Data Security Guidelines for Community-Based Research

A Best Practices Document Prepared by the
Ad-Hoc Committee for Data Security
Program for Global Disease Epidemiology and Control
Department of International Health

Members:
Alain Labrique (Chair)
Allan Massie
Andre Hackman
Christian Coles
Fred Van Dyk
Lee Wu
Luke Mullany
Mathilee Mitra

External Contributions
Joan Petit
Jonathan Links

Date: May 8, 2011 v5.0

Document Outline

Appendices

Summary: Data Security Review and Guidelines

The following document presents the summary discussions, findings and recommendations of an ad hoc committee formed by faculty and data-management staff of the Global Disease Epidemiology and Control (GDEC) and Human Nutrition (HN) Programs of the Johns Hopkins Bloomberg School of Public Health Department of International Health. The committee's charge was to review IRB recommendations, Information Systems recommendations, standard data practices across a number of overseas project sites, and identify 'risk' scenarios involving data with the aim of providing reasonable, "best practice" guidelines for the secure storage and transaction of research data. These guidelines are meant to provide tangible risk-minimization strategies for a broad set of stakeholders, from student researchers to project Principal Investigators. Extensive and detailed discussions with leadership of the JHSPH IRB was instrumental in the construction and content of this document.

I. Definitions and Basic Principles
Data Management

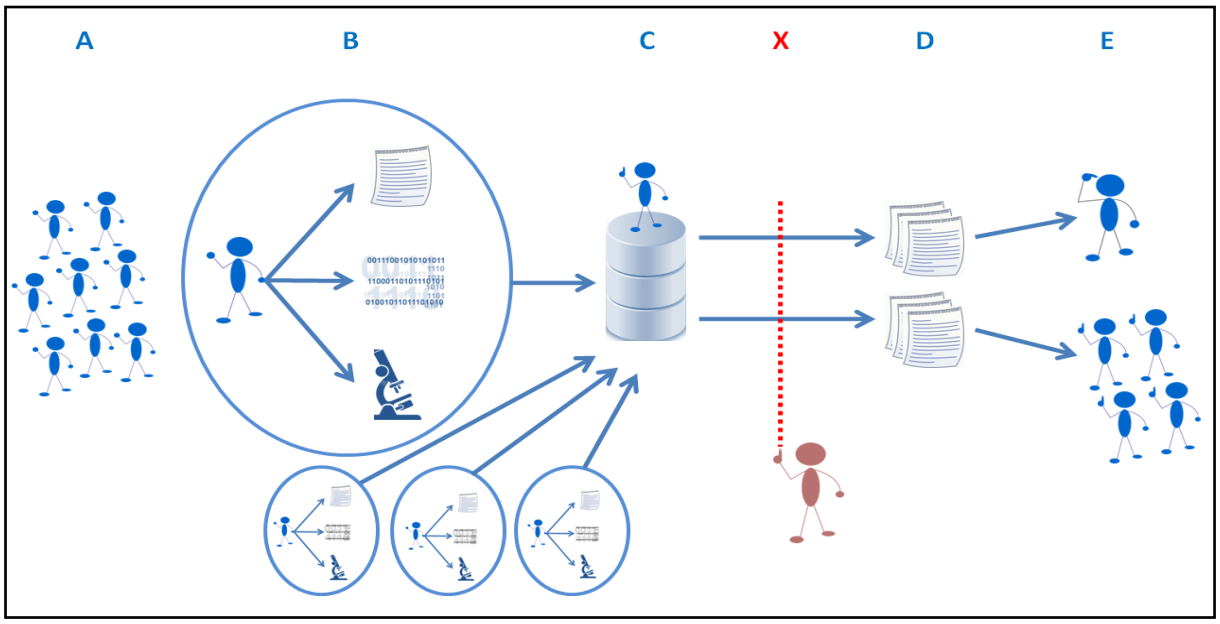
Data management of human subjects research data includes: data collection, data entry, and database repository oversight (controlling access, tracking use of analytic datasets). Theoretically, data collected for research may be used indefinitely, extending the obligation to protect that data, especially as long as the subjects remain identifiable as individuals or as a group. Good data security planning requires establishing controls and guidelines to govern access, use, and protection for all users.

Definitions

- Database: Generally refers to the "study database", e.g., repository of all study data, usually the repository of data as it is entered from original sources such as paper forms. Database access is usually limited to a small number of technical administrators, data entry personnel (often limited to write-only access) and less frequently, study investigators. Databases generally contain identifying information about participants, if it has been collected and subsequently entered. Databases may be simple, with all data residing in a single tabular format, or complex, with multiple tables containing different participant data, linked together by a common identifier (also referred to as a "relational database").

- Analytic dataset: A subset of the study database created by a data administrator or data gatekeeper in accordance to the data abstraction needs of the user. Analytic datasets may contain identifying information, only a unique identifier, or no identifier whatsoever.

