Objectives

- Review of available database and data management tools and their capabilities
- Demonstrate examples of good data management
- Discuss strategies and good practice for data security and backup
- Discuss data conversion and transfer options

Overview of Data Management Tools

- We will look at 4 types of "data management" tools used
  1. Spreadsheet (Microsoft Excel, Google spreadsheet)
  2. Desktop database (Microsoft Access, FileMaker Pro)
  3. Web database (REDCap, OpenClinica, CaBIG)
  4. Mobile data collection
Overview of Data Management Tools

• The spreadsheet
  • “Lowest common denominator”
  • Format that all users can open and edit files
  • Often used by labs to output their results
  • Disadvantages
    • Data can be corrupted by single column sorts or accidental keystrokes
    • Data entry interface is limited.
    • Sometimes program tries to “assist” the user with data
    • Not designed for multiple user entry
  • Helpful spreadsheet data management commands
    • Transpose data (Paste Special, transpose)
    • Remove duplicates (Data, Remove Duplicates wizard)
    • Match merge data (VLOOKUP)
    • Concatenation/substring (CONCATENATE, SUBSTRING)
    • Data Validation (Data, Data Validation)
    • “Forms” tool for data entry (Add button to Quick Access Toolbar)
    • “Table” to maintain integrity (Insert, Table)

Overview of Data Management Tools

• The desktop database
  • Microsoft Access and FileMaker Pro widely used
  • Include database, forms, reports, queries, custom programming
  • Easy to initially set up using wizards
  • Designed for simultaneous access from multiple users
  • Disadvantages
    • By default, limited to local network access
    • Built-in wizards can make it easy to start but tasks beyond wizards require training
    • Scalability can be an issue if amount of data is large

Desktop Database: Microsoft Access

All of the database elements are in one file
Desktop Database: Microsoft Access

Overview of Data Management Tools

- The web database application
  - REDCap, OpenClinica, CoBeeG are examples
  - Include database, forms, reports, queries, user/site management, audit trails, backup and security
  - Automatically build forms from codebooks
  - Accessible from a web browser – no additional software needed
  - Scale for very large studies and multiple sites
  - Most have large user communities as a resource

Disadvantages
- Features that YOU want may not exist yet
- Some changes require approval by system admin

http://www.project-redcap.org/
Mobile data collection

• Mobile data collection
  – Apple (iPhone, iPad) and Android (phones and tablets) are main options.
  – Data is entered directly on a smart phone or tablet
  – Data can be stored locally and transmitted manually OR it can be transferred in real time to a central database.
  – Smart phones work well for simple question types
  – Tablets are better suited for complex question types with many answers OR the ability to see previous answers
  – Considerations:
    • If data is stored on the device, is it protected in case of theft, damage, etc.
    • Do you need a mobile data connection to transfer data on-demand or will it be transferred via WiFi or USB?
    • Battery life of the device.
    • Where will the data ultimately be stored? Central server, website?

Mobile data collection

• Options for mobile data collection
  – iFormbuilder
    • https://www.iformbuilder.com/
  – doForms
    • http://www.doforms.com/
  – Magpi
    • https://www.magpi.com
  – Formhub (for Android only)
    • https://www.formhub.org/
  – Enketo
    • https://enketo.org/
  – Open Data Kit (ODK) (for Android only)
    • This is what our Center currently uses for mobile data collection.
Form & Database Design

- Recommendations for fields and codings on forms
  - Make codes for missing and refused data fields.
    - A common convention is ‘9’ for missing but only if it is not a valid value
    - Problematic if the 9’s get included in the analysis
    - Sometimes need to distinguish between
      - “missing – not obtained”,
      - “don’t know”,
      - “refused to answer”

- Use multiple identifiers on each form as a cross check
  - You don’t need to save all of them each time. It is simply a cross-check.
Form & Database Design

• Recommendations for fields and database design
  – Make codes for missing and refused data fields.
  – Consider requiring that all fields are entered unless they are conditional on another question.
  – Use multiple identifiers on each form as a cross check.
  – Code your responses whenever possible and be consistent in your code assignments. For example, use “0” for NO and “1” for YES.

  • Consistency makes it easier for users of your data. They know what to expect!

