Data Management Solutions: Where is the Grass Greener?

Diane Levy, MS
diane.levy@columbia.edu
April 22, 2014
My Background

• Pre-Columbia
  – Community Health Education
  – Consulting
  – Training

• At Columbia (2001)
  – Data manager in Biostatistics (MSPH)
  – Created/taught P8180
Data Management: [Finally] Accepted as Necessary by Investigators

• Columbia has put security requirements into place
  – Investigators must now indicate which secure server they will be using
  – Must be certified by Columbia’s IT department
    – https://secure.cumc.columbia.edu/cumcit/secure/security/scp_systems.html

• Funding agencies require a data management plan
Where is the grass greener?

• How do you decide which tools?
  – Clinical and Translational Science Award (CTSA)
    • Request a 1 hour consultation for data management
      – Best use: when writing a proposal
      – Alternatively: after receiving funding (but then you’re limited by the funding you received and \textit{maybe by the people you already hired})
      – Must be faculty and have a UNI and an eRA commons ID
      – \url{http://www.mailman.columbia.edu/academic-departments/biostatistics/consultation-service}
What are the tools?

• What you hear:
  – Flat files
  – Relational databases
  – REDCap
Flat Structures: Generally Worst Choice

- Excel
- Statistical Software Packages
  - SAS
  - SPSS
  - R

The purpose of these packages is for data analysis. Not for data collection and data management.
Why?

- **Excel: tries to be too helpful**
  - Assumes all data to be dates or numbers
    - If your data contains text that resembles a date, Excel assumes date
    - If your data contains text that resemble a number, Excel assumes a number
Opening a text file of gene names in Excel

These are now dates
RIKEN Clone Identifiers

• In the form of: nnnnnnnEnn (where n denotes a digit)
• Identifiers are comprised of:
  – Serial number of the plate that contains the library
  – Information on plate status
  – Address of the clone
• 2310009E13 would be converted irreversibly to the floating-point number "2.31E+19."
Example of Excel Assuming a Number

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Pasted as text</td>
<td>Pasted without formatting</td>
</tr>
<tr>
<td>2 2310009E13</td>
<td>2.31E+19</td>
</tr>
</tbody>
</table>

10
All Flat Files: Other Issues

• Structure of the data
  – Too wide
  VS
  – Mismatched records
Example of a Study

• Weight Control Study
  – Initial Visit
    • Collection of demographic data
    • First weigh in
  – Subsequent:
    • Weekly weigh in visits for x number of months
Too Wide For Ease of Use

In 5 Visits: Up to column Y
Difficult to add a new variable

• What if you needed to add a comments column to capture notes for each visit?

And difficult to retrieve data when creating a dataset (TBD later!)
Using Multiple Sheets

**Demographic Sheet:** One row per person

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>StudyID</td>
<td>DOB</td>
<td>Sex</td>
<td>Height</td>
<td>Address1</td>
<td>Address2</td>
<td>Zipcode</td>
<td>HomePhone</td>
<td>MobilePhone</td>
</tr>
<tr>
<td>2</td>
<td>123</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>345</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>456</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Visit Sheet:** One row per visit

(Requires unique combination of studyID and DOV)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>StudyID</td>
<td>DOV</td>
<td>Weight</td>
<td>WaistCircum</td>
</tr>
<tr>
<td>2</td>
<td>123</td>
<td>1/1/2000</td>
<td>192</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>345</td>
<td>1/14/2000</td>
<td>256</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>123</td>
<td>2/1/2000</td>
<td>189</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>456</td>
<td>1/23/2000</td>
<td>199</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>123</td>
<td>2/9/2000</td>
<td>188</td>
<td></td>
</tr>
</tbody>
</table>
Biggest Problem using Multiple Sheets

• Disconnected Records
  – Entering a visit with an incorrect Study ID
Limited Ability to Create User Friendly Interfaces

• Skip patterns are almost impossible to implement
• Validation is difficult to implement
  – Especially if validating one column based on info in another column
    • (e.g.: weight or height based on sex)
• Difficult to satisfy requirement of unique values
  – Especially if the combination of 2 columns makes a record unique
• Maintaining a regular structure requires discipline (no cutting and pasting over formats)
Relational Databases