- Data Gatekeeper: An individual who is authorized by the study PI to manage the issuance of analytic datasets, according to an explicit data security plan. The gatekeeper will be responsible to de-identify analytic datasets, to the maximum extent possible, as described below. The gatekeeper should also be tasked with logging the release and contents of analytic datasets, and request the return or deletion of data as defined in the data use agreement.



Some of these core principles are illustrated in the figure, above, with the following components: A, the population from which participants are recruited; B, the participants who contribute data / biospecimens; C, data that is entered and consolidated into a research database, managed by database administrators; D, and parsed into discreet, maximally de-identified analytic datasets, to be used by; E, various levels of end-users ranging from investigators, students, data analysts, quality control staff, course instructors, etc. The position of the data "gatekeeper" is illustrated as "X", serving as a management and control point between the database and the analytic datasets.

- Unique Nonsense Identifier: A non-sense random number that is assigned to a study participant upon enrollment into a research protocol. The number should not contain components that can be deciphered as a household number, cluster number or other geo-coded information.

- Unique Logical Identifier: A non-random number that is assigned to a study participant upon enrollment into a research protocol. The number is an aggregate of a series of numbers describing where the participant was enrolled in time and/or space. Eg. 11020223 – refers to household '1102' in cluster '022', where the participant is the 3rd member of the household census.

- Data identifiability: Data are "identifiable" to a specific investigator if **that** investigator is able to ascertain the identity of the participant. Use of a "unique nonsense identifier," and limiting access to the personal identifiers[1] through selective release of data by a

[1] * *"Personal Identifiers" include any data points which, when considered alone or in concert with other information, could identify an individual participant. Obvious identifiers include name, address, or telephone number. Less obvious data points could be considered as "identifiers", varying from case to case. For example, a geographic locater, such as the name of a village, plus the age and pregnancy status of an individual might be enough to permit association of coded data with its source individual. Each investigator must consider whether it is possible to identify an individual participant by using the data in the dataset in association with other readily available resources, to make a determination as to whether a dataset is sufficiently de-identified to protect the participant.*

designated “data gatekeeper,” add protection to participants. Investigators who can see, have access to, record, or use personal identifiers that go beyond a “unique nonsense identifier”, including a “Unique Logical Identifier" will be viewed as “using identifiable data.” Most researchers do not need personal identifiers for their work. Ideally, the study data management plan will maximize participant protection without unnecessarily hampering data utility. Access to “identifiable data” for a particular investigator will depend upon an investigator’s role in the study, and the investigator’s “need to know” those identifiers.

For example:

- Data collectors who directly obtain information from individual participants “know” the participants because of that direct interaction. The data will be considered “identifiable” to them even if identifiers are not recorded.
- If data are collected with personal identifiers, but those identifiers are recorded separately from the data with a unique study number assigned as a code linking the identifiers to the data (“Unique Nonsense Identifier”), the data are “coded.” Investigators who obtain the data with only the Unique Nonsense Identifier, and who have no access to or possession of the code linking the Unique Nonsense Identifier with the participant’s personal information, are most likely using a “de-identified” data set.
- Individuals who have access to, or possession of, the code to link data to identifiers beyond a study Unique Nonsense Identifier for research purposes are using “identifiable data.”
- A gatekeeper could hold the code and may create and distribute analytic datasets without identifiers, or with minimal identifiers, depending upon the needs of a particular analysis. The dataset recipients will only be responsible for protecting the identifiers provided to them by the gatekeeper. The data management plan should then state that only the PI and the gatekeeper will have access to the linkage between the Unique Nonsense Identifier and other personally identifying information.

- Data transport: The level of security required for physical transport of data will depend upon the identifiability of the data. Transport of data without identifiers, with only Unique Nonsense Identifiers and without access or link to extended personal identifiers, will require standard data transport precautions, but not enhanced security precautions.
- Data use agreement (DUA): Contractual agreement executed by the PI with the data user binding the user to terms and conditions governing the use of the analytic dataset, including security requirements, confidentiality, intellectual property/attribution,

restrictions on use, length of time use permitted, end of use requirements (return/destruction), etc. This agreement can be used to govern or manage sharing of data with external collaborators.

- Data Security Plan: Should provide detailed guidelines and decision rules for data administrators/gatekeepers to facilitate data sharing. Data should be treated like an asset, similar to project inventory. A data security plan includes provisions for defining (and restricting) levels of 'access' to types of data (e.g. identifiers, sensitive data, lab results, etc.) either by specific user or by 'class' of user. The plan should include a method for recording (tracking/logging) these individual permissions to access, access itself, and end of access. Each data sharing may be treated as a separate "transaction" and can have its own risk profile (see below). For higher risk transactions, such as data that contains identifiers or with external collaborators, data sharing may require more protection like data use agreements, time limited access to data, and possible review of proposed use by a data use committee created by the study leadership and authorized to screen such proposals.