Form & Database Design

Not the best format. Too many spelling errors:

1. What asthma medication was prescribed for the child to take for the last 7 days?
   1a. __________________________
   1b. __________________________
   1c. __________________________
   1d. __________________________
   1e. __________________________

Better format with most answers coded:

1. What asthma medication was prescribed for the child to take for the last 7 days?
   ☐ Albuterol
   ☐ Flovent
   ☐ Prednisone
   ☐ Other, describe: __________________________
Form & Database Design

Inconsistent coding for "Missing / Don't know"

1. How often did you exercise this week?
   0 – Did not exercise at all
   1 – 1 time
   2 – 2 times
   3 – 3 times
   4 – 4 times
   5 – 5 or more times
   6 – Missing / Don’t know

2. How many hours did you sleep last night?
   1 – less than 4 hours
   2 – 4 to 6 hours
   3 – 6 to 8 hours
   4 – Missing / Don’t know

In this example, I would recommend making the Missing / Don’t know a consistent value such as "9."

Form & Database Design

• Recommendations for fields and database design
  – Make codes for missing and refused data fields.
  – Consider requiring that all fields are entered unless they are conditional on another question.
  – Use multiple identifiers on each form as a cross check.
  – Code your responses whenever possible and be consistent in your code assignments.
  – Do not try to store more than 1 piece of information in a field.

Form & Database Design

If you store multiple pieces of data in a single field, it will be difficult to analyze.

<table>
<thead>
<tr>
<th>ID</th>
<th>Grade</th>
<th>Temperature</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>3</td>
<td>slightly elevated</td>
<td>round</td>
</tr>
<tr>
<td>1002</td>
<td>2</td>
<td>normal</td>
<td>oval</td>
</tr>
</tbody>
</table>

Storing the data in separate fields will make it much easier to analyze.

<table>
<thead>
<tr>
<th>ID</th>
<th>Grade</th>
<th>Separation</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>3</td>
<td>slightly elevated</td>
<td>round</td>
</tr>
<tr>
<td>1002</td>
<td>2</td>
<td>normal</td>
<td>oval</td>
</tr>
</tbody>
</table>
Field Naming

• Names should be meaningful
  – Meaningful names reduce reliance on labels and codebooks
    Example:
    “1. In the last week, has your child had a fever?”
    – Non-descriptive field name: Q1
    – More Descriptive field name: child_fever
  – Can be a challenge if there are many similar questions
    Example
    “1. In the last week, has your child had a fever?” (c_fvr)
    “2. In the last week, has your child had a fever with cough?” (c_fvr_cgh)
    “3. In the last week, has your child had a fever with cough and diarrhea?” (c_fvr_cgh_dia)

Field Naming

• Names should be reasonable length
  – A general rule is 10-15 characters
    Name Examples:
    • Too long: child_fever_in_last_week
    • Too short: flw
    • Just right: child_fvr_lstwk
  – Too many characters is cumbersome to use and some programs will truncate name when displayed. Better to keep name short and apply a text label.

Stata Example:

<table>
<thead>
<tr>
<th>Full name</th>
<th>Displayed as</th>
</tr>
</thead>
<tbody>
<tr>
<td>child_fever_in_last_week</td>
<td>child_fever</td>
</tr>
<tr>
<td>child_fever_cough_in_last_week</td>
<td>c<del>cough_in</del>k</td>
</tr>
<tr>
<td>child_fever_in_last_2_week</td>
<td>child<del>2</del>week</td>
</tr>
</tbody>
</table>

Field Naming – More Meaningful Names

<table>
<thead>
<tr>
<th>Child Visit form</th>
<th>Age in Years</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1800</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>73.5</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1800</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>76.3</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1800</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>76.0</td>
<td>8.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1800</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>76.3</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1800</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>76.0</td>
<td>8.2</td>
<td></td>
</tr>
</tbody>
</table>
Field Naming – More Meaningful Names

Field Naming

- Use prefixes and suffixes when appropriate
  - Prefix example: A prefix indicating the form name
    
    | Field | Example |
    |-------|---------|
    | _bf_  | lastwk_fever |
  - Suffix example: A suffix can indicate a time period for longitudinal data
    
    | Field | Example |
    |-------|---------|
    | lastwk_fever_12 | |
  - The suffix will help with reshaping data from wide to long format
  - Prefix/suffix can also help with extracting subsets of fields.

