurate measures of student learning than an independent cognitive assessment conducted by a researcher.

HOW TO FIND ADMINISTRATIVE DATA

Implementing partners are an invaluable resource for identifying sources of administrative data relevant to an evaluation. The partner may already have access to certain types of data through their standard work, creating a natural starting point for data discussions. For example, public health nonprofits may have existing partnerships with state Medicaid agencies that can be leveraged to expand an existing data use agreement.

Outside of this channel, J-PAL North America is working to compile a [catalog of administrative data sets](#) that may be used in randomized evaluations, focusing on data sets not for public use.

Ideas and examples of administrative data in the US include:

- The Research Data Assistance Center ([ResDAC](#)), which provides information and assistance with applying for access to data from the Centers for Medicare and Medicaid Services

- Credit reporting agencies: [Equifax](#), [Experian](#), and [TransUnion](#)

- The American Economic Association hosts resources enumerating [sources](#) and [procedures](#) for accessing US federal administrative data

- Researchers have compiled an inventory of data sets used to study [education](#)

COST OF ADMINISTRATIVE DATA

Administrative data sets range in price and vary in pricing strategy. Data providers may charge per individual record, per range of records (e.g., <10,000 records, 10,000–50,000 records, >50,000 records), or by file-year.

Researchers may be able to take advantage of their partner organization's existing data relationships to acquire data at a lower price. For example, a nonprofit organization that offers credit counseling may already have free access or discounted access to credit report data, in order to perform their normal business activities. Researchers may be able to leverage these existing relationships to lower the price of data needed for an evaluation.

In other cases, researchers may need to make new requests for data solely for research purposes. Prices vary widely: voter registration records in Pennsylvania are $20 per county; credit report data for a sample of 200,000–300,000 individuals for one year are over $5,000; one year of Medicare outpatient records for the full population of enrollees is over $10,000 per file. While the raw cost appears very high, the marginal cost per individual may be far less than costs for primary data collection. The ultimate decision is a balance between the sample size, the cost and value of administrative data available, and the cost and anticipated response rate of surveying.

Due to the high cost of data, some data providers (notably, the Centers for Medicare and Medicaid Services) require proof of funding and a letter of support from the funding agency prior to finalizing a data use agreement.

UNDERSTANDING DATA UNIVERSE AND CONTENTS

It is essential to understand the data universe and data contents when planning to use administrative data. A natural first step is to request a data dictionary from the data provider, assuming one is not available online. While some data are very well documented, it is not uncommon for documentation to be incomplete, even for commonly used data sets.

Understand the universe from which the data are collected. If appearing in the data set is contingent on some type of program enrollment or measurement (e.g., enrollment in a social service, being a credit union member), be sure to understand any applicable eligibility criteria. This is crucial to understanding any potential for bias from using the data set, as described in *Differential Coverage: Observability in Administrative Data*.

> Consider the universe from which public school attendance data are collected. Students must live in the school district or receive special assignment to the school, and be enrolled in that school. Living in the district may be correlated with income or engagement in local activities. Receiving special assignment may be correlated with parental involvement or ability. Enrollment in a public school, rather than a private school, may be correlated with income, student achievement, parental motivation, or the availability of a school voucher program. Given these factors, it may not be appropriate to study educational achievement using public school records if the intervention is likely to be correlated with remaining in the school district or enrolling in private school.

Understand how and why data are recorded. For example, this includes whether data are actively reported by an individual, or collected passively. If reported, understand what the incentives are for accuracy, and if any incentives exist for intentional misreporting.

Understand the data contents. Even when provided with a data dictionary that describes the contents of a data set, this step often requires iteration with the data provider to clarify the meaning of the data elements in the data set, as variable names or descriptions may be ambiguous.

Understand the available identifiers. The identifiers available in the administrative data set (e.g., name, date of birth, Social Security number, Medicaid identification number) will vary by data provider, as will the subset of these identifiers that the provider is willing to share. To ensure that they can match study records on treatment assignment to administrative records, researchers should verify that there is overlap between the identifiers that they have access to prior to random assignment and those present in any administrative data they plan to use. See _Which Identifiers to Use to Link Data Sets_ for more information.

While identifying treatment or control status is essential to performing analysis for a randomized evaluation, and other individual-level characteristics may improve the analysis, researchers do not necessarily need direct access to personally identifying information to complete a randomized evaluation. For further information of this process, see _Formulating a Data Request_.

ETHICS

As with any evaluation involving human subjects, randomized evaluations using administrative data must be carried out in accordance with the principles of ethical research. In the US, the three guiding principles are respect for persons, beneficence, and justice. These principles are incorporated as legal statutes in the Federal Policy for the Protection of Human Subjects (the "Common Rule") and were previously codified in the Belmont Report.

These principles are incorporated into the review performed by Institutional Review Boards (IRBs), and have implications for all types of research, including evaluations that involve only the use of administrative data. For example, respect for persons obliges researchers to consider whether seeking informed consent prior to accessing individuals' administrative records is necessary and appropriate. While this may not always be necessary, it may be required in some cases. See _Compliance_ and _Informed Consent_ for more details. Beneficence requires researchers to minimize possible harms, including minimizing the risk of a breach of confidentiality.

The National Institutes of Health (NIH) provides a free tutorial on Protecting Human Research Participants and the Collaborative Institutional Training Initiative (CITI) at the University of Miami also provides certifications in Human Subjects Protections. These courses provide more information on the principles of ethical research and their practical applications and implications for research. Most IRBs will require one of these courses to be completed by researchers who interact with human subjects _or_ have access to identified data.

COMPLIANCE

Research involving administrative data is very likely to be considered human subjects research, and thus subject to review by the Institutional Review Board (IRB) of the researcher's home institution.[6] Beyond the purview of the IRB,

---

[6] There are very limited exceptions to this rule, and the exceptions may vary by institution. Check with your IRB for details.

an overlapping web of federal laws and regulations, state laws and regulations, and institutional restrictions and procedures governs access, storage and use of administrative data.

Because compliance requirements and definitions of "identifiable" data differ by field, source, and geography, the guidance of an IRB and legal counsel at the researcher's home institution can be critical in sorting through compliance and reporting requirements. This guide will focus on the US context. Research conducted in any country by US-based researchers is likely subject to a similar level of IRB review based on funding or university policy. As with US-based research, local regulations may apply as well.

Certain types of administrative data contain particularly sensitive information and require additional documentation and compliance prior to their use for research. Some of this is based on protecting the privacy of individuals. For example, most health data in the United States are created or owned by entities that are subject to the HIPAA Privacy Rule, which mandates additional documentation and protections for data that fall within its scope. Educational data may be subject to the Family Educational Rights and Privacy Act (FERPA), which has special rules to protect the privacy of student records. Criminal and juvenile justice records are also subject to particular rules, which may vary by district. Some requirements are imposed in order to protect proprietary business information, irrespective of individual privacy. For example, an insurer may want to keep their reimbursement rates and contracts private, or businesses may want to guard their trade secrets. Data providers may apply additional restrictions or conditions beyond what the law requires for the researcher to use their administrative data, especially for data considered "sensitive" for any reason.

Compliance requirements for using administrative data generally depend on how closely the data can be linked to particular individuals and how sensitive the data are. _Table 1: Levels of Identifiable Data_ gives a very basic overview of the three main "levels" of identifiability. Ultimately, the responsibility for determining the level of identifiability, and the applicable requirements, lies with the data provider and the IRB. Researchers are expected to make a good-faith judgment of the identifiability of data, given their expertise in data analysis. Minimizing the extent to which the researchers themselves handle identified data can provide for a much smoother process of accessing and using administrative data. For information on how to avoid direct contact with identified data, while retaining individual-level data and treatment assignment, see _Data Flow_.

TABLE 1: LEVELS OF IDENTIFIABLE DATA

De-identified	Limited or Partially De-identified	Research Identifiable
Very difficult or impossible to identify	More difficult to identify, but still possible, especially with additional knowledge	Very easy to identify confidently to one individual

The remainder of this section describes the permissions and processes for obtaining data at each of three levels of data identification, with additional detail added for health data due to the overlay of HIPAA.

As there is significant overlap between the requirements, it is recommended that the reader use Tables 1 and 2 to determine which level is likely to apply, review Tables 3 and 4 for overview requirements, and then read the relevant section for more details.

- *Research Identifiable Data*

- *Limited Data Sets (HIPAA/Health Only)*

- *De-identified or Publicly Available Data*

TABLE 2: LEVELS OF IDENTIFIABLE DATA (HIPAA)

While HIPAA provides just one of many ways of defining "identifiable" or "de-identified" data, this table may be a useful reference in understanding when a data set is likely to be considered de-identified.