Risk Assessment for Data Use/Sharing

Risk of harm to participants due to a breach of confidentiality will vary, depending upon the sensitivity of the study information. Risk is never "zero" because the fact that a person has joined a study at all is no business of anyone outside the study, but it could be minimal, for example, if the information collected concerns a participant's food preferences. "Sensitive" information is that which, if disclosed, could expose a participant to harm, whether physical, social, economic, psychological, emotional, or legal. The chart below attempts to describe a continuum of risk associated with sharing analytic datasets with other investigators or students.

Database/dataset Identifiability: Highest risk (I) to lowest risk (VI) for study participants

Data Level I: Individual w/ personal identifiers

Data Level II: Individual w/out personal identifiers, but with a unique study identifier – a "Gatekeeper" holds code

Data Level III: Individual w/out personal identifiers, but with potentially identifying meta-data (GIS, distinguishing features, etc.)

Data Level IV: Individual, no identifiers, no code

Data Level V: Community[2], aggregate and identified by community

Data Level VI: Community, aggregate and not identified by "community"

[2] *"Community" may be an "identity" in itself. The Havasupai tribe case illustrates this point. Investigators must consider the risk of harm to the community through future data use. This protection may be achieved by reviewing the original consent document and determining whether the proposed future use of a dataset is consistent with its objectives.*

Sharing Analytic Datasets: Lowest security (I) to highest security risk (V)

User Level I: Co-investigators, same institution

User Level II: Other faculty w/in Hopkins (JHSPH, JHMI, JHU, Jhpiego)

User Level III: Co-investigators, outside institution (sharing requires electronic transport outside firewalls)

User Level IV: JHSPH, JHMI, JHU Students

User Level V: Colleagues (not co-investigators) outside of Hopkins

Data Sharing	User Level V	User Level IV	User Level III	User Level II	User Level I
Data Level I	Highest Risk				
Data Level II					
Data Level III					
Data Level IV					
Data Level V					
Data Level VI					Lowest Risk

II. General Recommendations: There are many aspects of computer security, ranging from intellectual property protection and the secure backup of information to safeguarding against viruses and/or unauthorized users. In this report, we focus specifically on data security issues focusing on guarding the confidentiality of research subjects. Personally identifying information is regularly collected and stored as part of the day to day operations of many of our studies. This document focuses on the handling and safeguarding of human research subject data.

As discussed above, the committee acknowledges that there may be a "continuum of consequences" and costs associated with a security breach involving different kinds of data, based on the level of sensitivity of the concerned datasets. These recommendations acknowledge this continuum and provide guidance for situations when identifiable data must be transported, transferred or utilized outside the security of the JHMI networks. There are many ways in which these measures could be implemented, considering that increased data security measures will have associated costs (monetary, investigator time, analyst time, technical staff time).

I. Extant Guidelines:

A. **Be familiar with Institutional Requirements**: As a foundation to understanding this document, we propose that each person involved in data management read and become familiar with the JHSPH policies relating to data security. This information is posted on the JHSPH Information Systems Data Security site, available at this link: https://my.jhsph.edu/offices/informationsystems/datasecurity

   This site contains pertinent policy information for all members of the Institution as well as a checklist and tips to prevent the inadvertent disclosure of confidential data, or data leakage. The following sections of these documents are especially important:
   -Descriptions of various user roles and responsibilities.
   -Understanding how data is classified into different risk groups.

B. **Protect Data in Transit**: As described in these documents, there are a number of safeguards that are in place while data is stored on a fixed-site computer, operating within a Johns Hopkins Information-Systems (IS) secured network. **When data leaves this protected zone on a laptop, USB device, PDA or other mobile device - a scenario frequently encountered by investigators and students in our research group - the level of risk increases significantly, and specific guidance is needed.**

   We propose several recommendations for devices traveling outside of a protected zone with data stored on them. These recommendations are classified into four main strategies. First, we suggest limiting, to the extent possible, the number of identifiers included in a dataset that is issued for analysis. To facilitate this, we recommend the identification of a data 'gatekeeper', at the outset of a research project, who is charged with the issuance of analytic datasets or tables according to *a priori* data access permissions, and the careful logging of these transactions. (This person is not necessarily database administrator, but is expected to work very closely with those handling the raw data.) Second, there are certain basic system-level security procedures, such as operating system passwords, and adhering to password complexity guidelines, which should be followed. Third, the datasets and associated analysis files should be maintained in encrypted folders or be themselves encrypted. Fourth, if the datasets contain identifiable data, hardware-level security measures should be

considered (system tracking and remote shutdown software, hardware-level encryption, etc.)

C. **Provide Research-Specific Data Security Protocols**: We found that when research applications lack clearly documented data security protocols, data management is at risk of being *ad hoc,* with decisions driven by specific situations. Data security protocols should be thought through as part of study protocol development and should provide enough detail to cover the use of, access to, and transfer of data, including how these transactions should be documented, approved and made available to all members of the research team. Adherence to these protocols becomes a responsibility of the study PI and individual members of the research team.

