David Thiemann Advice to Investigators

What is a Data Management Plan?
Why Do I Need One?
Why Do I Have to Jump Through All These %$%!$Q# Hoops, Anyway?

Caveat investigator: This document focuses strictly on data protection, management and security. From the perspective of NIH and the scientific community, a data plan also includes data dissemination and sharing. Many studies now and in future will require depositing de-identified, patient-specific data into resources such as the NIH dbGaP (http://www.ncbi.nlm.nih.gov/gap)

In the digital era, data breach is an increasingly common catastrophe for all concerned. Johns Hopkins Medicine takes data security seriously, for obvious reasons:

1. We have a moral and legal duty to safeguard the confidentiality of our patients, who could lose control of their privacy and could suffer financial harm, such as identity theft and adverse insurance selection.

2. Data breach involves huge financial and reputational risk to JHM as an enterprise, and directly to individual departments and researchers. The Health Insurance Portability and Accountability Act (HIPAA) requires that, in the event of a data breach involving >499 patients, both the news media and the federal Department of Health and Human Services (HHS) be notified of the breach without unreasonable delay and in every case within 60 days. In the event of a breach, HHS has assessed civil penalties of (literally) millions of dollars ($4.8M against Columbia/NY Presbyterian, $865,000 against UCLA). Patient notification after data breach costs >$150 per patient, which for a 25,000-patient dataset means $3.75M. These costs would be born by the responsible division or department, **NOT** by Johns Hopkins Medicine or JHU SOM.

The goal of JHM policy for research data management is to balance necessary data security with researcher needs and capabilities—which mostly means balancing the relative unhappiness of the data-security and research communities.

How does a data breach happen?

In clinical research, data breaches usually don't involve fancy firewall breaches or phishing exploits by overseas hackers. The root causes of research data breaches usually are low-tech carelessness, bad data-system design and bad practice. Her are common examples, all of which which violate JHM IT and IRB policy:

1. Research data often is stored on unencrypted, un-backed-up standalone devices rather than centrally administered servers. A desktop hard-drive crash can permanently destroy a research database. The most common cause of data breach is lost or stolen portable devices, such as laptops, thumb drives and CDs; the second most common cause is exposing internal databases to public Internet/web access.

2. Research databases often are located in shared folders/drives, so that anyone with access to the folder can copy research data. Passwords often are shared, such as a common password for a study.

3. Even when access is limited to the research team, research databases (typically .xlsx files) often allow global access by the entire study team, so that a rogue user--a disgruntled research assistant, a student--can remove data without leaving a trace, by coping a file to a portable device or emailing it outside the institution. (Unlike clinical systems, research databases seldom limit users to row-level access.)

4. Research databases seldom are encrypted at rest/on disk.

5. Research data access seldom is logged (even at the file level), so it is impossible for the principal investigator or the JHM technical team to determine who did what to the data and when, or to know what data has been potentially breached.

Assessing the riskiness of research data

No single factor determines the riskiness of a research database. There are several considerations:

* Research topic. Regardless of individual data fields, a database of patients with sensitive conditions (substance abuse, HIV or STDs, psychiatric conditions, adolescents, for example) involves high risk.

* Number of subjects. Due to both common sense and the HIPAA breach-notification requirement, a database with thousands of patients poses a greater risk than one with a handful of patients. JHM Legal Counsel has determined that the IRB (through its consultants) shall review a Data Security Profile and if indicated a data-management plan for all studies with >=500 patients and for any study with high-risk subjects or data.

* Database content. Regardless of the number of patients, a database that contains patient ID data (any combination of medical record number, date of birth, name, address and phone number (or even worse, social security number) is high risk, as is a database that contains sensitive information, such as infectious-disease serology or diagnosis, medical complications, some quality improvement/legal data, substance abuse or addiction, or psychosocial data.

* Number and nature of staff with database access. A database that involves dozens of research assistants or abstractors involves greater risk than a database with a handful of investigators.

Developing a data-management plan

Because data management plans necessarily are tailored to particular study methods and risks, there is no cookie-cutter, check-the-box data management plan. In escalating fashion, here are some routine precautions:

Basic Precautions

1. All data will be stored in one of three locations: a secure institutional environment (such as REDCap or the JHM SAFE secure research data desktop, described at http://ictr.johnshopkins.edu/clinical/clinical-resources/clinical-research-informatics-

core/secure-research-data-desktop/); a separate folder on a LAN-administered, JHED/active-directory-enabled server (typically a divisional or departmental server), with access limited to the study team; or JHBox.