DATA ELEMENTS ALLOWED BY HIPAA	DE-IDENTIFIED	LIMITED DATA SET	RESEARCH IDENTIFIABLE
Names or initials			X
Geographic subdivisions smaller than a state			X
Street address			X
City, County, and/or Precinct		X	X
ZIP code (5+ digits)		X	X
Equivalent geocodes		X	X
ZIP code (3 digits)* ** Provided that the geographic unit formed by combining all ZIP Codes with the same 3 initial digits contains >20,000 people*	X*	X	X
Dates directly related to an individual		X	X
Year*	X*	X	X
Any elements or dates indicative of age >89 ** e.g,. year of birth indicative of age >89 would not be permitted in a de-identified data set, though years are otherwise permitted*		X	X
Birth date		X	X
Admission or discharge date		X	X
Death date		X	X
Contact information			X
Telephone and/or fax numbers			X
Email addresses			X
Account numbers and codes			X
Social Security numbers			X
Medical record numbers			X
Health plan beneficiary numbers			X
Account numbers			X
Certificate/license numbers			X
Vehicle identifiers and serial numbers, license plate numbers			X
Device identifiers and serial numbers			X
URLs or IP addresses			X
Visual/biometrics			X
Biometric identifiers			X
Fingerprints or voiceprints			X
Full-face photographs or comparable images			X
Any other unique identifying number, characteristic, or code			X

1. See the National Institute of Health's guidance for more information.
2. HIPAA also defines a Minimum Necessary Requirement: Covered entities and business associates must make reasonable efforts to limit disclosures of protected health information to the minimum necessary to accomplish the intended purpose of the use, disclosure, or request. Though a limited data set is permitted to contain, for example, an individual's exact birthdate, the birthdate should only be included if it serves a specific research purpose.
3. CMS and ResDAC define limited data sets and research identifiable files slightly differently by excluding most dates from limited data sets.

Requirement Summary Tables

TABLE 3: SUMMARY OF HEALTH (HIPAA) DATA REQUIREMENTS

	De-identified or Publicly Available Data (HIPAA)	Limited Data Sets (HIPAA/Health Only)	Research Identifiable Data (HIPAA)
Data Use Agreements*	Process is typically simple or not required.	Almost always required. Somewhat less intensive than DUA for identified data.	Almost always required; usually involves significant review.
IRB Oversight	Frequently determined exempt by the IRB.	Frequently determined exempt by the IRB.	Likely requires ongoing review by the IRB.
Informed Consent or IRB Waiver	If research is determined exempt, not required by the IRB.		
	Likely possible to obtain a waiver of informed consent from the IRB as the risk to subjects is low.	Likely possible to obtain a waiver of informed consent from the IRB. More justification may be necessary than obtaining a waiver for de-identified data.	May be possible to obtain a waiver of informed consent from the IRB. Significant justification and proof of secure data protocols is likely necessary to obtain the waiver.
Individual Authorization for Research (HIPAA) or IRB Waiver	If research is determined exempt, not required by the IRB.		
	Likely possible to obtain a waiver of individual authorization as the risk to subjects is low.	Likely possible to obtain a waiver of individual authorization. More justification may be necessary than obtaining a waiver for de-identified data.	May be possible to obtain a waiver of individual authorization. Significant justification and proof of secure data protocols is likely necessary to obtain the waiver.
Data Security*	Intensive protocols typically are *not* required by either data provider or IRB.	Intensive protocols typically are required by either data provider or IRB.	Intensive protocols typically are required by both data provider and IRB.

*This table summarizes requirements based on the *identifiability* of administrative data. Researchers and data providers may take additional precautions based on the *sensitivity* of the data that does not necessarily pertain to individuals or identifiability. For example, trade secrets such as recipes do not pertain to individuals, and yet may never be released even under the strictest of data use agreements and data security protocols.

TABLE 4: SUMMARY OF GENERAL DATA REQUIREMENTS

	De-identified or Publicly Available Data (non-HIPAA)	*Research Identifiable Data* (non-HIPAA)
*Data Use Agreements**	Process is typically simple or not required.	Almost always required; usually involves significant review.
IRB Oversight	Frequently determined exempt by the IRB.	Likely requires ongoing review by the IRB.
Informed Consent or IRB Waiver	If research is determined exempt, not required by the IRB.	
	Likely possible to obtain a waiver of informed consent from the IRB as the risk to subjects is low.	May be possible to obtain a waiver of informed consent from the IRB. Significant justification and proof of secure data protocols is likely necessary to obtain the waiver.
Individual *Authorization for Research* (HIPAA)	Not applicable for nonhealth (i.e., non-HIPAA) data.	
Data Security*	Intensive protocols typically are *not* required by either data provider or IRB.	Intensive protocols typically are required by either data provider or IRB.

*This table summarizes requirements based on the *identifiability* of administrative data. Researchers and data providers may take additional precautions based on the *sensitivity* of the data that does not necessarily pertain to individuals or identifiability. For example, trade secrets such as recipes do not pertain to individuals, and yet may never be released even under the strictest of data use agreements and data security protocols.

Research Identifiable Data

Research identifiable data contain sufficient identifying information such that the data may be directly matched to a specific individual. Usually, such data contain direct individual identifiers, such as names, Social Security numbers, physical addresses, e-mail addresses or phone numbers. However, data that contain a combination of elements (such as date of birth, ZIP code, and gender) may be considered "identifiable" *even if they do not contain names, Social Security numbers, or any other single direct identifier*.

In general, research involving such data requires active IRB oversight, formal *Data Use Agreements* (DUAs), robust *data security* measures, and compliance with any additional requirements, such as those imposed by the data provider, or federal or state law.

Additional Considerations for Health Data
Research identifiable health data obtained by certain data providers (e.g., a health plan, health insurer, or health-care provider) are often subject to the specific requirements and restrictions of the HIPAA Privacy Rule. These data, which are also known as *Protected Health Information (PHI)*, are particularly sensitive. Disclosure of these data is subject to tight restrictions and may pose substantial risks, including legal liability, for the data provider. A list of data types that make a data set "identifiable" under the HIPAA Privacy Rule can be found in *Table 2: Levels of Identifiable Data (HIPAA)*.

Researchers are advised to confirm whether their implementing partner or the data provider is subject to HIPAA. Some organizations will incorrectly assume they are subject to HIPAA, thus invoking unnecessary review and regulation. For guidance on determining whether an organization is subject to the HIPAA Privacy Rule (i.e., is a "covered entity"), the US Department of Health & Human Services (HHS) describes the definition of a covered entity and provides a flowchart tool.

Because the HIPAA Privacy Rule imposes liability on health data providers when identified data are involved, health data providers may be particularly cautious about how and with whom they share data. The Privacy Rule's complex nature may make it difficult for health data providers (and researchers) to understand their obligations, causing them to err on the side of caution. Successful research relationships with these data are often predicated on a substantial amount of trust and goodwill.

HHS provides a detailed guide to the requirements associated with research identifiable health data and how the HIPAA Privacy Rule applies to research. For other resources on HIPAA, see *Resources for HIPAA*.

Limited Data Sets (HIPAA/Health Only)

Under the HIPAA Privacy Rule, data sets that exclude specific data items, as listed on *Table 2: Levels of Identifiable Data (HIPAA)*, are called limited data sets. In general terms, limited data sets exclude direct identifiers of individuals of record, or of their relatives, employers, or household members, such as their addresses and Social Security numbers. These data do contain, however, data elements (e.g., birthdate, ZIP code, and gender) that may allow others to make a reasonable guess as to the identity of the subjects.

Compared to research using identifiable data, research using limited data sets may benefit from a simpler process of obtaining the data from the provider and less restrictive oversight. However, these data are still considered *Protected Health Information* (PHI) and are still much more sensitive than de-identified data.

While the risks associated with limited data sets are lower than those associated with identifiable data, data providers may still face substantial risk if the data are mishandled. While the HIPAA Privacy Rule imposes fewer restrictions on the sharing of limited data sets than identified data sets, the complexity of the privacy rule may cause providers of health data to err on the side of caution.

De-identified or Publicly Available Data

De-identified data are data that do not contain sufficient identifiers to link to specific individuals with certainty. Relative to research using identified data, research using de-identified data is likely to benefit from a simpler process of obtaining the data and minimal IRB or compliance oversight. The lower risk of re-identification with de-identified data reduces the risk (i.e., legal liability) to data providers of releasing de-identified data. As a result, data providers are less likely to impose additional restrictions or requirements on data requestors.

While it is often necessary to link individual subjects to their treatment status and outcome measures, it may be possible to obtain the necessary individualized detail while avoiding identifiers by having a party other than the researcher (e.g., the data provider or a partner organization) link data sets. A description of this process can be found in the *Data Flow* section.

However, even data that do not pertain to individuals at all may be highly classified or confidential; consider the recipe for Coca-Cola or the Google search algorithm.

FORMULATING A DATA REQUEST

In order to extract data from their database, data providers will often need their programmers to write a query to pull the specific data necessary for the research. Researchers should not assume that the programmers have background knowledge about their particular research project, and therefore should write clear and concise data requests that include:

- Timeframe for request, in calendar months/years

- Acceptable or preferable data format (e.g., ASCII, SAS, Stata)

- Data structure (e.g., multiple tables with unique keys for merging versus single, pre-merged data set, long or wide form)

- List of variables requested

- Clarification or notes for each variable

Specify whether identified data or other elements likely to be "sensitive" are necessary. Many data providers will assume you need access to explicit identifiers, such as name or Social Security number, and may stop conversations short based on this understanding. If you do not need access to identified data, be explicit about this, as it will likely speed the process.