We advise that each IRB research application include, to the extent possible, a comprehensive data security plan, describing in detail not only how data will be collected, entered and stored, but also acceptable protocols for:

1. data transfer from user to user
2. data transfer from study site to JHSPH
3. who, by name or position, may have access to various forms of the data such as:
    i. Raw, complete electronic or paper data
    ii. Identifiable, analytic datasets
    iii. De-identifiable analytic datasets
    iv. Summary performance statistics
4. physical and data-security protocols to be implemented for physical data, electronic data, and hardware on which data may reside; and
5. if shared with investigators or students outside the study, user agreement template which outlines permitted uses of data, and disposition after use.

This data security plan could also contain a detailed list of principal investigators, co-investigators, students and staff, with clearly delineated "access permissions" to different forms of the dataset within the scope of the proposed research application. This list can then serve as a principal guideline for the data gatekeeper, where any deviations from or additions to the specified access permissions requiring prior authorization from the Principal Investigator. Figure 1, above, illustrates some basic limitations of data access by research team member role.

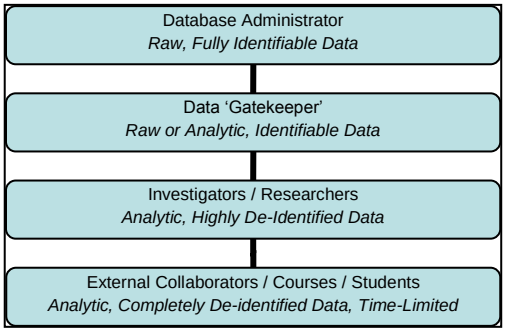


Figure 1. *Illustrative hierarchy of levels of access to various forms of data in a research study.*

II. Specific Guidance: Confidential Data and De-identification

The next sections provide basic information about confidential data and de-identification as well as guidance according to level of user, from database administrator to investigator to student. At each level, our aim is the minimization of RISK, with consideration of what is reasonable with limited resources and effort.

A. Confidential Data.

JHSPH data security policy describes 4 levels of data sensitivity. Our human subject research data falls into the class requiring the highest level of security -- *Confidential*:

Confidential *– data with this classification requires the highest level of protection. In general, this will be data that is covered by government law or regulation*

Protecting Data Confidentiality: Ideally, these data should be maintained exclusively within the network or JHSPH Protected Zone, with only de-identified data being released to any outside use. This general rule however, should be interpreted within the working definitions presented in the initial section of this document – where the "identifiability" of a dataset is defined by the ability of the individual possessing that dataset to link the data back to an individual research participant. Thus, the committee recommends limiting the release of data with ANY form of identifier to any level of investigator or analyst. This "control at source" strategy will reduce the number of 'risk points' for management. This also implies that compiled data (e.g. analytic datasets) should not contain identifiers (beyond a random identifier) unless specifically requested for an analysis.

When identifier-containing data is released, specific control measures will need to accompany the release of that dataset. These are described in section IV, below. The recipient of the dataset should be held responsible for complying with the recommendations that accompany the dataset. This also implies that data for analytic use should be centrally controlled and distributed to users and investigators. Data releases should be tracked in a log file. When appropriate, data use agreements (DUA) may be employed to govern the terms and conditions of data sharing, especially with non-JHSPH collaborators.

Solutions:

1. ***Use de-identified or minimal-identifier data as often as possible***
2. ***Minimize risk by maximizing work "within network"***
3. ***Control release of identifier-data at source / identification of a "Gatekeeper"***
4. ***Track data releases (approval documentation and logging)***
5. ***Use Data Use Agreements (DUA) with recipients when appropriate***

B. De-identification of Data

De-identification is the process of removing personal information that could be used to identify an individual research subject. Once data has been de-identified, it is no longer classified as confidential and does not require a high level of data security. This process, however, may limit the future usability of the dataset and re-integration with other potentially linked data, unless a 'key' to the unique identifier or nonsense identifier has been preserved by the person issuing the dataset (See elaboration on Data Identifiability in BACKGROUND Section).

CAUTION: Even if all methods of de-identification of data protect an individual research participant from a breach of confidentiality, the identity of the data as a "pool" associated with a group or community remains. The interests of the group should be protected as well.

a. Methods

There are at least three common approaches to de-identification of individual-level information:

a) The removal of all personally identifying information from data, except for a nonsense identifier, such that it can only be associated with an individual study participant by a person OTHER than the individual holding the dataset (e.g. Gatekeeper or PI).
b) The removal of all personally identifying information from data, replacing identifiers with a new nonsense, unlinked, identifier, such that it CANNOT ever be re-associated with an individual study participant.
c) Using a nonsense identifier to replace the subject identifier in the dataset, a key for which is preserved by the Gatekeeper who created the database, should a later re-merger of the dataset be required to the source dataset.