- Access
- FileMaker
- FoxPro
- SQL Server
- MySql
- Oracle
- SIR
Can Accommodate Any Data

• Data are stored in consistent tables
  – Design is based on set theory so that:
    • Each table has a subject
      – Each record is a description of that subject
    • Each record in each table is unique
  – Additional tables can be added at any time
    • No redesign of database necessary
Example of Weigh-In Database
Add Current Medications (with lookup)
Relationships

• Represented by a line between 2 tables
• Programmed into the database
  – Automatically enters the StudyID into the visits table
    (*Access is automatic; otherwise programmable*)
    • Automatically updates the StudyID in the visits table if user
      changes it in the demographic table
  – Prevents orphaned records
    • Can’t add a visit for a person who isn’t in the demographic
      table
    • Can’t delete a person from the demographic table if visits
      exist
Strong Data Type Enforcement

• Decimals
• Integers
• Dates/times
• Text
• Memo
Enables User Friendly Interfaces

• Access: Has user interface programming ability built in

• MySQL, SQL Server, Oracle (and others)
  – Allow web interfaces to be built/programmed
SQL Query Language

• Universal querying language
• English based structure
• Allows easy querying of well designed and structured relational databases
  – Datasets for analysis can be compiled with minimal effort
Simplicity of SQL

To find all the men born on or after January 1, 1982:

```
Select * from tblDemographic where Sex=1 and DOB >= '1/1/1982'
```
Relational Database Data Storage

Want to retrieve all subjectIDs of people who had visits in 2013
### Peek at the Subject Visit Data

**SELECT distinct subjectID FROM tblSubjectVisit where**

DateVisit\(\geq\) ‘1/1/2013’ and

DateVisit\(\leq\) ‘12/31/2013’

**Query never changes even when more people or more visits for a person are added**

<table>
<thead>
<tr>
<th>SubjectID</th>
<th>DateVisit</th>
<th>VisitType</th>
<th>Weight</th>
<th>WaistCircum</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7/4/2009</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11/5/2009</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3/3/2010</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7/8/2010</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11/9/2010</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6/26/2009</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10/28/2009</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2/28/2010</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5/4/2010</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1/11/2009</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5/20/2009</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9/3/2009</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9/1/2010</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4/27/2009</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8/29/2009</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>12/26/2009</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5/5/2009</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9/2/2009</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1/6/2010</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5/5/2010</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10/31/2009</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Poorly Designed Data: Repeated Columns

Difficult to apply SQL to repeated column design used with flat files or incorrect table design

List the studyID of each participant who had a weigh in visit between 1/1/2013 and 12/31/2013:

Select StudyID from tblDemographic
where (DOV1 >= 1/1/2013 and DOV1 <= 12/31/2013)
or (DOV2 >= 1/1/2013 and DOV2 <= 12/31/2013)
or (DOV3 >= 1/1/2013 and DOV3 <= 12/31/2013)
or (DOV4 >= 1/1/2013 and DOV4 <= 12/31/2013)
or (DOV5 >= 1/1/2013 and DOV5 <= 12/31/2013)
Query will Grow

• The more visits a person can have (maximum number of visits for the study), the longer that query will have to be
SELECT Demographic.SubjectID, Sex, DOB, Count(SubjectID) AS NumberOfVisits
FROM LEFT JOIN SubjectVisit ON Demographic.SubjectID = SubjectVisit.SubjectID
GROUP BY Demographic.SubjectID, Sex, DOB
Produces this Dataset

Query results are a combination of demographic data and aggregated visit data

Can be exported to:
• Text
• Excel
• SAS
Access

• Access must be installed on an encrypted computer *(for use by 1 user at a time only)*
  - or -

• Can be installed on a secure server
Access: Pros

• Has front end built in
• Easy to implement relational structure
  – Fast development
Concerns about using Access

• It is a file (1 file) and can be copied and moved
• Although it allows for password encryption, it can be skipped in design or broken
• Doesn’t allow for different levels of security
• Can be used incorrectly
• Data can get too large and it can become corrupted
Other Options: Industrial Databases