Initial data request:

- Data for the year before and year after the intervention
- Distance from home to school
- Number of children in the household enrolled in school
- Date of birth
- Age

Data provider comments:

"How do you want us to calculate distance? We have an indicator whether there is a school-bus route, and which route it's on. Do you really need both date of birth and age? We assume you want only elementary school records. When is the intervention?"

Updated, refined data request:

- School address
- Enrollment list for all district schools at all levels for the 2014–15 and 2015–16 school years
  - Student identifier
  - Date of birth
  - Student home address
  - Bus route (yes/no)
  - Bus route code

Beware of requesting "all of the data." While it may be appealing to request "all" of the data or "as much as possible," there are dangers with this approach. Without specifying the necessary data elements, researchers may receive less data than they anticipate, or may complete a data use agreement that does not guarantee access to all needed elements. Such a request may also raise red flags with the data provider's legal team, and create unnecessary hurdles if sensitive data are not truly required.

Further, data sets that are not necessarily sensitive or identifiable on their own may become more sensitive when combined with additional external data sets.

Request specific variables of interest rather than general subjects. For example, researchers may wish to measure the effect of a job training program on the propensity of individuals to hold a "professional" job. Rather than requesting general indicators of having a professional job, researchers should request specific variables, such as job codes and titles. Given a list of specific variables, the data provider may be able to suggest a closely related additional variable of interest: for example, the industry code.

Ensure researchers and data providers understand the data and the request in the same way. Often, researchers make a request for data with a specific project and set of assumptions in mind. Without understanding this background, the data provider may interpret a seemingly straightforward request in a vastly different way. For

example, a request for "all checking transactions" may be interpreted as all transactions that occurred through a written check, or all transactions from the checking account. A phone call or detailed email exchange can help to identify and resolve any problems.

Identify points of contact within the data provider organization. Many data requests are made unnecessarily complicated due to confusion from staff working outside of their normal area of expertise. Successful relationships and requests often involve a high-level owner, director, or policy representative, who understands the general need and importance of the data request and can push the request through any bureaucracy; a data or IT expert who understands the available data elements and would work through the data extraction, matching, and transfer; a programmatic or institutional expert who understands how the data relate to the research questions; and a legal expert who can provide detail on compliance issues and map the pathway to creating a data use agreement.

Leave as little room for interpretation as possible. Request raw inputs rather than calculated or aggregated outcomes. For example, rather than requesting "age", request the date of service and the date of birth, from which age can be calculated. Rather than requesting "number of transactions above $100," request the dollar amount of all transactions.

DATA FLOW

Taking into account any ethical, legal, or resource restrictions, a data flow strategy that maps how data will be linked across administrative, survey, and intervention data should be developed during the design phase of a randomized evaluation. The data flow strategy should include:

- How identifying information will be gathered for the study sample

- Which identifiers will be used to link intervention data and treatment assignment with administrative data

- Which entity or team will perform the link

- What algorithm will be used to link data

- What software will be used to link data

In the simplest case, a data provider may be willing and able to send a fully identified data set directly to the researchers, leaving all other steps of the linking process in the research team's hands. Often, ethical or legal restrictions will preclude this. The following sections describe data flow strategies for more complex scenarios.

To motivate this section, we will use a hypothetical example where random assignment will take place from an existing list of individuals that includes baseline characteristics, and outcomes data will be collected from two external sources.

LIST OF INDIVIDUALS FOR THE RANDOMIZED EVALUATION:

NAME	SSN	DOB	INCOME	STATE
Jane Doe	123-45-6789	5/1/50	$50,000	FL
John Smith	987-65-4321	7/1/75	$43,000	FL
Bob Doe	888-67-1234	1/1/82	$65,000	GA
Adam Jones	333-22-1111	8/23/87	$43,000	FL
James Trudu	123-45-9876	5/17/60	$50,000	FL
Joyce Gray	587-157-8765	7/28/57	$43,000	FL
Philippe Zu	224-85-6879	3/30/80	$65,000	GA
Alanna Fay	341-78-3478	11/10/79	$50,000	FL

DATA AVAILABLE FROM DATA PROVIDER A:

NAME	SSN	DIABETIC?
Jane Doe	123-45-6789	Y
John Smith	987-65-4321	N
Bob Doe	888-67-1234	N
Adam Jones	333-22-1111	Y
James Trudu	123-45-9876	N
Joyce Gray	587-157-8765	Y
Philippe Zu	224-85-6879	N
Alanna Fay	341-78-3478	Y

DATA AVAILABLE FROM DATA PROVIDER B:

NAME	DOB	OWN A CAR?
Jane Doe	5/1/50	Y
John Smith	7/1/75	Y
Bob Doe	1/1/82	Y
Adam Jones	8/23/87	N
James Trudu	5/17/60	Y
Joyce Gray	7/28/57	N
Philippe Zu	3/30/80	N
Alanna Fay	11/10/79	Y

How Identifying Information Will Be Gathered for the Study Sample

Researchers and/or their implementing partner may have direct access to a set study sample--including identifying information--through existing administrative data. For example, this may be a list of students in a classroom, existing customers at a bank, physicians in a hospital, or residents of a town with an existing, complete, identified census.

In other cases, researchers may need to proactively identify or recruit a sample that has not been previously defined or identified. For example, new customers at a store, clients at a bank, patients who visit a doctor's office, or

residents of a town for which researchers do not have census information. In this case, identifying information is likely to be obtained proactively through a baseline survey or intake form.

Which Identifiers to Use to Link Data Sets

Researchers must select which identifiers to use to link data sets based on the set of identifiers available in both the administrative data and program-implementation data. Given that constraint, there are several tips to take into account when deciding which of these identifiers to use to match data sets:

Use only those identifiers available prior to random assignment (e.g., those collected at intake/baseline) to link to participants' administrative records. Through engagement with individuals through the intervention, it may be possible to obtain more detailed or accurate identifiers for some individuals. If this information is more readily available for treatment group than control group members, using updated information may introduce bias, as described in the section on *Differential Coverage: Observability in Administrative Data*.

Study participants may be hesitant or unwilling to provide sensitive numeric identifiers at intake (e.g., Social Security number, student ID, account numbers). Even if willing, they may not know this information off-hand. While this problem should be balanced across treatment and control groups, as it applies prior to random assignment, a lack of adequate identifiers can reduce effective sample size. Quickly establishing rapport with participants, specifying what information they will be asked to provide, and explaining how the privacy and confidentiality of their data will be protected throughout the course of the study, may help to maximize the number of identifiers that you are able to collect at baseline.

It may be possible to confirm individual's identifiers, such as SSN or account numbers, through administrative data, provided that participants' consent is obtained prior to randomization and that the records themselves pre-date the random assignment.

Use numeric identifiers if possible. These are less prone to typos, misspellings or alternative versions than string identifiers (i.e., words or text such as name or address), making them easier to match.

Use identifiers that are not subject to frequent change. Date of birth and SSN do not change with time. In contrast, names are subject to change, especially among women through marriage. Sometimes, unique identifiers that are used as a part of the program or intervention may change over time. For example, participants may be assigned a key tag with a unique code that grants access to library services. If a participant loses the key tag, they may be assigned a new code and key tag during the intervention and evaluation period. Having related records and backup identification methods can insure against this problem.

Which Entity Should Perform the Link

Some agencies have experience accommodating requests for data extraction and matching to a specific list of study participants, and are willing and able to perform these matches following a relatively straightforward discussion or agreement with the researchers. Other agencies may be less willing to perform data matches due to fears about confidentiality or legality, or they may be resource-constrained in their technical ability or personnel time. In these situations, researchers may want to suggest creative solutions to providing hands-on assistance with the matching while maintaining confidentiality.

Data agencies may prefer to use their own staff to match administrative records with lists of study participants, or they may allow researchers to do the matching themselves, either on-site or on a secure device provided and monitored by the data provider.

The decision of which entity should perform the match, and which data flow strategy should be chosen, should be made while considering applicable regulations, confidentiality, data security capabilities, and the potential need to maintain the ability to re-identify the data in order to add follow-up data or new data sources.

Data Flow Strategies

Data flow option one. For a defined sample of specific individuals, researchers may send the data provider a "finder file" that contains participants' unique study IDs (as assigned by the researcher) and personal identifiers. The data provider will use the personal identifiers in the finder file to select the matching administrative records. The data provider will then strip off all _Personally Identifiable Information_ (PII), and send back the matched data, including the study IDs. When the data is shared in this way, the data provider often requires the researcher to store the finder file and the de-identified analysis file separately, and to agree never to match them.