Our general practice is to use the first method, in low-risk, field trials of micronutrient deficiency.

b. Examples of identifying information

To completely de-identify a dataset, identifiers such as those listed below would need to be removed for individuals, household members, or relatives:

1. Names
2. All geographic subdivisions (e.g. smaller than a U. S. State., such as city, street address, zip, etc.)
3. Birthdates (Age by year, or a general age category is permissible).
4. Telephone and Fax numbers
5. Medical record numbers
6. License numbers (including vehicle numbers)
7. Full face photographic images
8. Ethnic group
9. Any other characteristic or information which could enable data to be reconnected to its human source.

c. Practical Steps

In day to day practice, it is often necessary to work with identifiable data. In general, study subject names are not needed, but there are many cases where personal information such as Date of Birth (DOB) or specific geographic location is needed by investigators. When working with this data, *partial de-identification* may be acceptable, providing either the linkage between the identifier and the specific identity of a subject is maintained by a data gatekeeper, or if these data are not removable, specific protections are in place on the machine containing identifiable data. Almost always, names can be removed, with others being expunged, as possible: date of birth, age, geographic location, ethnic group, etc.

III. Guidelines for handling data at various user levels:

The data security plan must outline the user roles and "access permissions" for data associated with the specific research project.

A. Administrators/Programmers

Responsibilities: Understand the general principles of data security. Familiarize yourself with the available tools and best practices. Follow the best practices. As technology changes rapidly, it is necessary to continually update this knowledge.

Understand the rules and procedures for distributing identified data.

- De-identify data as much as possible when distributing.
- Distribute the data using secure methods.
- Log the data release
- Obtain PI approval for exceptions prior to release
- Ensure documentation of data use agreement for outside recipients

➔ Use a secure workstation for database and software development. When possible, do not use a laptop for development. If a laptop is required, all data stored on the laptop should be encrypted. See the laptop and external device recommendation below (in section III).

➔ An alternative method is to use a remote access software such as GoToMyPC to access a terminal that exists within the Institutional firewall, and perform all coding operations without storing any data on the external laptop. This approach has been successful for several users within our group.

Paper and Electronic Forms Development:

Often development is a collaborative effort, and there is significant exchange of information (drafts of forms, comments, data samples, etc.) between developers and with clients. If security procedures are not followed closely, there is a significant risk of data leakage.

➔ One method to eliminate this risk is to use a database with dummy data. However, in real day to day work, this can be very cumbersome due to the constantly changing nature of a database that is under development. There is often the need to look up errors referring to a specific record from a certain data entry form, or the need to validate or check on data entry progress. It is possible to do this without a copy of the live data, but it is not practical. In day to day work, Developers and Administrators need regular local access to copies of the full research database. This requires that security measures are in place and security methods are followed.

B. Investigators

Liability for data security rests with the Principal Investigator who is considered the data owner. **The data owner may delegate management of data security, but may not delegate the liability for protecting the data.**

Investigators require regular access to data files that represent both the raw data (i.e. data that are reflective of the database where data were originally or are currently stored)

and “analytic data files” (i.e. data file [or set of files] derived from the raw data files). Such derivatives may be identical to the raw database files, or may reflect further manipulations of the raw data including, but not limited to: labeling variables, renaming variables, recoding of variables, generation of new variables.

- ➔ Unless there is a specific need for the investigator to have access to fully identifiable data, the investigator could designate a specific individual, using a fixed, access-limited workstation to preserve the complete database, and issue working datasets, as needed, following the guidelines described above.

- ➔ An investigator using a portable device (laptop, tablet, usb drive, etc) should also follow the system securing recommendations described below in Section IV.

C. Field Staff

Data security practices at any field site should reflect best practical standards in data security. Even in the most remote settings, basic guidelines must be in place, carefully defining the specific parameters and controls for hardware and software security. Training should be conducted to ensure field staff have the appropriate knowledge to successfully implement and maintain data security. Involve field staff in security planning and in the decision making process so they fundamentally understand the underlying risks in, reasoning for, and strategy for securing data.

- ➔ As described earlier, clear access guidelines should be in place at the beginning of a study, delineating specific individuals (or positions) with authorization to access or obtain particular kinds of data. Procedures for obtaining approval to deviate from the access guidelines should also be specified – e.g. who has authorization to change the level of access or to provide permissions to access to any kind of data (irrespective of level of de-identification).

- ➔ Often, field sites have many visitors —students, investigators, in-house staff and external visitors. It is important that the individuals who control access to both hardware and data understand and are empowered to enforce the appropriate level of access for each person who may be ‘exposed’ to the data. Simple decision trees may be created to illustrate the guidelines to verify if data access rules are being adhered to. Periodic audits can also be used to check if unauthorized data is available on official study laptops, data transfer devices, etc, being used by staff or students.

D. Students and Non-JHSPH Scientists

- ➔ Any individual or position not explicitly listed in the study “data access permissions” table of the Research Protocol, written at the outset of the study must go through a process to obtain PI approval, demonstrate human subjects certification, and be added (with clear access permissions) to the permissions list.

- ➔ It is recommended that the terms and parameters of the data sharing be carefully delineated, including the level of identifiable data being provided, and any associated physical security measures that have to be in place for identifiable data. We also recommend that there be clear “expiry” dates for the data sharing, such that upon completion of a collaborative project, data may be deleted or withdrawn.