Notes:

a) For small projects, the simplest, easiest, cheapest (because free, underwritten by the ICTR) method of research data storage is SAFE. REDCap is ideally suited to larger projects, especially those that need extensive data entry or chart abstraction, or involve high-risk data.
b) Divisional/departmental file shares are allowed but often an expensive nuisance. The LAN administrator must enable file-level logging for critical (data) files, such as .xlsx and .mdb files. The necessary logging is described at http://www.it.johnshopkins.edu/restricted/standards/WindowsServerStandardRevisedAPPROVED_07062015.pdf, starting on Page 9. The server administrator will maintain operating-system, application and anti-virus patches per IT@JohnsHopkins standards. LAN admins are familiar with these standards. (JHED-enabled servers effectively eliminate shared passwords and also mean that, should a user's JHED credentials be revoked, all data access is immediately blocked.)
c) JHBox for file storage is discouraged and included only for backward compatibility. JHBox originally was designed as a file-transfer utility, and in that capacity remains an excellent well-logged tool, but it has significant limitations as a data-storage system, including poor application file connectivity (for statistical applications, for example); limited backup, so corrupted/inadvertently deleted files may be lost forever; and no staff support for security settings or file recovery, so the user (rather than a LAN administrator) is responsible for everything (thus fudging a basic data-security principle—the honor system is not best practice).

Precaution #1 above implies an important prohibition: Raw identifiable research data shall not be stored on desktop workstations or laptops. Although desktops and laptops historically have been used for research data, in the current era they are inappropriate because file access and email/portable device connectivity are minimally logged, and files that are corrupted (eg hard drive crash) or inadvertently deleted are not recoverable (a liability that should terrify every researcher). Research data needs server storage.

2. The study will maintain separate raw-data and analytic files, with a link table connecting PHI identifiers (for example, medical record number) in the raw data file(s) to an anonymized study ID in the analytic file(s). Access to raw data files containing PHI will be limited to a single trusted person (typically a database administrator), with passwords escrowed in the event that the administrator is unavailable or incapacitated. Analysts will work with de-identified analytic files. (Analytic files may contain HIPAA limited datasets, which can include dates.)

3. Data must not be stored on portable devices, such as laptops, thumb/USB drives and CDs. (There are exceptions for approved encrypted devices but far better simply to never put PHI on portable devices. Disk/device encryption does not just mean a password. It requires special tools.)

4. Unencrypted data will not be transmitted by email. This means that **any** electronic data transmission must be encrypted. Encryption can be done in various ways: By using native application

utilities (for example, Excel's file-level password protection (Prepare-Encrypt Document, then enter a password), by using JHBox, or by SFTP). If file-level passwords are used, the password should be communicated separately from the data file, by a different method (eg verbally or text rather than email).

5. Data shall not be transmitted outside the JHM firewall, except in conformance with a HIPAA Business Associate Agreement and/or approval of the JHM Data Trust. For multi-center studies, data submission to a coordinating center requires Data Trust review.

6. Research data (including lists for prospective enrollment) shall not be acquired with clinical reporting tools, such as Epic Reporting Workbench; the results provided by such tools are unencrypted, insecure and unauditable, and involve email or download to portable media. Similarly research data shall not be acquired by informal/unregulated back-channel arrangements, such as friendly administrative analysts.

Moderate security
All the basic precautions, plus:

1. Individual research assistants/abstractors will not have global data access, and will not be able to copy the research database. (N.B. This effectively means that the study will not use Excel or similar spreadsheet tools for data storage.) Instead the study will use an application (such as REDCap, available via the ICTR; Cold Fusion; a Visual Basic for Applications app; or JavaScript) to provide row-level access to a back-end database (such as a SQL database).

2. Servers will be physically located in an enterprise data center (such as the Mt. Washington or 1830 data centers).

High security
All of the basic and moderate precautions, plus:

1. All system-, table- and row-level database access will be logged (the same standard as for clinical information systems, which log all data access, including creating, reading, updating and deleting records).

2. Logs will be periodically scanned/monitored for unusual/unauthorized activity (eg failed login attempts, unexplained bulk downloads or queries). DBAM (Database Activity Monitoring) software may be used to monitor the database for unusual queries, connections and activities. All extracts/exports are logged.