FIGURE 2: DATA FLOW OPTION ONE

Researcher	Finder file (Study ID, name, DOB) →	Data Provider
	← De-identified analysis file	
• Researcher agrees **never** to match finder file and de-identified analysis file		• Conducts match • Strips identifiers off of data set (but keeps study ID).
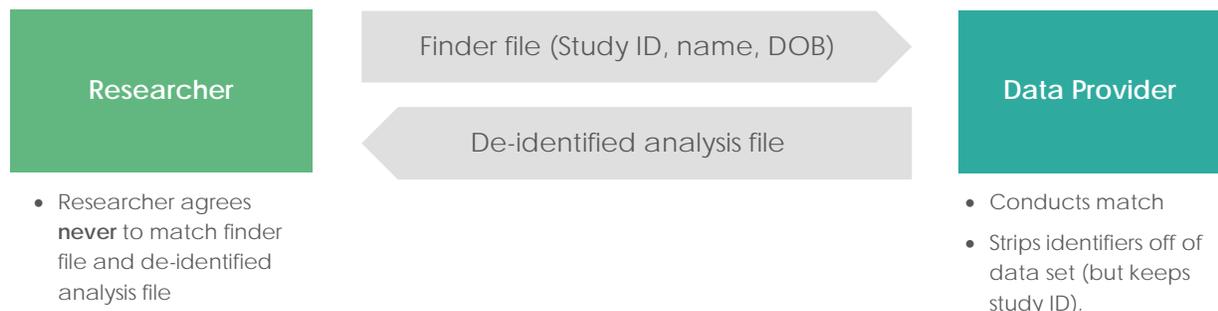
Using our example, the researcher will send the following finder file to both Data Provider A and Data Provider B. Since Data Provider A only has Social Security numbers (SSN), and B only has dates of birth (DOB), the researcher may alter the finder file for each data provider. In this scenario, the researcher creates the Study ID.

FINDER FILE: DATA FLOW OPTION ONE

NAME	SSN	DOB	STUDY ID
Jane Doe	123-45-6789	5/1/50	1
John Smith	987-65-4321	7/1/75	2
Bob Doe	888-67-1234	1/1/82	3
Adam Jones	333-22-1111	8/23/87	4
James Trudu	123-45-9876	5/17/60	5
Joyce Gray	587-157-8765	7/28/57	6
Philippe Zu	224-85-6879	3/30/80	7
Alanna Fay	341-78-3478	11/10/79	8

Data Providers A and B will return the following files:

ANALYSIS FILES: DATA FLOW OPTION ONE

STUDY ID	DIABETIC?		STUDY ID	OWN A CAR?
1	Y		1	Y
2	N		2	Y
3	N		3	Y
4	Y		4	N
5	N		5	Y
6	Y		6	N
7	N		7	N
8	Y		8	Y

The researcher will create the following files:

ANALYSIS DATASET

TREATMENT ASSIGNMENT	STUDY ID	INCOME	STATE	DIABETIC?	OWN A CAR?
Treatment	1	$50,000	FL	Y	Y
Control	2	$43,000	FL	N	Y
Treatment	3	$65,000	GA	N	Y
Control	4	$43,000	FL	Y	N
Treatment	5	$50,000	FL	N	Y
Control	6	$43,000	FL	Y	N
Treatment	7	$65,000	GA	N	N
Control	8	$50,000	FL	Y	Y

IDENTIFIERS-STUDY ID CROSSWALK

NAME	SSN	DOB	TREATMENT ASSIGNMENT	STUDY ID
Jane Doe	123-45-6789	5/1/50	Treatment	1
John Smith	987-65-4321	7/1/75	Control	2
Bob Doe	888-67-1234	1/1/82	Treatment	3
Adam Jones	333-22-1111	8/23/87	Control	4
James Trudu	123-45-9876	5/17/60	Treatment	5
Joyce Gray	587-157-8765	7/28/57	Control	6
Philippe Zu	224-85-6879	3/30/80	Treatment	7
Alanna Fay	341-78-3478	11/10/79	Control	8

In this method, the researcher maintains full control over the crosswalk between the identifying information, the Study IDs, and the random assignment indicator. This ensures that they will have the technical ability to seek additional data sources in the future without relying on continued cooperation from external partners.

However, if the administrative data are very sensitive, the data provider may not feel comfortable with the researcher having access to both the finder file and the de-identified analysis file, because the study ID could be used to re-identify the analysis file. Options two and three detail data flow strategies that address this concern.

Data flow option two. A close research partner (e.g., an implementing partner or trusted third party) may be made responsible for maintaining the participant list and sending the finder file to the data provider, without ever sharing the identities of the individuals with the researchers. The data provider would conduct the match and then send the de-identified analysis file directly to the researcher. In this scenario, the research partner never has access to the administrative data, and the researcher never has access to the personal identifiers.

FIGURE 3: DATA FLOW OPTION TWO
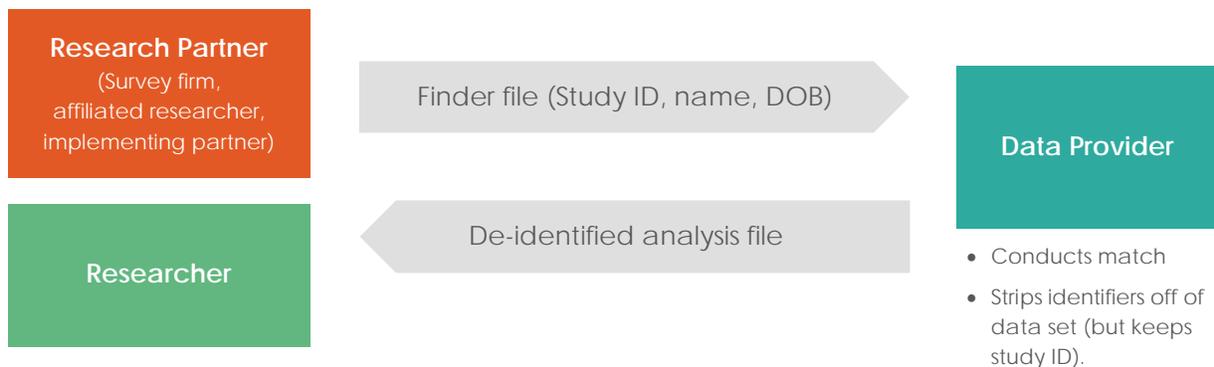
Research Partner never has access to administrative data.
Researcher never has access to personal identifiers.

If the research partner is willing and able to perform this function, option two is a good fallback for maintaining separation of identified data from outcomes. When considering this option, researchers should consider the strength of the relationship and set realistic expectations for future collaboration. Because the partner, and not the researcher, maintains the identified list of the study sample, the partner's cooperation would be necessary for any follow-up or long-term evaluations. To ensure that the crosswalk between identifiers and Study ID is maintained, researchers may consider supporting the cost of IT infrastructure and data storage facilities.

The implementer (or third party) will send the following finder file to both Data Provider A and Data Provider B. Since Data Provider A only has Social Security numbers (SSN), and B only has dates of birth (DOB), the partner may alter the finder file for each agency. In this example, the implementer generates the Study ID.

FINDER FILE: DATA FLOW OPTION TWO

NAME	SSN	DOB	STUDY ID
Jane Doe	123-45-6789	5/1/50	1
John Smith	987-65-4321	7/1/75	2
Bob Doe	888-67-1234	1/1/82	3
Adam Jones	333-22-1111	8/23/87	4
James Trudu	123-45-9876	5/17/60	5
Joyce Gray	587-157-8765	7/28/57	6
Philippe Zu	224-85-6879	3/30/80	7
Alanna Fay	341-78-3478	11/10/79	8

Data Providers A and B will send the following files to the researcher (and not necessarily to the partner). The partner or implementer will send a similar file including Study ID, Treatment Assignment, Income, and State to the researcher.

ANALYSIS FILES: DATA FLOW OPTION TWO

STUDY ID	DIABETIC?	STUDY ID	OWN A CAR?
1	Y	1	Y
2	N	2	Y
3	N	3	Y
4	Y	4	N
5	N	5	Y
6	Y	6	N
7	N	7	N
8	Y	8	Y

The researcher will create the analysis dataset as in the previous example. The implementer/partner will create and maintain the crosswalk.

Data flow option three. For very sensitive data, providers may not be willing to release any information that could, hypothetically, be linked to additional characteristics or specific individuals. For example, some jurisdictions protect vital statistics natality data in this way. These data must be analyzed in isolation, or with only a limited set of additional characteristics approved by the data provider.

The identified list may originate from the researcher, implementer, or a partner. Baseline data may be included in the finder file. The data provider will match the finder file with their data, and de-identify the file to its standards. This

may include the removal, censoring, or categorization of certain baseline variables. For example, ages may be converted into ranges, or indicators of low-probability events such as arrest or rare disease may be removed.

In this method, the data provider retains all control over the crosswalk between the finder file and outcomes or administrative data. The data provider may (or may not) maintain this crosswalk internally. If the provider does maintain this crosswalk, they will be able to provide updated administrative data in the future that can be linked to their previous data transfer.

The implementer (or third party) will send the following finder file to both Data Provider A and Data Provider B. Since Data Provider A only has Social Security numbers (SSN), and B only has dates of birth (DOB), the partner may alter the finder file for each agency. In this example, the implementer generates the Study ID.

FINDER FILE: DATA FLOW OPTION THREE

NAME	SSN	DOB	INCOME	STATE	TREATMENT ASSIGNMENT
Jane Doe	123-45-6789	5/1/50	$50,000	FL	Treatment
John Smith	987-65-4321	7/1/75	$43,000	FL	Control
Bob Doe	888-67-1234	1/1/82	$65,000	GA	Treatment
Adam Jones	333-22-1111	8/23/87	$43,000	FL	Control
James Trudu	123-45-9876	5/17/60	$50,000	FL	Treatment
Joyce Gray	587-157-8765	7/28/57	$43,000	FL	Control
Philippe Zu	224-85-6879	3/30/80	$65,000	GA	Treatment
Alanna Fay	341-78-3478	11/10/79	$50,000	FL	Control

In this example, it would not be possible to combine the data from Data Providers A and B. Data Provider A will send the following file to the researcher. Additionally, they may choose to censor the income or geographic data, if they believe it is too specific and potentially identifying.