The committee discussed a number of technological solutions to potentially control duration of access, but found that all currently available solutions to be limited by the honesty of the end user. For example, a centrally controlled, cloud storage solution could be used to remove individual access to a dataset upon completion of a collaboration. However, the end-user could copy files to their personal system or storage device, rendering the solution useless.

E. Students (Course-level interactions)

Data provided to students in course-level interactions should always be completely de-identified. No identifiable data should be provided to students who require data for exclusive use for course requirements. The data should be reviewed and approved by the study PI, and the specific dataset issued should be logged by the data analyst.

IV. Use of Portable Electronic Devices (Laptops and External Devices)

When data are stored on a laptop, or external device (thumb drive, USB external hard drive, etc.), the file or folder should be password protected and encrypted. Backups are not excluded from this requirement. Backup software usually has an encryption option. These recommendations apply to any situation when an identifiable dataset exists on a computer or computing device that is mobile and may leave the physical premises of JHSPH. We recommend that the PI of the project require all authorized individuals with access to data to follow these recommendations as part of the data security plan.

Why password protect? Password protection prevents others from easily gaining access to your computer through the standard user interface: screen, keyboard, and mouse. This is important for preventing unauthorized access to your operating system in situations where you briefly step away from your computer, or when your computer is lost. If your computer is lost or stolen and your hard drive is encrypted, but you do not have password protection turned on, anyone who finds your computer can turn it on and access all your data. System-level encryption, in the absence of password protection, is useless. Password protection alone, however, does not prevent someone with a minimal amount of skill from accessing data stored on your hard drive.

Why encrypt? If the data on your hard drive is not encrypted, you computer can be opened physically, the hard drive removed, and connected to another computer, from which that data can be accessed, without any password. If the data on the drive is encrypted, the data on the hard drive cannot be accessed in this way without knowledge of the encryption key.

A. Encryption

JHSPH provides PointSec encryption for Windows laptops. This software is available for a license fee of $69. This is recommended and has the benefits of complete encryption of the hard drive and a centrally managed key system for data retrieval in situations where data needs to be recovered (such as operating system corruption).

For Macbooks, hard drive encryption is also recommended. JHSPH does not offer a system for encrypting macbooks. At a minimum, folders containing confidential data should be encrypted. A practical approach is to use the Mac's built in File Vault encryption.

More information on encryption is available on the JHU Guide to encryption:

http://www.it.johnshopkins.edu/security/crypto.html

B. Password Recommendations:
All computers, irrespective of their mobility, should require password entry at start up and after a short period of inactivity (5-15 minutes).

On windows, this can be accomplished by setting the screen saver to turn on after 5-15 minutes of inactivity, and also selecting the "On resume, password protect" option. When stepping away from your computer, actively lock the computer by pressing "Windows key-L".

On Macs, under System Preferences→Security, select the option to require password after sleep or screen saver begins.

The use of a strong password must be required.
-The longer the better.
-Use a combination of letters, numbers, and symbols
-Use at least 14 characters

Avoid:
-Using a single word from a dictionary in any language, these may be hacked by dictionary-accessing software.
-Personal information (birthdays, middle names, children's names, locations)
-Sequences or adjacent keyboard keys: "123456789" or "qwerty"
-Words spelled backwards or abbreviations

Tips:
-Long passwords are hard to remember, but passphrases are easier to remember and still difficult to crack. If a password policy allows it, use the space character, but passphrases can be created without a space also.
Examples:
StereoMusicSoundsGood!ToMe
Stereo Music Sounds Good To Me.
(woWthaTcakEsmelleDgooD2me)
(woW thaT cakE smelleD gooD2me)
0iCu812

- There are online services to test password strength. If you use these, do not use your actual password, use a password with a similar structure and length. Here are some links to password checking tools:

http://www.microsoft.com/protect/fraud/passwords/create.aspx
https://www.microsoft.com/protect/fraud/passwords/checker.aspx?WT.mc_id=Site_Link

- Sometimes you are only allowed to enter 10 or 12 characters. In this case, use the maximum number possible and be sure to include a mixture of letters, numbers, and symbols.
- Change your password yearly.
- Be wary about using your password. Avoid logging in at internet cafes or on other people's computers. Be especially careful not to save your password in the browser by selecting "remember me" or "keep me signed in".

JHSPH enforces a password policy for accessing email and objects under its Active Directory managed network.

JHSPH password creation policy

1. Minimum length is 8 characters. Some accounts will require longer passwords.
2. Utilize both upper and lower case characters (e.g., a-z, A-Z)
3. Have at least one digit or character from the following list: !@#$%^&*()_+|~-=\`{}[]:";'<>?,./. Depending on the system or software, special characters may not be allowed.
4. Never Expires

C. Tracking Software

If budgets permit, beyond the encryption of data, hard drives and strict password policies, there are software and hardware solutions which allow remote tracking and disabling of a lost or stolen computer. Examples of these include Lojack®, or Computrace®. There are privacy concerns and the software is extremely difficult or even impossible to remove. These cons should be weighed against the value of using this software.

