J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
NORTH AMERICA

DATA SECURITY PROCEDURES FOR RESEARCHERS

Elisabeth O'Toole, Laura Feeney, Kenya Heard, Rohit Naimpally

J-PAL North America, August 2018

Summary: This document provides a primer on basic data security themes, provides context on elements of data security that are particularly relevant for randomized evaluations using individual-level administrative and/or survey data, and offers guidance for describing data security procedures to an Institutional Review Board (IRB) or in an application for data use. This is not an exhaustive or definitive guide to data security and does not replace the guidance of professional data security experts nor supersede any external data security requirements.

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

Acknowledgements: This document is an updated excerpt from "Using Administrative Data for Randomized Evaluations" (J-PAL NA, 2015). We are grateful to the original contributors to that work, as well as Patrick McNeal, and James Turitto for their insightful feedback and advice. Chloe Lesieur copyedited this document and Laurie Messenger formatted the guide, tables, and figures. This work was made possible by support from the Alfred P. Sloan Foundation and the Laura and John Arnold Foundation. Any errors are our own.

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.

TABLE OF CONTENTS

DATA SECURITY OVERVIEW

Data security is critical to protecting confidential 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.[1] Additionally, the information revealed in this process could be damaging in unexpected ways. For example, computer scientist Arvind Narayanan successfully re-identified a public-use de-identified data set from Netflix. Through this, he was able to infer viewers' 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. Researchers should also ensure that 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 illustrate how this calibration may function in practice.[2]

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, which may expose them to identity theft, embarrassment, and financial, emotional, or other 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. A breach will likely trigger additional compliance requirements, including reporting the data breach to the Institutional Review Board (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.[3]

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, device loss, file-sharing software such as Google Drive, Box or Dropbox, or improper erasure of files from hardware that has been recycled, donated or disposed of. All hardware that comes into contact with study data should remain protected including: laptops, desktops, external hard drives, USB flash drives, mobile phones, and tablets. Theft or a cyber-attack may target either a researcher's

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

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

[3] 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.

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.

Reduce the data security threat-level a priori by acquiring and handling 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 of J-PAL NA's resource: Using Administrative Data for Randomized Evaluations.

DEIDENTIFYING DATA

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, analyzed separately, and transmitted separately.[4] 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 1, 2, and 3 illustrate this separation. J-PAL hosts programs for searching for PII in Stata and in R on a GitHub repository.

TABLE 1: 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 2: 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

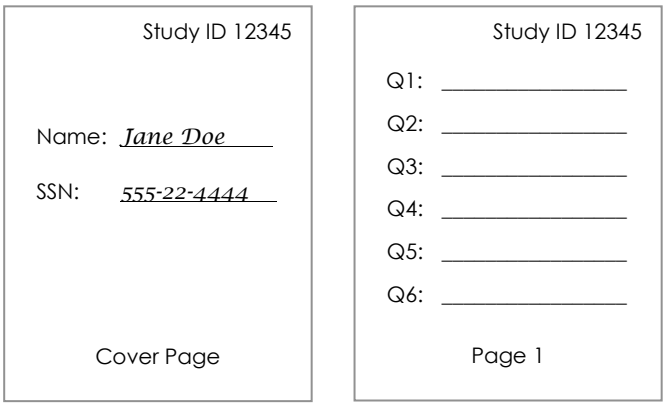
TABLE 3: ANALYSIS DATASET

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

[4] Separating & encrypting identifiers is a minimum requirement in J-PAL's Research Protocol Checklist

Paper-based surveys should also be designed in a way such that PII is removable. Refer to Figure 1 for a mock-up of survey design. All direct identifiers such as name or Social Security Number and contact information such as address or telephone number should appear on a separate cover sheet. The cover sheet and any consent form should be separated from the main questionnaire as soon as possible – ideally within 24 hours. As described below, participants will be assigned study IDs (listed on both sheets) so that the research team can match and re-identify the data if needed. A crosswalk document will contain the study ID link between these two sections and will be stored in a separate location from both survey halves as to ensure confidentiality.

FIGURE 1



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[5]

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

[5] 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.