ANALYSIS DATASET

TREATMENT ASSIGNMENT	INCOME	STATE	DIABETIC?	OWN A CAR?
Treatment	$40,000 – $60,000	FL	Y	Y
Control	$40,000 – $50,000	FL	N	Y
Treatment	> $60,000	GA	N	Y
Control	$40,000 – $60,000	FL	Y	N
Treatment	$40,000 – $60,000	FL	N	Y
Control	$40,000 – $60,000	FL	Y	N
Treatment	> $60,000	GA	N	N
Control	$40,000 – $60,000	FL	Y	Y

The data provider may maintain the crosswalk.

Data flow option four: researchers perform match on-site at data provider. Some agencies may not have the capacity to extract and match data for researchers or the ability to freely share a fully identified data set. It may be possible for a researcher to conduct or assist with the match on-site at the data provider, under the supervision of provider staff, and leave with only a de-identified data set.

This may be an informal process, or the researcher may sign forms to act as a volunteer employee of the data provider for a day or more, or sponsor an assistant seated at the data provider. The provider may wish to restrict the researcher's access to the internet or storage devices in order to prevent the release of confidential data during the match. This approach reduces the burden on the data provider while maintaining a high level of confidentiality.

In this case, the researcher may or may not be permitted to retain a crosswalk between identifiers and the Study ID. The data provider may analyze the resulting de-identified data set and request further action, as in option three.

This option allows the researcher to maintain greater control over the matching algorithm. However, while the burden of data matching is lifted off of the data provider, the researchers must invest in travel, and may have a very limited timeframe to complete the linkage and solve any issues that may arise.

Algorithms for Linking Data

In general, there are two types of matching strategies: "exact" (deterministic) and "fuzzy" (probabilistic). Regardless of whether the researcher, the data provider, or a third party is performing the match, it is important to understand and document the matching strategy. Using an exact matching strategy minimizes the number of false positives (i.e., a "match" is found, but is in fact not the same individual), but may maximize the number of false negatives (i.e., a "match" is not found, but the same individual is present in both files).

Exact matching. With "exact" matching strategies, specific identifiers in the finder file are matched to the identifiers in the administrative data sets. If there are minor discrepancies (e.g., reversed month and day in date of birth, or typos in the last name), records will not be identified as a match even though they may be.

Probabilistic, or fuzzy matching. With fuzzy matching strategies, a more sophisticated algorithm is used to account for the fact that identifiers in the finder file may not match exactly to those in the administrative data sets, but may be close enough to be considered a good match. For reproducibility and transparency, the algorithm or protocol for determining a fuzzy match should be explicitly stated and followed. Manually reconciling data sets and making judgment-based decisions is *not* a reproducible method.

For example, when linking two data sets, which records are identified as "matches" depends heavily on the variables that are used for the match, and the types and amounts of variances allowed for in a fuzzy match.

RECORD	NAME	SSN	DOB
A	Jane Doe	123-45-6789	5/1/1950
B	Jonathan Smith	987-65-4321	7/1/1975
C	Bob Doe	888-67-1234	1/1/1982
D	Adam Jones	333-22-1111	8/23/1987
E	Maria Anna Lopez	532-14-5578	10/15/1965
F	Sarah Franklin	333-22-1111	8/23/1987

RECORD	NAME	SSN	DOB
1	Jane Doe	123-45-6789	1/5/1950
2	John Smith	987-65-4321	7/1/1975
3	Bob Doe	888-67-1243	1/1/1982
4	Adam Jones		8/23/1987
5	M. Anna Lopez	532-14-5578	10/15/1965
6	Sarah Franklin	333-22-1111	8/23/1987

RECORDS	FUZZY MATCH	EXACT MATCH ON SSN ONLY	EXACT MATCH ON NAME, SSN, DOB	EXACT MATCH ON NAME, DOB
A+1	Could be matched	Match	No Match	No Match
B+2	Could be matched	Match	No Match	No Match
C+3	Could be matched	No Match	No Match	Match
D+4	Could be matched	No Match	No Match	Match
E+5	Could be matched	Match	No Match	No Match
F+6	Matched	Match	Match	Match

Depending on the structure of the data and the treatment of nonmatched records, the matching errors may decrease the precision of impact estimates. Under some circumstances, the (lack of) balance between the rate of false positives and false negatives may even bias the estimates. For example, consider the case of matching housing-voucher recipients with their arrest records. Because arrest records only contain those individuals who were arrested, there is no file that will confirm whether an individual was *not* arrested, so the researcher would likely attribute a value of "no arrests" to any records that do not have a positive match with arrest records. Relying on an exact-match strategy will minimize the risk that crimes are falsely attributed to individuals who did not commit crimes. However, false negatives may lead to individuals who did commit crimes being attributed a clean slate. Though this type of matching error may be random across treatment and control, it may still lead to bias and/or to decreased precision in impact estimates (Tahamont et al. 2015; Dynarski et al. 2013).

Consider researchers who have the following list of program participants:

RECORD	NAME	SSN	DOB
A	Jane Doe	123-45-6789	5/1/1950
B	Jonathan Smith	987-65-4321	7/1/1975
C	Emer Blue	547-94-5917	8/15/00
D	Amy Gonzalez	431-54-9870	12/2/89
E	Bob Doe	888-67-1234	1/1/1982
F	Adam Jones	333-22-1111	8/23/1987
G	Maria Anna Lopez	532-14-5578	10/15/1965
H	Sarah Franklin	333-22-1111	8/23/1987

They receive the following arrest records from the state criminal justice agency:

RECORD	NAME	SSN	DOB	ARRESTS
1	Jane Doe	123-45-6789	1/5/1950	5
2	John Smith	987-65-4321	7/1/1975	3
4	Amy Gonzalez	431-54-9870	12/2/89	2
6	Adam Jones		8/23/1987	9
7	M. Anna Lopez	532-14-5578	10/15/1965	2

Depending on which matching strategy is used, a clean arrest record may be attributed to an individual, even if they were, in fact, arrested.

Arrests attributed based on exact match

NAME	ON NAME ONLY	SSN ONLY	ON NAME, SSN, DOB	ON NAME, DOB
Jane Doe	5	5	0	0
Jonathan Smith	0	3	0	0
Emer Blue	0	0	0	0
Amy Gonzalez	2	2	2	2
Bob Doe	0	0	0	0
Adam Jones	9	0	0	9
Maria Anna Lopez	0	2	0	0
Sarah Franklin	0	0	0	0

Software for Data Linkage

There are several options that may be used for computing a probabilistic ("fuzzy") match. Stata's "reclink" provides one option that may be familiar to social-science researchers. Software programs that specialize in matching include Merge Toolbox and Link Plus. Other researchers have used Elastic Search and Open Refine (formerly Google Refine).

DATA USE AGREEMENTS

A Data Use Agreement (DUA) documents the terms under which a data provider shares data with a researcher's home institution for use by the researcher. This agreement, which typically must be approved by legal counsel at the researcher's home institution, contains a number of provisions that can significantly impact the underlying research. Many universities have a standard template that includes terms and conditions that are acceptable to the university, and were created with researchers' needs in mind. Using a pre-vetted template may simplify the review process at the institution that created the template. MIT has made samples of data use agreements and nondisclosure agreements available.

Described below are certain items commonly found in DUAs that are particularly important for researchers to understand when negotiating with a data provider:

- **Review periods and academic freedom to publish.** Data providers may request the right to review the research manuscript prior to publication, either to preview the results or to identify the inadvertent disclosure of confidential information. Such review may be acceptable so long as the researcher retains the final freedom to publish at their sole discretion, after reviewing any data provider comments in good faith. Researchers may also want to set reasonable limits on the time allotted to review prior to publication. This provision is important, as providers may attempt to suppress unfavorable results.

- **Limits on the publication of summary statistics.** Some data providers may impose strict cutoffs for the number of individuals in a "cell" or disaggregated summary statistics. Researchers should be aware of these limits and be sure they are reasonable.

- **Individual liability.** DUAs are typically entered into by the data provider and the researcher's home institution and should not personally expose the researcher to the risk of lawsuit if the agreement is breached. Researchers should take note of any forms or addenda requiring individual signatures and consult with their legal counsel prior to signing.

- **Destruction or return of the data.** Data providers typically require that researchers return or destroy personally identifiable information (or all information) after a particular period of time. Researchers should be sure the time allotted is sufficient to cover the research, publication, and academic journal review process, taking into account whether identifiers may be necessary to link with additional data obtained in the future, or whether data may need to be reviewed at a later date.

Below, common elements and tips for negotiating a DUA are described:

- **Data description**. While some data providers require a detailed list of data elements requested, others will accept a general description of the data. A general description may allow additional elements to be added more easily without an official amendment, though the researcher may not be ensured access to all necessary elements.