Afterthoughts/addenda:

1. **Data de-identification.** Researchers sometimes over-promise data de-identification, adopting in the name of data security steps that are bad practice for research: For example, not entering medical record number or date of birth into a research database, or destroying protected PHI after the research database is compiled. Such steps are bad because de-identification precludes data auditing, data cleanup/validation, and *any* downstream linkage (to survival data, pathology specimens or genomic data, for example). Far better to rigorously secure the raw research database, which contains PHI; and to de-identify analytic extracts. In the words of Mark Twain: "Behold, the fool saith, 'Put not all thine eggs in the one basket' - which is but a matter of saying, "Scatter your money and your attention; but

the wise man saith, 'Pull all your eggs in the one basket and - WATCH THAT BASKET.'" - Pudd'nhead Wilson's Calendar.

2. **Data integrity.** Any research database (whether in Excel, RedCap, SAFE or SQL) needs to be designed for quality control, downstream cleanup and auditability. Among other things, this means that:

a) The database needs to record who entered every data row, and when; and who subsequently modified (updated) the row, and when. (Ideally there also should be a history table with every iterative update.)

b) The database and user interface should be designed to minimize data-entry errors--for example, by requiring multiple patient identifiers (MRN+DOB, or study ID + DOB, for example) before allowing data entry. Without dual data entry or rigorous validity checking, there will be a 3-5% rate of fat-finger mistakes.

c) Database schema and permissions should distinguish between classes of users or sites. This is especially important with multi-site databases--the database should not allow users from one site to modify data for patients of another site.

d) The database and data management plan should include primary keys (to block duplicative data entry), constraints and/or queries for out-of-range dates, invalid heights/weights/BMI, missing data, etc.

What are the most common researcher mistakes in data management plans?

1. The data management plan is not actually a plan (which clearly states what the study is actually proposing to do), rather merely a list of options and a vague promise of good intentions, comprised mostly of weasel-words connected by "or". The IRB cannot vet or approve non-plans.

2. The eForm A/B fails to specify whether the study/project will acquire clinical data via manual chart review (whether of paper or electronic charts) or bulk query (or both). Ambiguity leads to misunderstandings. Bulk query needs to be clearly identified and defined.

3. Researchers often plan to store data on a desktop/laptop. Such expedients may be technically allowable (if the device is encrypted) but they are bad practice, for the reasons described in **Basic Precautions** #1 above.

4. Researchers often plan to store data on a personal folder on a divisional/departmental server. This is safer than a laptop or desktop but remains inadequate, for several reasons: The server-side logging (described above and at http://www.it.johnshopkins.edu/restricted/standards/WindowsServerStandardRevisedAPPROVED_07062015.pdf) for research data is different than for personal data; commingling personal and research files means that audit or external review may compromise personal files; and if project scope or resources expand beyond a single investigator, data from a personal folder cannot easily be shared with new researchers or assistants.

5. Researchers often misunderstand data linkage issues. Sometimes the eForm A/B overpromises absurd security (saying everything will be anonymized, which renders research unauditable/inextensible (see **Data Integrity** #2 above); sometimes the eForm plans to use MRN as an identifier in the analytic file (with associated HIPAA/PHI issues and 3-4% fat-finger errors). MRNs are neither unique nor stable, and are HIPAA PHI, so they are an unacceptable primary key; a study-specific patient ID should be used.

6. Researchers often use Excel simply because it's easy and familiar. Excel may be usable for small simple studies but it's inappropriate for complex/high-risk data, for reasons ranging from auditability to data corruption and data breach.

7. Researchers often assume that REDCap is a panacea for all data security problems. REDCap is a great data entry tool but it's poorly suited to data cleaning and to automated bulk clinical-system extracts. So the data management plan needs to anticipate server-side pre- and post-REDCap storage for many projects. In addition, REDCap by itself doesn't ensure adequate data security; much depends on exactly how REDCap is configured, especially in matters of data export and account management. N.B. RedCap is NOT (yet—remediation in progress, we hope) JHED-enabled. This has concerning implications: It encourages use of shared passwords and it means that, in the event that a study team member is fired, access is not automatically terminated.