Field Naming – Suffix Useful for Reshaping

Field Naming

- Use prefixes and suffixes when appropriate
  - Prefix example: A prefix indicating the form name
    
    | Field | Example |
    |-------|---------|
    | _bf_  | lastwk_fever |
  - Suffix example: A suffix can indicate a time period for longitudinal data
    
    | Field | Example |
    |-------|---------|
    | lastwk_fever_12 | |
  - The suffix will help with reshaping data from wide to long format
  - Prefix/suffix can also help with extracting subsets of fields.
**Field Naming**

- Avoid blanks in names: They will be problematic!
  - Example:  
    "first name" is better renamed as "firstname" or "first_name"

- Decide on the character convention:
  - all lower case
  - ALL UPPER CASE
  - CamelCase
  - use underscores

- Be consistent in choices!
  - Consistent choices will make it much easier for those who are new to your dataset to anticipate the structure of the data.

**Codebooks**

- The codebook describes the data fields on each form and their codings.
- This document serves as a map to anyone working with the data.
- Log at end of codebook is a good place to list changes that occur during the study related to that form.

**Codebook sample**

- The codebook header should always have the date of change so that a user can tell if they have the current version.
- The Notes column is where field-specific change notes should be listed.
• The last page of the codebook should have dated and signed notes.
• Anyone working with this dataset should be able to look at the codebook and clearly understand any issues or changes that occurred with the data.

Programs like REDCap use a defined template as a codebook.
Because file gets read into REDCap, codebook changes and notes have to be managed separately.

• Data quality strategies
  – Include validation checks during data entry
  – Set fields as “required” when appropriate
  – Double-entry of records
    • The amount of double entry depends entirely on the study environment
    • Some studies double enter 100% of all data
    • Other studies enter a random, sub-sample
  – Develop data programs to review records
    • Write programs that check for errors

• Audit trails should be created for changes
  – On paper, create a standard process of change and initialing
  – If electronic entry, then application should record changes made.
Data Security and Handling

• Data Security
  – Protect the subjects!
    • They have consented to data collection.
  – Protect yourself and your institution:
    • IRB requires that you follow best practices
    • You can be liable in the event of a data breach
  – De-identify the data
    • When data is being stored or being prepared for distribution, identifiers should be removed
    • Datasets with identifiers should be stored in a secured environment with limited access — Not carried around on portable devices
    • Identifiers are usually not needed for analysis
    • See next slide for HIPAA list of 18 identifiers

HIPAA Protected Health Information (PHI):
List of 18 Identifiers

1. Names
2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Phone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including fingerprints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data)

HIPAA Breaches >500 since 2009

<table>
<thead>
<tr>
<th>Breach Type</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hacking/IT incident</td>
<td>46</td>
<td>14</td>
</tr>
<tr>
<td>Improper Disposal / Pasen</td>
<td>75</td>
<td>23</td>
</tr>
<tr>
<td>Lost / Stolen Computer/Server</td>
<td>41</td>
<td>13</td>
</tr>
<tr>
<td>Lost/stolen media or portable electronic devices</td>
<td>47</td>
<td>15</td>
</tr>
<tr>
<td>Lost/Stolen Laptops</td>
<td>77</td>
<td>25</td>
</tr>
<tr>
<td>Unauthorized access</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Email</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>18</td>
<td>6</td>
</tr>
</tbody>
</table>
| TOTAL                                           | 314    | 100%

http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrules/coveredentities.html
Data Security and Handling

Laptop Encryption Options

• Mac’s
  – Lion: use FileVault2, whole disk encryption
  – Pre-Lion: use FileVault or TrueCrypt folder encryption
• Windows XP – Checkpoint encryption through Hopkins (often pre-installed) or TrueCrypt FDE
• Windows7 – (Enterprise or Ultimate) MS Bitlocker or TrueCrypt FDE

Data Security and Handling

• Encrypting your data
  – It is best practice to encrypt any identifiable or sensitive data
  • Depends on the data management system – You may not have direct control of the storage device
  • If the data is on a system or device that you control, you can encrypt the data using the operating system or software such as TrueCrypt which is an open source program.
  • TrueCrypt can encrypt entire drives as well as create encrypted areas on a drive.
  • Many hard drives now have encryption built-in
  – To send sensitive data via email
  • Use software such as 7-zip which can create strong encryption zip files.
  • Do not send the password for the encrypted file to the same email address!
  • A better choice is a phone call, text, or sending the password to a separate email address
  • Flash drives
  • For safely transporting data via flash drive, you should create an encrypted area.
  • When buying, check if the flash drive has encryption built-in.
  • Otherwise, you can create an encrypted area on the drive using TrueCrypt or similar software.