- MySQL
- SQL Server
- Oracle
Industrial Databases: Pros

• Powerful
• Sophisticated
• Flexible
• Complex
• Enhanced security
  – Different security levels for different users
• Can be programmed onto the web
  – Easy collaboration (e.g.: multi-center studies)
Industrial Databases: Concerns

• Expensive
• Require server installation
• Requires working with a data programmer
• Longer development time to create data entry screens (usually web based)
All Databases: Development Steps

• Start with data structure first
  – Tables are designed to accommodate all data needed for the project

• Data entry screens follow
  – Duplicate of investigators’ data collection tools
    • Data are entered into data entry screens and parsed (behind the scenes) into the tables

• These steps ensure ease of data going in cleanly and being retrieved easily
Sample Data Entry Screen: Weigh In Study

1. Infinite number of visits
2. All automatically assigned correct subject ID
Sample of An Access Screen: List of Participants

Preloaded Study IDs

<table>
<thead>
<tr>
<th>EDC ID#</th>
<th>Version:</th>
<th>Subject Initials</th>
<th>Interviewer</th>
<th>Interview Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED001</td>
<td>24 to 36 mos</td>
<td>CN</td>
<td>JRC</td>
<td>3/9/2012</td>
</tr>
<tr>
<td>ED001</td>
<td>36 to 48 mos</td>
<td>CN</td>
<td>JRC</td>
<td>3/6/2013</td>
</tr>
<tr>
<td>ED002</td>
<td>24 to 36 mos</td>
<td>EP</td>
<td>JRC</td>
<td>2/28/2013</td>
</tr>
<tr>
<td>ED002</td>
<td>36 to 48 mos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED003</td>
<td>36 to 48 mos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED004</td>
<td>24 to 36 mos</td>
<td>YNF</td>
<td>JRC</td>
<td>4/5/2012</td>
</tr>
<tr>
<td>ED004</td>
<td>36 to 48 mos</td>
<td>RD</td>
<td>YNF</td>
<td>3/21/2013</td>
</tr>
<tr>
<td>ED005</td>
<td>36 to 48 mos</td>
<td>JE</td>
<td>YNF</td>
<td>3/27/2013</td>
</tr>
<tr>
<td>ED006</td>
<td>36 to 48 mos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED007</td>
<td>36 to 48 mos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED008</td>
<td>36 to 48 mos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED009</td>
<td>36 to 48 mos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED010</td>
<td>24 to 36 mos</td>
<td>BRG</td>
<td>JRC</td>
<td>4/21/2012</td>
</tr>
<tr>
<td>ED010</td>
<td>36 to 48 mos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED011</td>
<td>36 to 48 mos</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Page 1: Duplicate of Paper Questionnaire
Page 7: Data are Saved into Separate Tables

1. Saved in Early Intervention table
2. Data entry is only allowed if previous question’s answer is YES
Example of Web Based Screen

HEALTH EFFECTS OF ARSENIC

13. Have you ever noticed that your water causes your bathtub, toilet or sinks to turn black?
- Yes
- No

14. When the child is home, does he/she drink the tap water from this well? (Include all drinks mixed with water such as tea, hot chocolate, kool-aid, etc.)
- Yes
- No

15. When your child is home, does he/she have any other source of drinking water?

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
<th>Add</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bottled Water</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Filtered Water</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Edt Detaile 2</td>
<td>under sink filter (lead micro organism)</td>
<td></td>
</tr>
</tbody>
</table>

- Answered yes but no fixtures on questionnaire
- Answered yes but no sources on questionnaire

16. If the child has an alternative source of drinking water, how often does the child use the alternative source?
- All the time (100%)
- Often (75%)
- Occasionally (50%)
- Rarely (25%)
- Never

Other sources: saved to another table
Summary

• Table design is first in process
• Data entry screens (with endless validation) follow
• Data are stored in various tables in an organized manner (relational structure)
  • Users are oblivious
• Data set building (retrieval) using SQL is easy
• Datasets themselves are stored
• When using server based database technology, the entire database cannot be copied to a flashdrive or other computer
Welcome to REDCap!