8. Researchers often assume that because Microsoft Access, Filemaker Pro and similar software are database programs, they're a panacea for data security problems. They're good programs but need expert setup; otherwise they involve many vulnerabilities, including shared passwords and risk of copying/corrupting the entire database (eg .mdb file).

9. Researchers sometimes don't realize that technical details of data sharing/export to non-JHM entities (a commercial sponsor/partner or an academic multi-site study, for example) require special precautions (including specific encryption) and approval by the Research Subcouncil of the JHM Data Trust, chaired by Stuart Ray, MD, and Chris Chute, MD; staff contact is Valerie Smothers. (For commercial partners, JHU/JHM legal approval, a Research Collaboration Agreement (via the Office of Research Administration) and/or a HIPAA Business Associate Agreement may also be needed.)

10. Researchers sometimes assume inappropriate data access, eg "I have a friendly analyst/administrator who will query Datamart" or "I already have Datamart access, and can get whatever data I want," or "I'll run my own Reporting Workbench report on Epic." Research data needs to come from defined, auditable, institutionally controllable sources, not ad hoc back-door arrangements.

11. Research projects sometimes assume that free-text data (such as radiology and pathology reports) exists in categorical/granular form. Text is not data. Retrospective projects also sometimes wrongly assume that Epic contains legacy data from systems such as Meditech and Sunrise Clinical Manager. (Epic does contain essentially all data from the old JHH EPR system.)

12. Researchers sometimes assume that they can download data extracts directly from Epic or Datamart to a statistical package such as Stata. Dealing with clinical data extracts is like trying to sip from a very dirty fire hose. Clinical data extracts need extensive cleaning and separate storage as part of the data management plan.

13. Researchers sometimes fib about sample size, saying that their database will have 495 patients (to avoid the 500-patient HIPAA breach-notification threshold), when in actuality they plan to download and winnow through a database of 10,000 patients to get to 495. From an IRB/regulatory/data breach/HHS threshold, it's the initial/largest cohort of eligible patients that matters, not the final group.

14. In eForm A/B researchers sometimes broach a subject in very general terms, with no indication of exactly what data they plan to collect. The IRB needs a data dictionary, with specificity about scope and data elements, rather than "we're going to collect lots of data about patients with XX disease."

15. Researchers sometimes blithely promise to de-identify data. De-identification is tricky and technical, **NOT** simply a matter of substituting a study ID for a MRN. **All** dates (not just date of birth but service date and lab dates) need offset or obfuscation; zip codes need selective masking. The JHM Data Trust typically requires that de-identification be done by an expert honest broker such as the Center for Clinical Data Analysis, not by a PI using Excel. Details at this url: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

16. Researchers often submit IRB protocols which assume that cases (and/or controls) already have been identified, without specifying exactly how patients will be identified. From a data-breach standpoint, patient/case identification is one of the riskier steps in the research process, because it involves hundreds or thousands of patients; from a statistical/epidemiological standpoint, the sampling methods used to identify cases and controls are the leading source of biased results and bad science. For both IRB scientific review and data security, the exact method of case/patient identification needs to be fully described.

17. Researchers often believe that Excel file password protection confers data security. Excel password protection (except for one-time encryption passwords for data export, described above) inherently involves a shared password, thus violating a basic precept of data management and security. Research databases need robust access control and logging, typically with JHED logins and Active Directory.

18. Researchers sometimes use insecure file-sharing and data-storage methods, such as DropBox or Gmail. Research data needs to be stored and shared on logged servers inside the JHM firewall.

19. Helpful urls/forms (also embedded in the IRB eForm B):
    IRB Data Security Profile form (needed for all studies involving data from >=500 patients):
    http://www.hopkinsmedicine.org/institutional_review_board/forms/DataSecurityProfile.doc

    IRB Data Security Checklist form (for studies that do not comply with or meet the requirements of the Data Security Profile):
    http://www.hopkinsmedicine.org/institutional_review_board/forms/Data_Security_Checklist.doc

    JHM Privacy Office Use of Data Agreement (for data extracts from JHM enterprise databases, such as Epic, EPR2020, SCM and CaseMix):
    http://intranet.insidehopkinsmedicine.org/privacy_office/_docs/additional_information/Use%20of%20Data%20Agreement_012816_clean.pdf

    JHM encryption standards:
    http://www.it.johnshopkins.edu/restricted/standards/EncryptedStandardsRevisedAPPROVED030116.pdf

dt 7/13/2017