J-PAL's randomization exercise in Stata includes the creation of Study IDs using this process.

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. Innovations for Poverty Action (IPA) has a publicly-available Stata program on GitHub that automates the separation of PII from other data and the creation of a crosswalk.

DATA STORAGE & ACCESS

Researchers have many options for secure data storage and access. Relevant considerations for choosing among these options 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, and is a requirement for all J-PAL-implemented projects[6]. 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 and/or primary data collection may be whole-disk encrypted. This method protects all files on the device, and requires a password upon device start-up. For tablets used in primary data collection, an application such as AppLock can be used to prevent users from accessing other applications during data collection. This protects data from being transferred across applications. The research team should also enable remote wiping capabilities on these devices in case of loss or theft.[6] 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.

CLOUD STORAGE. Many cloud storage providers including Dropbox, Box, and Google Drive have configured their platforms to comply with various industry, federal, and international regulations to keep files secure on the cloud (while data will still need to remain encrypted at all endpoints). Some of these services allow users to "upgrade" to a specific level or type of data security compliant with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) or the Family Educational Rights and Privacy Act (FERPA). Without a formal agreement to store data in compliance with a specific set of regulations or other file-level encryption, simply storing files using these services is not a fully secure option for sensitive data. The original data provider and any reviewing IRBs should be consulted prior to initiating agreements with cloud-storage providers.

[6] Refer to Google's instructions on how to find, lock, or erase a lost Android device. Individual device manufacturers (such as Samsung) may have their own procedure that you could separately enable. Generally, steps must be taken when initially setting up each study tablet, such as creating accounts for individual devices, and should be approved by IRB.

For more information on cloud storage services & security:

- Box (and HIPAA Specific Overview)
- Dropbox (and HIPAA Specific Overview)
- Google Drive (and HIPAA Specific Overview)

FOLDER-LEVEL ENCRYPTION. While cloud storage tools encrypt the connection and files "at rest" on their systems, they retain the encryption keys, which technically gives their employees read access to all files saved on their servers. To address this, tools including Boxcryptor[7] and VeraCrypt[8] encrypt files before they are stored in the cloud. Boxcryptor is a paid subscription model, whereas VeraCrypt is free and open-source.

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. Options for file-level encryption include PGP-Zip[9] and 7-zip.

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 encrypting all network connection and requiring two or more types of authentication (e.g., a password and a code sent via text message). Data will still need to be encrypted at both endpoints – i.e., the server or files on the server must be encrypted, and any data transferred to or from the server to another server or hard drive must be encrypted at those points.

[7] This is Innovations for Poverty Action's recommendation (as of November 2015) for file-level encryption, per IPA's Best Practices for Data and Code Management.

[8] J-PAL recommends VeraCrypt, and has updated the Truecrypt Stata command to work with VeraCrypt. J-PAL also developed a guide to installing and using VeraCrypt software.

[9] PGP-Zip is MIT's current (as of May 2018) recommendation for file-level encryption.

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. Seek out your IT department's official recommendations regarding passwords and permission, such as IS&T Policies for MIT projects.

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 include:

- 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 include:

- Secure Shell File Transfer Protocol (SFTP), including Secure Shell (SSH) or Secure Copy (SCP). MIT provides SFTP support for SecureFX.
- Uploading an encrypted file to Dropbox or Box
- Emailing an encrypted file, and sharing the password separately and securely
- Mailing encrypted files loaded onto encrypted devices
- Survey software with encryption features, such as SurveyCTO, that supports encryption during data collection and transmission to a central server

COMMUNICATION AND DATA SHARING WITH PARTNERS

Many research partners, such as service providers, survey enumerators, and holders of administrative data, have had minimal prior exposure to data security or data sharing protocols. It is best practice to develop a data sharing and security protocol with these partners, and to guide them in understanding their role in data security. All partners handling or transmitting data should be informed of and trained on data collection, storage, and transfer policies agreed upon for the study. Request that partners notify the research team before sharing any data to ensure compliance with the data protocol. Teams should communicate with each other and with partners by referencing Study ID numbers rather than using PII. Consider developing standard operating procedures for checking for and responding to breaches in following the agreed upon method for sharing data. For example, if partners share data in a non-secure way or if unauthorized data are disclosed to researchers or partners. This will allow staff to respond quickly in the event of a breach. A standard operating procedures document should include:

1. Process for sharing data and receiving updates
2. Process for verifying data set does not contain unauthorized information prior to downloading, if possible
3. Timeline for reviewing new data for unauthorized information or PII
4. Plan for notifying the source of the breach and requesting corrective action to prevent future breaches
5. How files with unauthorized information will be removed and destroyed

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. These steps include:

- Use a password-locked screensaver and timeout lock.
- Install and maintain antivirus software. MIT currently recommends Sophos; other institutions may support or recommend alternatives. Keep this software up to date, and allow it to perform regular checks.
- Use a firewall. Most operating systems (including Windows 10, macOS, 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 encrypted files should all be different.

The National Institute of Standards and Technology (NIST) published revised guidelines for passwords in 2017. These guidelines and the underlying rationale are explained in more approachable language in a NIST staff blog post.

In general, strong passwords should:

- Be at least eight characters, but preferably much longer
- *NOT* contain or solely comprise:
    - 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)
    - The name of the service or related words
    - Your name, username, email address, phone number, etc. (forwards or backwards)
    - Repetitive or sequential letters or numbers

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

Store and share passwords securely. An unencrypted, password-protected Excel file of passwords is *not* a secure way to store or share passwords. Passwords should never be shared using the same mechanism as file transfer, nor should they be shared over the phone.

Password storage systems such as LastPass offer a secure way to create, manage, and store passwords online. On this webpage and mobile application, notes and passwords also can be securely shared with specified teammates. A hard copy of a password list, locked in a safe, is another secure option.

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 PII should be backed up regularly in at least two separate locations, and passwords must not be forgotten.

Cloud-based backup tools such as CrashPlan and Carbonite 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. Cloud-based storage tools such as Box, Dropbox, and Google Drive offer packages to back up data for several months or more, and may insure against unintentional erasure of data if it is noticed within the backup time period; these storage tools are not true backup tools as they do not keep deleted files forever. Institutional servers may also have data backup plans, and device-level backup plans are also available. Backing up data to an encrypted external hard drive (stored in a *separate location* from daily computers) is an option for low-connectivity environments.

ERASING DATA

The IRB or data provider may dictate whether and when data must be retained or destroyed. PII 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. For paper-based surveys, J-PAL recommends that hard copies of PII cover sheets,

questionnaires, and study ID crosswalks be destroyed within 3-5 years of the end of a project (or as committed on IRB protocol).[10]

The data provider will need to be confident that all files have been securely removed and no additional copies have been retained. In order to document data erasure, some researchers have taken screenshots of the removal process.

EXAMPLE LANGUAGE FOR DESCRIBING A DATA SECURITY PLAN

Data Use Agreements (DUAs) and IRBs often require researchers to provide a description of their data security and destruction procedures, and some may include specific requirements on these processes dependent on the sensitivity of requested data.[11] This section provides examples of descriptions of data management plans drawn from approved DUAs. This language is provided for informational purposes only; it is not necessarily comprehensive nor feasible in all environments. Please refer to your university's (and/or department's) research support center, libraries, or IT department for detailed protocols, potential templates, and descriptions of what is feasible, required, and sufficient at your location. Additional external resources on describing data security plans are in this section: *Resources for Data Security*.

EXAMPLE LANGUAGE: SECURE DATA STORAGE

> The Department maintains a Unix/Linux-based research computing environment for its students and faculty members. The research computing systems utilize enterprise-level hardware and are managed by a dedicated staff of IT professionals. The Department leverages additional resources provided by the institution centrally, such as network infrastructure and professional co-location services in institutional datacenters. Department IT staff fully support private research servers purchased by individual faculty members. This support includes account management, security patching, software installation and host monitoring. Secure servers will be utilized for the purposes of processing and analyzing data. All computations and analytical work will be performed exclusively on these servers. File based permissions will be set to restrict data access to the research team.
>
> All project data will be stored on a network attached storage (NAS) device. A dedicated volume will be created on the NAS for exclusive storage of all data related to this research project. Data on this volume will be served using the NFSv4 protocol and restricted to authorized hosts and users using IP-based host lists and institutional credentials. Data on this volume will be accessible only to authenticated users on the project servers described below. Data is backed-up to a secondary NAS device which is accessible only by IT personnel.
>
> All network traffic is encrypted using the SSH2 protocol. A VPN provides an additional level of encryption/ access restriction for off-campus connections. All server logins require two forms of authentication, a password and an SSH key pair. SSH Inactivity Timeout is used as the session timeout protocol.