Data Security and Handling

• Backing up your data
  – Create a regular backup schedule
  – Backup often and develop a plan for managing the backups
  – Test your backup to make sure that it works and that you know what to expect. Periodically try to restore some files to test the backup
  – If you are responsible for your own backups, consider:
    – Online backup using services like CrashPlan or Carbonite
    – External hard drive with encryption
  – Flash drives should NOT be a primary backup method. When they fail, they are rarely recoverable.

• The differences between backup services and sync services is blurring
  – Services like Dropbox will allow you to save versions and recover deleted files
  – Backup services will not have storage quotas. Typically, unlimited backups for x number of computers or mobile devices.
Data Security and Handling

Privacy & Security Contacts at JHU

- hipaa@jhmi.edu
- network.security@jhmi.edu
- IT Help Desk 410-735-4357
- Darren Lacey (dll@jhu.edu)
  - Chief Information Security Officer

Data Transfer Methods

- Data transfer options
- Email is not always an ideal method because of file size or security issues
  - JHU offers JSHARE (my.jhu.edu)
    - Web-based, 5GB of space
  - Can grant access to non-Hopkins users
  - Dropbox or SpiderOak
    - Web-based, 2GB
    - Allows sharing to collaborate and synchronization
    - Data is encrypted on Dropbox site BUT site administrators can access your data
    - Add a level of security by encrypting files BEFORE posting to Dropbox.
    - SpiderOak has encryption that you control. They cannot access your data.
    - JHU alternative supposed to be coming but still not available.
  - SharePoint
    - Allows for file sharing, discussions, announcements, calendars
    - Integrates closely with Microsoft Office
    - Can be challenging to use with external collaborators

Data Conversion

- Most programs can export data to a common format
  - Comma Separated Values (CSV) is a text format that can be used by most programs.
  - eXtensible Markup Language (XML) a text format that is structured with each data element surround by tags. Can become quite large because of the overhead of the tags.
  - Microsoft Excel (.xls and .xlsx)

- Stat/Transfer
  - Very useful if you frequently convert data between programs
  - Both a graphic interface and a command line interface
  - Command line interface can process many files quickly
  - Student pricing available from Stata through GradPlan
Data Conversion

Stat/Transfer Version 11 supports the following file formats:

- 1-2-3
- Access (Windows version only)
- ASCII - Delimited
- ASCII - Fixed Format
- dBASE and compatible formats
- Data Documentation Initiative (DDI) Schemas
- Excel
- Epi Info
- FoxPro
- Gauss
- HTML Tables (write only)
- JMP
- LIMDEP
- Minitab
- Minim
- NLOGIT
- OpenDocument Spreadsheets
- Paradox
- Quattro Pro
- R
- RATS
- SAS Data Files
- SAS CPORT (read-only)
- SAS Transport Files
- S-PLUS
- SPSS Data Files
- SPSS Portable
- Stata
- Stat/Transfer Version 11 supports the following file formats.
- Statistica (Windows version only)
- SYSTAT
- Triple-S
- Tableau

Contact Information

- Andre Hackman
  Research Associate, Biostatistics Center
  Department of Biostatistics
  Johns Hopkins Bloomberg School of Public Health
  615 North Wolfe Street, Room E3146
  Baltimore, MD 21205
  Phone: (410)614-5646
  Fax: (410)955-0958
  email: ahackman@jhsph.edu
Resources mentioned

- **Online backup**
  - Crashplan: [www.crashplan.com](http://www.crashplan.com)
  - Carbonite: [www.carbonite.com](http://www.carbonite.com)

- **Encryption and Compression**
  - TrueCrypt: [www.truecrypt.org](http://www.truecrypt.org)
  - 7-Zip: [www.7-zip.org](http://www.7-zip.org)

- **Data Conversion and Cleaning**
  - Stat/Transfer: [www.stattransfer.com](http://www.stattransfer.com)

- **Data Transfer**
  - Dropbox: [www.dropbox.com](http://www.dropbox.com)
  - Spider Oak: [www.spideroak.com](http://www.spideroak.com)