- **Data timeframe.** Many data providers allow researchers to request several years of data prospectively; others will require an amendment each year to request new data.

- **Frequency and time schedule of data transfers.** The first data extraction and transfer may take more time and require more communication and iteration than subsequent data pulls. It can be helpful to schedule and plan a "practice" or "sample" data transfer.

- **Research subject privacy**. Data providers may require the researchers to agree not to use the data within the data set to contact the subjects, their relatives, or obtain additional information about them (i.e., conduct a "follow-back" investigation).

- **Personnel.** Some data providers require a list of each individual who will have access to the data, while others allow principal investigators to identify "classes" of individuals (e.g., students or employees) who will have access. Identifying classes rather than specific individuals may reduce administrative burden by minimizing the need for future amendments when staff are added or removed from a project.

In addition to the elements described above, DUAs may require researchers to provide a study protocol, and typically contain provisions related to data security, confidentiality, IRB or Privacy Board review, the rerelease of data, and the allocation of the liability between the data provider and the researcher's home institution.

TIMELINE

Gaining access to administrative data is a multifaceted process that should be initiated during the design phase of any research project. The time required to establish a data use agreement can vary widely, and depends on the data provider's capacity to handle such data requests, the sensitivity of the data, and the levels of review that must be undertaken by both sides before a legal agreement can be signed. In a 2015 analysis of data acquisition efforts with 42 data agencies, the MDRC found that it typically takes 7 to 18 months from initial contact with a data provider to the completion of a legal agreement.[7]

Much of this time is spent in a tandem process of obtaining both IRB and legal approval for a data request. IRBs often require researchers to furnish signed DUAs before approving a study protocol, and data providers often require IRB approval before signing a DUA. Both processes can involve lengthy review periods, and changes made by one entity must be reviewed and approved by the other.

Many administrative data sets are available on a lagged basis, to allow time for data entry and cleaning. Beyond this planned data lag, researchers should plan to allow additional time for the data provider to extract and transfer data. This may take several weeks or months, depending on the provider's capacity and workload. Especially in the first transfer, researchers should expect to iterate with the data provider on matching strategies and data definitions.

Having clear specifications in the data use agreement or memorandum of understanding may help to set expectations between the data provider and the researcher regarding extraction date(s) and frequency.

ENCOURAGING DATA PROVIDER COOPERATION

Additional incentives may encourage the data provider to cooperate and prioritize a request for administrative data. Depending on the frictions preventing cooperation, incentives may range from a letter of support from a high-ranking official within the data provider's organization, monetary compensation for staff time and IT resources, or in-kind

---

[7] See Lee et al. (2015), especially Figure ES.1 Data Acquisition Process: Typical Length of Time to Complete Each Step in The Mother and Infant Home Visiting Program Evaluation-Strong Start report.

assistance through providing data analysis relevant to the provider's needs. To solve issues of trust, researchers have sought co-investigators from within a data provider's organization, and brought on consultants who the data provider organization trusts or who previously worked for the provider.

Data providers may have access to a wealth of data, but they may not have analytics as their primary objective. For example, credit unions, government departments of health and education, and school systems may have access to data that are extremely useful for researchers. However, these organizations are focused on providing services, and may have very small (or nonexistent) IT and data analysis teams. Supplying the provider with relevant and interesting analyses of their own data on an ongoing basis may be very valuable, and may help them prioritize the researcher's needs.

A more intensive approach is to offer an intern or research assistant who could work directly for the provider. The research assistant's role would be to assist with the data extraction for the randomized evaluation, but also to assist the provider in other ways. When offering this type of support, it's best to concretely define and explain how the provider could benefit from the research assistant's time, and how this would be a logistically viable solution for the provider (e.g., researchers may offer to conduct background checks, oversee the onboarding process and payments, etc.).

DATA SECURITY

Data security is critical to protecting confidential administrative data, respecting the privacy of research subjects, and complying with applicable protocols and requirements. Even seemingly de-identified data may be re-identified if enough unique characteristics are included,[8] and the information revealed could be damaging in surprising ways. For example, Arvind Narayanan successfully re-identified a public-use de-identified data set from Netflix. Through this, he was able to discover their apparent political preferences and other potentially sensitive information (Narayanan and Shmatikov 2008).

Many research universities provide support and guidance for data security through their IT departments and through dedicated IT staff in their academic departments. Researchers should consult with their home institution's IT staff in setting up data security measures, as the IT department may have recommendations and support for specific security software.

In addition to working with data security experts, researchers should acquire a working knowledge of data security issues to ensure the smooth integration of security measures into their research workflow and adherence to the applicable data security protocols. As well, researchers should ensure their research assistants, students, implementing partners, and data providers have a basic understanding of data security protocols.

Data-security measures should be calibrated to the risk of harm of a data breach and incorporate any requirements imposed by the data provider. Harvard University's classification system for data sensitivity and corresponding requirements for data security illustrates how this functions in practice.[9]

---

[8] For a review, see "The Re-Identification Of Anonymous People With Big Data."

[9] The full Harvard Research Data Security Policy can be found here.

This section provides a primer on some basic data security themes, as well as context on elements of data security that are particularly relevant for randomized evaluations using individual-level administrative data. For more in-depth coverage of how to implement data security, reference the *Data Security* appendix.

Data Security Breaches: Causes and Consequences

A data security breach can result in serious consequences for research subjects, the researcher's home institution and the researcher. Research subjects may suffer unintentional disclosure of sensitive identified information and may be exposed to identity theft, embarrassment, and other financial or emotional harms. Both the researcher's home institution and the researcher may suffer reputational damage and may have more difficulty obtaining sensitive data in the future. Additional compliance requirements will likely be triggered by a breach, including reporting the data breach to the IRB, and, in certain circumstances, to each individual whose data was compromised. The data provider may require additional security protections or terminate access to the data. There may, in some cases, be financial and/or criminal liability to the data provider and/or the research subjects.[10]

Sensitive data are vulnerable to both inadvertent disclosure and targeted attacks. If data security protocols are not adhered to, data may be disclosed through email, laptop loss, file-sharing software such as Box or Dropbox, or improper erasure of files from hardware that has been recycled, donated or disposed of. Theft or a hacking attack may target either a researcher's specific data set or the researcher's home institution more generally and inadvertently sweep up the researcher's data set in the course of the attack. Sensitive data must be protected from all of these threats.

Minimizing Data Security Threats

Minimizing the research team's contact with sensitive individually identifiable data may substantially reduce the potential harm caused by a data breach and the required data security measures that need to be put in place. This will often simplify and accelerate the research data flow.

Researchers may reduce the data security threat-level a priori by acquiring and handling directly only the minimum amount of sensitive data strictly needed for the research study. Researchers may, for example, request that the data provider or a trusted third party link particularly sensitive individualized data to individual treatment status and outcome measures, so that the researchers themselves do not need to handle and store the sensitive data. A description of this process may be found in the *Data Flow* section.

Researchers should separate *Personally Identifiable Information* (PII) from all other data as soon as possible. Data pose the most risk when sensitive or confidential information is linked directly to identifiable individuals. Once separated, the "identifiers" data set and the "analysis" data set should be stored separately (i.e., in separate folders or servers with different permissions), analyzed separately, and transmitted separately. Once separated, the identifiers should remain encrypted at all times, and the two data sets should only meet again if necessary to adjust the data matching technique. Tables 5, 6, and 7 illustrate this separation.

---

[10] For example, Bonnie Yankaskas, a professor of radiology at the University of North Carolina at Chapel Hill, experienced legal and professional consequences after the discovery of a security breach in a medical study she directed, though she was not aware of the breach and no damage to participants was identified. See the Chronicle of Higher Education article for more details, and a joint press release from the University and Professor Yankaskas describing the final result of the incident.

In order to maintain the ability to re-identify the analysis data set, a unique "Study ID" can be created by the researcher, data provider, or implementing partner. This ID should be created by a random process, such as a numbered list after sorting the data on a random number, or through a random number generator. This ID should *not* be:

- Based on any other characteristic of the data, such as numerical or alphabetic order, or a scrambling or encryption of Social Security numbers or other uniquely identifying codes

- An arbitrary mathematical function

- A cryptographic hash[11]

---

[11] For example, data from NYC taxi trips were released, with the drivers' hack license and medallion numbers obscured using a standard cryptographic hash. The data were de-anonymized within two hours. See Goodin 2014, Panduragan 2014, and Berlee 2015 for more information on this case. A cryptographic hash plus a secret key may be a more secure option, but the best is an entirely random number, unrelated to any identifiers.

TABLE 5: INITIAL DATASET

NAME	SSN	DOB	INCOME	STATE	DIABETIC?
Jane Doe	123-45-6789	5/1/50	$50,000	FL	Y
John Smith	987-65-4321	7/1/75	$43,000	FL	N
Bob Doe	888-67-1234	1/1/82	$65,000	GA	N
Adam Jones	333-22-1111	8/23/87	$43,000	FL	Y

TABLE 6: IDENTIFIERS DATASET

NAME	SSN	STUDY ID
Jane Doe	123-45-6789	1
John Smith	987-65-4321	2
Bob Doe	888-67-1234	3
Adam Jones	333-22-1111	4

TABLE 7: ANALYSIS DATASET

STUDY ID	INCOME	STATE	DIABETIC?
1	$50,000	FL	Y
2	$43,000	GA	N
3	$65,000	MN	N
4	$43,000	GA	Y

One method for creating Study IDs is:

1. Create a random number from a physical source (e.g., dice) or from a pseudo-random number generator (e.g., in Stata).

2. Use the first random number as a "seed," and use a pseudo-random number generator (e.g., in Stata) to sort the observations.