V. Using Data Files, Data Storage, and Data Transfer

A. Using Data

Use of database files: Raw database files often contain immediately identifying information (first and last names) of participants. In general, most users do not require access to copies of the raw database (e.g. SQL). In the rare instance that such access is required/requested, ideally the user should connect to a SQL server within the network (i.e. on a network server), rather than downloading and restoring a copy of the database to a local SQL server on a specific machine. If the latter path is chosen (for example, if access outside the network is required), the user should make all efforts to restrict use to a temporary nature, and remove the database from his/her machine when finished using it.)

Exporting from raw SQL database to analytic files: In general, bulk export of individual tables (or views) from the raw SQL database into external software (i.e. SAS, STATA, EXCEL, etc) produces data that are identical to the raw database, including the incorporation of immediately identifying information (first and last names) and secondary identifying information (ID numbers, addresses, date of birth, etc). Therefore, it is advisable, if the export from raw SQL database to analytic files is done in a manner that strips the immediately identifying variables (first and last names) from the dataset. For example, if the investigator (or database administrator) has access to a network SQL server containing a copy of the live/original database, he/she should ensure that the scripts written to export the individual tables drop/exclude variables that are not necessary for analysis.

Documentation and reproducibility of manipulations from the raw data: Users should use basic principles of documentation of data management and analyses so that not only analytical results can be reproduced from the raw database, but any creation of primary analytic files can also be reproduced

Using and Managing Analytic Files: Users (especially investigators) within our group need access to analytic files on a continual basis, including periods away from the School's network. Therefore, we require solutions that allow for maintaining analytic files on local machines, even if only for temporary use. Some basic guidelines that should be adhered to, when possible:

- Check to ensure that analytic files do not contain first/last names.
- When possible, utilize dedicated password protected, restricted-access space on shared network drives and conduct and save analytic work directly into these locations.
- The use of cloud-based servers to store and manipulate data is also a feasible, secure option to manage data without permanent local storage.
- The use of remote access software is also an interesting solution – provided users maintain a primary machine which resides within the JHSPH secure network, and remotely access the machine, conducting any analysis and storage of data on the secure machine 'virtually'.
- If a local copy of the data is required
- Follow recommendations in section IV. "Laptop and External Device" to secure your device
- Work within the secure network as much as possible, and when possible, upload your work to dedicated space on our shared network drives, as above
- After uploading to our network drives, remove local workspaces when no longer needed
- Remove local copies of the data when no longer needed

B. Data Storage

In this section we review a number of options for the storage of data, analytic files and other sensitive documents.

Local File Servers

Physically, the file servers are in locked rooms and only authorized personnel have access. The systems are password protected, but the data is not encrypted. These systems have regular OS patching and antivirus updates. Network security is provided by the JHSPH firewall and network security systems. User access is granted through shared folders. Users do not have execute permission on the server and cannot run or install software on the server.

Currently, user-access is only available through the school's secure network, but we are planning to explore file access through secure internet technology. This openness should decrease the instances where users need to save files to their local computers. MS Sharepoint will be used for this feature, but this service is not yet available.

Cloud Storage

Currently, we are using cloud storage (Dropbox.com) for active, collaborative data analysis projects. These are mostly projects that are in start-up phase and involve varying kinds of research data. Users must be specifically invited to share a folder, which can either be accessed online only, or be copied to the local hard drive of the invitee. If shared folders contain data, and any end users maintain their dropbox files

on a mobile device / laptop, all the specific policies and procedures described in Section IV should apply.

It is difficult to completely audit cloud system security, but it is rapidly becoming an accepted standard. Data is heavily encrypted, all transfers are logged and handled securely (via https or sftp). Details about Dropbox.com's Amazon S3 security is available here: http://www.dropbox.com/help/27

Amazon has policies regarding data security, and uses various certifications and audits to support their claims. Details can be found at: http://aws.amazon.com/security/

Microsoft claims their cloud is secure and reliable, but their policies are difficult to locate. Here is a PDF regarding their general cloud security practices: http://www.globalfoundationservices.com/security/documents/SecuringtheMSCloudMay09.pdf

Google, specifically, Google Docs, is not explicit about its security policies and practices. They offer general reassurance, but do not claim to adhere to standards that would be acceptable for confidential information. http://docs.google.com/support/bin/topic.py?topic=15143

C. Data Transfer

In general, online transfer is adequately secure if it uses strong encryption. **Email and FTP are not commonly encrypted, and as such, are not acceptable means of data transmission.** Normal web transactions (browsing web pages, typing data into forms) are not encrypted, unless specified. Web transactions are adequately secure when SSL is used. For browser traffic, this means the "HTTPS://" prefix is visible at the beginning of a URL.

Recommended Procedure for Distribution of Non-identifiable Data.