[10] J-PAL Research Protocol Checklist

[11] In addition to these elements, 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.

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.

HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT (HIPAA)

HIPAA provides regulation for healthcare data security, holding health care providers, insurance providers, researchers and others accountable for safeguarding protected health information (PHI) in the United States. Compliance requirements differ based on the party, such as individuals, covered entities, or researchers; the purpose of the data usage; and on stipulations or structure of data use agreements.

Resources:

- HIPAA Privacy Rule specific to research use
- De-identification of Protected Health Information in compliance with HIPAA
- Definition of a covered entity under HIPAA
- Other privacy & security resources, and security risk assessment tools
- J-PAL NA's Administrative Data Catalog: Compliance section (pg. 11-17) for information on HIPAA-compliant de-identification considerations, and other HIPAA data requirements.

FAMILY EDUCATIONAL RIGHTS AND PRIVACY ACT (FERPA)

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. FERPA may have implications for how researchers conduct evaluations and report results, in particular, as related to obtaining individual consent from study participants.

Resources:

- FERPA Resources for researchers
- FERPA-Recommended best practices for data security, such as a Data Breach Response Checklist, and Best Practices for Data Destruction
- Using Financial Aid Information for Program Evaluation and Research

EXTERNAL RESOURCES

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 DATA SECURITY

1. J-PAL's resources on working with data, which includes additional resources for VeraCrypt
2. For researchers at MIT, Dr. Micah Altman, Director of Research at MIT Libraries regularly presents talks on Managing Confidential Data.
3. MIT's Information Systems & Technology Department provides resources on:
    - Protecting data
    - Data risks
    - Secure Shell File Transfer Protocol: SecureFX
    - Encryption (including software recommendations) and whole-disk encryption
    - Removing sensitive data
    - Password protection
    - Virus protection software: Sophos
4. Harvard University's Research Data Security Policy (HRDSP) is an excellent resource for security level classification and security requirement examples.
5. J-PAL's Research Protocol Checklist
6. IPA's Best Practices for Data and Code Management
7. J-PAL and IPA provide sample code to:
    - Scan for PII data in Stata and in R
    - Separate PII from other data in Stata
    - Generate random Study IDs in Stata
    - Use VeraCrypt with Stata
8. The National Institute of Standards and Technology's (NIST) paper on De-Identification of Personal Information, and explained in their presentation on Data De-Identification.

9. The National Institute of Standards and Technology's (NIST) revised guidelines for passwords, explained in more approachable language in a NIST staff blog post.
10. Several institutions provide guidance on developing data security plans, and describing the plans for grant proposals or data use agreements. Resources include:
    - Inter-university Consortium for Political and Social Research's (ICPSR) Framework for Creating a Data Management Plan and Guidelines for Effective Data Management Plans.
    - MIT Libraries' guide to Writing a Data Management Plan.
    - NC State University Libraries' Data Management Plan Examples.
    - Rice Research Data Team's resource for Developing a Data Management Plan.
    - UNC Carolina Population Center's tools and resources on Security Plans for Restricted-Use Data.
    - University of California Curation Center's Data Management Plan Tool (DMPTool) enables users to organize data management plans according to templates, for example, to adhere to funding requirements. This resource is subscription-based. Please refer to the list of DMP Participants to see if your university or institution already enables an institutional sign-in.

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/.

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/.

Mangan, Katherine. 2010. "Chapel Hill Researcher Fights Demotion after Security Breach." *Chronicle of Higher Education.* Accessed March 7, 2018. https://www.chronicle.com/article/chapel-hill-researcher-fights/124821/.

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/.