3. Use another random number as a second "seed," and use a pseudo-random number to generate a Study ID.

4. Ensure each Study ID is unique (e.g., using the *-isid-* command in Stata).

Depending on how the Study ID is created, it may be essential to maintain a secure crosswalk (i.e., mapping/decoding) between the Study ID and PII. This crosswalk should be guarded both to ensure confidentiality, and to insure against data loss.

Erasing Data

The IRB or data provider may dictate whether and when data must be retained or destroyed. Personally identifiable information linkages should be erased when they are no longer needed. Simply moving files to the "recycle bin" and emptying the bin is not sufficient to thoroughly erase sensitive data. There are several [software options](#) for removing all files. For example, MIT maintains [recommendations for removing sensitive data](#). Some IT departments may offer support for secure removal and disposal services.

The data provider will need to be confident that all files have been securely removed and no additional copies have been retained. Some researchers have documented the data erasure by taking screenshots of the removal process.

EXTERNAL RESOURCES

General Resources for Administrative Data

1. The Administrative Data Liaison Service supports administrative data research in the UK.

2. The "Cheaper, Faster, Better: Are State Administrative Data the Answer?" report provides an overview of working with administrative vital records and Medicaid data in the Mother and Infant Home Visiting Program Evaluation–Strong Start (MIHOPE-Strong Start) project. As such it provides an excellent overview of all steps in the administrative data process.

Resources for Data Classification and Data Security

1. Harvard University's Research Data Security Policy classifies data according to five levels of sensitivity and defines data security requirements that correspond to each sensitivity level.

2. US Department of Health & Human Services' Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

3. The National Institutes of Health's "How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?"

4. 45 CFR 164.514 – Describes the HIPAA standard for de-identification of protected health information (original text).

Resources for HIPAA

1. The US Department of Health & Human Services provides a detailed guide to the requirements associated with research identifiable health data and to how the HIPAA Privacy Rule applies to research, and a guide to understanding HIPAA for all types of users.

2. 45 CFR 164.502 – Uses and disclosures of protected health information (original text).

3. NIH guidance on complying with HIPAA, including de-identified health information, Authorizations, and Authorization waivers.

4. 45 CFR 164.514 – Describes the HIPAA standard for de-identification of protected health information (original text).

5. 45 CFR 160.103 – Defines Individually identifiable health information.

Resources for Informed Consent and Authorization

1. 45 CFR 164.502 – Uses and disclosures of protected health information (original text).

2. 45 CFR 164.508 – Uses and disclosures for which an authorization is required (original text). HIPAA regulations pertaining to authorizations for the release of health information, and requirements of the authorization.

3. NIH guidance on complying with HIPAA, including Authorization for research and waivers of Authorizations.

4. The US Department of Health & Human Services' guide to understanding the HIPAA Privacy Rule's relationship to research includes descriptions of the specific requirements of an Authorization for research.

5. [45 CFR 46.116](#) – Common Rule requirements for informed consent (original text).

   o Paragraph (a) describes the basic, required elements of informed consent;

   o (b) describes additional elements that may be included or required depending on the study;

   o (c) and (d) describe the conditions under which IRBs may waive or approve an alteration of informed consent.

6. The US Department of Health & Human Services maintains a [tips sheet](#).

7. MIT's Committee on the Use of Humans as Experimental Subjects (COUHES) provides [template forms and instructions](#) for obtaining informed consent and authorization.

8. J-PAL researchers can reference checklists and resources on Sharepoint.

9. IPA researchers can reference checklists and resources on Box.

10. The University of Colorado's IRB provides [examples and discussion](#) of waivers of informed consent.

Resources for IRB Procedures

1. The Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects (COUHES) provides examples of what various types of [application forms](#) might look like, and includes instructions.

2. COUHES also has an [investigator quick guide](#) overviewing when review is needed, and [guidelines](#) for various procedures in the IRB review process.

3. The University of Colorado has several [guidance documents](#) relating to all elements of the IRB process.

4. The US Department of Health & Human Services presents [decision charts](#) for understanding IRB requirements.

5. J-PAL researchers can reference checklists and resources on [Sharepoint](#).

6. Innovations for Poverty Action (IPA)'s IRB publishes FAQs and resources on their [website](#).

Resources for Data Sources

1. J-PAL North America is developing a [Catalog of Administrative Data Sets](#) for use in randomized evaluations.

2. Medicare and Medicaid data: The Research Data Assistance Center ([ResDAC](#)) at the University of Minnesota provides explanation and assistance with applying for access to data from the Centers for Medicare and Medicaid Services.

3. Credit reporting agencies: [Equifax](#), [Experian](#), [TransUnion](#)

4. US Federal data: The American Economic Association hosts resources enumerating [sources](#) and [procedures](#) for accessing US federal administrative data.

5. Education data: Researchers have compiled an inventory of data sets used to study [education](#).

Resources for Data Security

1. For researchers at MIT, Dr. Micah Altman, Director of Research at MIT Libraries regularly presents talks on [Managing Confidential Data](#).

2. MIT's Information Systems & Technology Department provide resources on:

- o [Protecting data](#)
- o [Sensitivity of data](#)
- o [Data risks](#)
- o Secure Shell File Transfer Protocol: [SecureFX](#)
- o [Encryption](#) (including software recommendations) and [whole-disk encryption](#)
- o [Removing sensitive data](#)

3. Harvard University's [Research Data Security Policy](#) (HRDSP) is an excellent resource for security level classification and security requirement examples.

Resources for Data Use Agreements

1. MIT has sample data use and nondisclosure agreements [here](#).

References

Berlee, Anna. 2015. "Using NYC Taxi Data to identify Muslim taxi drivers." *The Interdisciplinary Internet Institute* (blog). Accessed November 30, 2015. http://theiii.org/index.php/997/using-nyc-taxi-data-to-identify-muslim-taxi-drivers/

Dynarski, Susan, Steven Hemelt, and Joshua Hyman. 2013. "The Missing Manual: Using National Student Clearinghouse Data to Track Postsecondary Outcomes." w19552. Cambridge, MA: National Bureau of Economic Research. http://www.nber.org/papers/w19552.pdf.

Goodin, Dan. 2014. "Poorly anonymized logs reveal NYC cab drivers' detailed whereabouts." *Ars Technica*. Accessed November 30, 2015. http://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/

Finkelstein, Amy, and Sarah Taubman. 2015. "Using Randomized Evaluations to Improve the Efficiency of US Healthcare Delivery." http://www.povertyactionlab.org/publication/healthcaredelivery/reviewpaper.

Lee, Helen, Anne Warren, and Lakhpreet Gill. 2015. "Cheaper, Faster, Better: Are State Administrative Data the Answer?" OPRE Report 2015-09. The Mother and Infant Home Visiting Program Evaluation-Strong Start Second Annual Report. MDRC. http://www.mdrc.org/sites/default/files/MIHOPE-StrongStart-2yr_2015.pdf.

Ludwig, Jens, Greg J Duncan, Lisa A Gennetian, Lawrence F Katz, Ronald C Kessler, Jeffrey R Kling, and Lisa Sanbonmatsu. 2013. "Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity." *American Economic Review* 103 (3): 226–31. doi:10.1257/aer.103.3.226.

Meyer, Bruce, and Nikolas Mittag. 2015. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness and Holes in the Safety Net." w21676. Cambridge, MA: National Bureau of Economic Research. http://www.nber.org/papers/w21676.pdf.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-Anonymization of Large Sparse Datasets." presented at the Proceedings of 29th IEEE Symposium on Security and Privacy, Oakland, CA, May. http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf.

Pandurangan, Vijay. 2014. "On Taxis and Rainbows: Lessons from NYC's improperly anonymized taxi logs."
    Medium. Accessed November 30, 2015. https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1

Tahamont, Sarah, Shi Yan, Ryang Kim, Srivatsa Kothapally, and Leslie Kellam. 2015. "The Consequences of
    Demographic Matching of Administrative Data." Poster presented at the APPAM Fall 2015 Conference,
    Miami, FL, November 12.

Taubman, S. L., H. L. Allen, B. J. Wright, K. Baicker, and A. N. Finkelstein. 2014. "Medicaid Increases Emergency-
    Department Use: Evidence from Oregon's Health Insurance Experiment." *Science* 343 (6168): 263–68.
    doi:10.1126/science.1246183.

APPENDICES

CONSENT AND AUTHORIZATION

Consent and authorization are related topics, and both are reviewed by Institutional Review Boards or Privacy Boards. While authorization may be obtained during the informed-consent process, consent and authorization are not substitutes.