1) Prior to data access, students and external investigators not explicitly listed in the IRB data access listing should:

a) Get permission from the Principal Investigator (PI) and fill in a Data Request Form (DRF). This must all be documented electronically or on paper.
b) Write a brief research and analysis plan.
c) Obtain approval from the IRB (if appropriate, or necessary).
d) Sign Data Use Agreement.
e) Provide a timeline for Data Use.

2) All these documents should be approved by the PI. The data manager must be copied on all correspondence. The data manager will create an individual folder on the secure server or cloud server for each separate study project.

3) The data manager will create the data set with a mapped set of Identification numbers (NEWID). This data set will not contain any personal identifiers (names, addresses, etc.). The data manager will safeguard the Identification key. The data set will be stored in the individual folder on the secure server. It will not be on a personal computer.

4) When the study project ends, the PI and data manager will send an e-mail to the student or scientist and ask them to delete the data set. This email will be kept on file for the record.

VI. Data Security Risk Scenarios

Scenario #1: Graduate Student requests additional data related to thesis:

Email: "...can you check the database for the following two maternal ID numbers to see if you can obtain either the NAMES or ETHNICITY:"

ID NUM	Address
00001	228
00002	229

Appropriate Response:

-Verify with Data Manager or Data owner that student is allowed access to request this specific data
-If *ETHNICITY* is available, release this in a file on a secure server. It is preferable to release ETHNICITY in place of NAMES when possible, as this is less identifying information.
-If NAMES are needed, be sure to caution the student regarding the use of names outside any secure network, and that they must comply with enhanced security regulations. Remind the student to work with the data on the server and not to save it to their local machine, if possible.

Scenario #2: Investigator (data owner) requests dataset for analysis:

Verbal request at meeting: "Can you give me all the ALRI data to date for all the men enrolled in the study, with age between 30 and 40, in study areas 4, 7, and 9?"

Appropriate response:

-Find out if any identifying information is needed, DOB, Address, etc.?
-If no identifying data is required, dataset can be created and transferred to the investigator by secure server.
-Email **may** be used to transmit the de-identified dataset, but this is not optimal.

Scenario #3: Staff for an overseas study wrote to an investigator in an email:

> "...Please find the updated data set as requested. I have entered 103 new records of deaths into the existing data set (ref: mail below). So the total number of records in the attached file is 432 (329+103).
>
> *Attn: Field Data Manager and Field Manager:* I am attaching an email with some data clarifications. Please preserve these communications in a folder. We would need these communications/explanations in future.
>
> The attached datasets had ID, Name and other data for each woman. Not only was this data being emailed without a password, it was on the Physician's laptop in an excel spreadsheet which most likely had no security.

Appropriate response:

- Refresh Physician's data security training and identify the lapses in security policy.
- Delete the email with the attachment from all recipient inboxes.
- Encrypt the file the Physician stores this data in.
- Administrative action may be reserved for repeat violations of security policies.

Below is an example of providing data for a student. This example models appropriate procedures and possible technologies to be used for distributing data.

Example: Student requests DOB and date of visit information for study X.
Procedure:

- Obtain documented approval for release of data.
- Log the transaction and terms of data release (e.g. Expiry date of Data Use Agreement)
- Remove unnecessary identifying information.
- Export data from database to spreadsheet or CSV file. Copy file to thumb drive.
- Because the file contains partially identifying information, a higher level of security must be maintained.
    - Encrypt the thumb drive.

Example: Student requests previous pregnancy histories for women enrolled in study X to be shared electronically.
Procedure:

- Obtain documented approval for release of data.
- Log the transaction and terms of data release (eg. Expiry date of Data Use Agreement)
- Remove unnecessary identifying information.
- Export data from database to spreadsheet or CSV file.
- Copy file to an encrypted folder, or encrypt the file itself, on a cloud server or file-sharing system, with password-protected access

Appendixes

Appendix A. JHSPH IS Document “Data Security Measures When Using Personal Identifiers” – available online

Appendix B. JHSPH IS Document “ Data Security: How Should Investigators Protect Confidential, Identifiable Study Data?” – available online

Appendix C. JHSPH IS Document “Data Security Checklist” – available online

Appendix D. Data Request Form Example

Appendix D: Data Request Form (Sample)

August 24, 2006 / AL

DRF v1.0 / August 24, 2006 / JiVitA

Data Release Form (DRF)

Name:_____ JiVitA ID [ ] Designation:_____

Date (dd/mm/yy) ___ ___ _____

Dataset type: ☐ SQL (Complete) ☐ Frozen (All or Specify):

_____

Freeze Date: [ ]

Format: ☐ STATA ☐ SAS ☐ SPSS

Data Issued by: Signature Needed JiVitA ID [ ] Date:___ ___ _____

Approval (if needed) by: Signature Needed JiVitA ID [ ][ ][ ][ ] Date:___/___/_____

Analysis Plan Attached ? ○Y ◉N

COMPLETED DRFs MUST BE FILED IN THE DMC IN THE DRF REGISTRY FOLDER.

Comments regarding the dataset provided (ie. Merges / Exclusions / Masking):

SAVE A COPY

CLICK TO PRINT