Informed Consent

Informed consent is a process by which research subjects are informed of the research procedures, goals, risks, and benefits, and consent to participate voluntarily. The Common Rule (45 CFR 46.116) and institutional policy dictate certain elements of informed consent. The US Department of Health & Human Services also maintains a tips sheet for informed consent.

Only individuals who are legally adults may give consent. While parents may give consent for their children, the IRB may determine that researchers obtain the assent of the children to participate in the research, in addition to the consent of their parents.

Data providers may want to review the study's informed consent form to ensure it accurately describes to prospective study participants what data from their organization will be shared and with whom, when the data will be released, and how it will be protected. The organization's review may result in changes to the study's informed consent form, which must be approved by an IRB. Researchers who request administrative data for a previously consented study sample may be required to re-obtain for each participant before a data provider will agree to release his/her administrative records (Lee et al. 2015).

For more information, see *Resources for Informed Consent and Authorization*.

Authorization for Research (HIPAA)

An authorization is a signed record of an individual's permission to allow a HIPAA *Covered Entity* to use or disclose their *Protected Health Information (PHI)*. The authorization must describe the information requested and the purpose, and must be written in plain language that is readily understood by the individual. This is similar to the concept of *Informed Consent*, and is often embedded within an informed-consent document. However, an authorization has a distinct set of criteria, and may be a separate written agreement obtained outside of the informed-consent process.

For more information, see *Resources for Informed Consent and Authorization*.

DATA SECURITY

Data storage and access

Researchers have many options for secure data storage and access. Relevant considerations for choosing among these include the sensitivity of the data, applicable compliance requirements, the research team's technical expertise, internet connectivity, and access to IT expertise and support.

Encryption
Encryption is the conversion of data to code that requires a password or pair of "keys" to decode. Data may be encrypted at many levels, at multiple stages of the data lifecycle, and through a variety of software and hardware packages.

More information on encryption and software recommendations from MIT are available [here](#).

Device-level (whole-disk) encryption. Computers, flash drives, tablets, mobile phones, and any other hardware for data storage may be [whole-disk encrypted](#). This method protects all files on the device, and requires a password upon device start-up. After installation and implementation, whole disk encryption should not materially affect the user experience. Methods and software available for whole-disk encryption vary by hardware type. Researchers are advised to contact their institution's IT department for advice or assistance.

File-level encryption. While whole-disk or device-level encryption encrypts all files *on a device*, it does not protect the files once they leave the device--for example, while they are in transit or being shared with another researcher. File-level encryption applies to specific files, and facilitates data sharing. Proper use of file-level encryption requires strong protocols for password sharing and for unlocking and relocking files before and after use.

Some options for file-level encryption include:

- [PGP-Zip](#)[12]
- [Boxcryptor](#)[13]
- 7-zip
- TrueCrypt

Note that while cloud storage tools such as Dropbox and Google Drive have built-in encryption, Dropbox, Google and their employees have access to the encryption keys--and therefore have access to all data and files stored on their servers.

---

[12] This is MIT's current (as of August 2015) recommendation for file-level encryption.

[13] This is Innovations for Poverty Action's current (as of August 2015) recommendation for file-level encryption.

Cloud storage

Both [Box](#) and [Dropbox](#) (and potentially other cloud storage providers) are willing to sign HIPAA-compliant "business associate agreements" and have configured their platforms to comply with HIPAA regulations--but without this type of agreement or other file-level encryption, simply storing files using these services is not a fully secure option for sensitive data.[14] The original data provider and any reviewing IRBs should be consulted prior to initiating agreements with cloud-storage providers.

IT-administered Options

For researchers with access to a professional IT team, and whose team members have access to reliable, fast internet, IT-administered options may be preferable. These options allow researchers to delegate the administration of a data access and storage solution to IT experts. IT administrators may also be able to provide several additional levels of data protection. As with cloud storage, IT staff may have access to all data on a server, including PII. Researchers should be sure to understand who has access to the data, and maintain as much direct control as possible to prevent compliance issues or accidental data breaches.

Institutions may offer space on a server or provide a location to host a server. Storing data on such a server may be preferable to relying on laptops or desktops and cloud storage to maintain data. Depending on the institution, the IT department may be able to provide secure remote access for off-campus users, automated secure backups of data, and encryption.

Access to these servers is typically automatic when connected over an official institutional internet connection. Off-site access requires the use of a Virtual Private Network (VPN). This may provide additional layers of security by requiring two or more types of authentication (for example, a password and a code sent via text message).

Additional features that may be available upon request include:

- Inactivity timeouts for remote access

- Nonretrievable passwords. If a user forgets his or her password, the password is reset by the system, rather than the original password being returned.

- Password expiration settings that require a new password be created on a regular basis.

- Restriction on the number of password guesses permitted before account lockout.

- Access logs that describe who signed in, from where, and when.

IT or data managers may be able to grant access permissions to specific users for specific files or folders on the server. This level of control would enable teams to share general access to a folder while limiting access to identified data to a specific subset of the team.

---

[14] Dropbox also complies with a number of other international [data regulations,](#) as detailed by their help center.

Data Transmission and Sharing

Data must be protected both when at rest and in transit between the data provider, research team members, and partners. Data that are encrypted while at rest on a whole-disk encrypted laptop, or on a secure server, will not necessarily be protected while being transmitted. The options presented below may vary in their level of security.

Unsafe transmission methods

- Email without encryption

- Uploading unencrypted data to Dropbox or Box (no matter how quickly the data are deleted afterwards). See *Cloud storage*

- Mailing unencrypted media devices (e.g., CDs, USB memory sticks, flash drives, external hard drives)

- Password-protected Excel file

Safer transmission methods

- Secure Shell File Transfer Protocol (SFTP). MIT provides support for [SecureFX](#)

- Uploading an encrypted file to Dropbox or Box

- Emailing an encrypted file, the password for which is shared separately and securely

- Mailing encrypted files loaded onto encrypted devices

Personal Device Security

There are several simple steps researchers and their staff can take to ensure their machines remain secure and to minimize possible weak points.

- Use a password-locked screensaver or timeout lock.

- Install and maintain antivirus software. Keep this software up to date, and allow it to perform regular checks.

- Use a firewall. Most operating systems (Windows XP, Vista, Windows 7, Mac OS X, and Linux) have built-in firewalls.

- Keep all software up to date. Most computers and platforms regularly check for new versions of software. New versions are often created to fix security problems or other known issues.

- Don't install or run programs from untrusted sources.

IT departments generally have recommended software to help secure personal devices and may be able to assist with updating this software or may push automatic updates.

Password Policies

Strong passwords are essential to ensuring data security. A different password should be used for each high-value account. For example, the passwords for Dropbox, email, institutional servers, and encryption files should all be different. Passwords should also be changed regularly, even if not required by the account itself.

Strong passwords should:

- Be at least six characters, but preferably much longer

- Include a mixture of uppercase and lowercase characters, numbers, and symbols
- *NOT* contain
    - Dictionary words in any language, even with a varied capitalization scheme or with numbers or symbols substituted for letters (e.g., 1 for l, @ for a, 0 for O)
    - Names
    - Your username, user ID, email address, phone number, etc. (forwards or backwards).
    - A sequence of four or more numbers

Do not forget your password. Strong passwords may be difficult to remember. When using some encryption software, such as Boxcryptor, a forgotten password is completely irretrievable and means the loss of all project data.

An unencrypted, password-protected Excel file of passwords is *not* a secure way to store or share passwords. There are a variety of password storage systems; [LastPass](#) offers a secure way to create, manage, remember, and share passwords. A hard copy of a password list, locked in a safe, is another secure option.

Passwords should never be shared using the same mechanism as file transfer.

Preventing Data Loss

In addition to securing against outside threats, preventing data loss is an essential component of data security. Data and crosswalks between study IDs and personally identifiable information should be backed up regularly in at least two separate locations, and passwords must not be forgotten.

Cloud-based storage providers (e.g., Box, Dropbox, Google Drive) offer a range of options for data backups and may offer additional packages to back up data for longer periods of time to protect against the unintentional erasure of data. Institutional servers may also have data backup plans, and device-level backup plans are also available. Backing up data to an external hard drive (stored in a *separate location* from daily computers) is an option for low-connectivity environments.

DEFINITIONS

Personally Identifiable Information (PII)

PII is any piece of information or combination of information that can be used to identify a particular individual with a reasonable amount of certainty.

Examples:

- A Social Security number on its own is PII.

- An age, gender, and location combination may or may not be PII, depending on the age and size of the geographic area. "A 35-year-old man in Boston, MA" is not PII, but "A woman in her 90s in Tanana, AK" is PII.

Protected Health Information (PHI)

The HIPAA Privacy Rule protects "individually identifiable health information," and such information that the Privacy Rule protects is termed Protected Health Information (PHI). Individually identifiable health information is information that relates to an individual's health, health care, or health-care payments, created or held by a _Covered Entity_ that is identifiable, whether directly or indirectly.

Individually identifiable health information is defined in 45 CFR 160.103.

Covered Entity

A covered entity must comply with HIPAA, and with the HIPAA Privacy Rule's requirements to protect the privacy and security of health information. Covered entities are defined in 45 CFR 160.103.

University researchers are generally not considered covered entities, but in the course of health research they will often interact with covered entities in order to obtain data.