REDCap is a secure, web-based application for building and managing online surveys and databases. Using REDCap's stream-lined process for rapidly developing projects, you may create and design projects using 1) the online method from your web browser using the Online Designer; and/or 2) the offline method by constructing a 'data dictionary' template file in Microsoft Excel, which can be later uploaded into REDCap. Both surveys and databases (or a mixture of the two) can be built using these methods.

REDCap provides automated export procedures for seamless data downloads to Excel and common statistical packages (SPSS, SAS, Stata, R), as well as a built-in project calendar, a scheduling module, ad hoc reporting tools, and advanced features, such as branching logic, file uploading, and calculated fields.
REDCap: What is it?

• Developed by Vanderbilt University’s Informatics Department
  – Funded by the NIH

• “Free” download for researchers
  – Not to be used for non-research purposes
    • Never in a for-profit environment
      – Including clinical practices
More about “Free”

• Must be installed on a secure server
  – Deemed secure by CUIT
• Servers must be maintained by network administrators
  – Not programmers
• Must be upgraded several times a year
• A REDCap administrator must be the one to:
  – Create new projects
  – Create new IDs
  – Approve significant changes to a project
CTSA and REDCap

- CTSA supports the use of REDCap at CUMC
  - Biostatistics is funded to maintain it
  - Anyone eligible for CTSA support can request:
    - a consultation (required if REDCap is to be used)
    - a new project
    - necessary IDs
- Users of our REDCap installation must cite:
  - the CTSA in their publications
  - REDCap in their publications
Programming REDCap

• Quick learning curve
• Training videos
• Programming concentrates on the user interface
  – Online designer
  – Spreadsheet data dictionary
    • User learns how to do this after using data designer and downloading into spreadsheet
Programming REDCap

• REDCap starts with the user interface
• It builds a flat table structure into MySQL
  – The data are not stored in a relational structure
  – Downloaded data look much the same as a large repeated column spreadsheet
  – Creation of datasets (using aggregation) is more difficult
    • Can’t easily run SQL against the data to create a dataset for analysis
REDCap: Creating Datasets

• Investigators export to a specified format (e.g.: Excel, SAS, SPSS)
  – Dataset creation occurs in that software package
  – That snapshot of the data cannot be saved in REDCap
    • Finding the dataset that was used for a specific publication is harder
    • Recreation of the dataset in REDCap may be impossible
      – **Important**: the data in REDCap itself are constantly evolving!
        (Not static)
REDCap: Building User Interface

• Can create screens that look similar to questionnaires
  – Vertical listing of questions only
    • Not possible to have 2 questions on 1 line
• Use of the following allow for consistent and (somewhat*) clean data:
  – Coding (drop down choices)
  – Data types (only dates allowed in date variables)
  – Skip patterns
*Validation is limited*

- Validation rules cannot be based on values entered in previous fields
  - i.e.: weights or heights based on sex
- Required and range checking are only suggestions
  - System will warn user if data are being left out or entered out of range
    - But will allow the data to be entered and record to be saved
- Data cleaning will be necessary
Warning Message

User can save the record without entering required variables
Difficult to Program Repetition

- Medications
- Allergies
- Information about children
- Adverse events
- Anything unscheduled
Example of Capturing Medications
Medications

Maximum of 3 medications
Data Entry for Multiple Medications

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are you taking medications?</td>
<td>[Yes]</td>
</tr>
<tr>
<td>How many medications are you taking?</td>
<td>[2]</td>
</tr>
<tr>
<td>Name of first medication</td>
<td>[Blank]</td>
</tr>
<tr>
<td>Dose of first medication</td>
<td>[Blank]</td>
</tr>
<tr>
<td>Reason for taking first medication</td>
<td>[Blank]</td>
</tr>
<tr>
<td>Name of 2nd medication</td>
<td>[Blank]</td>
</tr>
<tr>
<td>Dose of 2nd medication</td>
<td>[Blank]</td>
</tr>
<tr>
<td>Reason for taking 2nd medication</td>
<td>[Blank]</td>
</tr>
</tbody>
</table>

Notes:
- Enter a number from 1 to 3
- Enter the dose including units and # times/day
Problem

• If more than 3 medications, the application would have to be modified in order to accept another medication

Very different from the relational model where a grid would be available to enter as many medications as needed
The Medications: Behind the scenes (table storage)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>participantid</td>
<td>redcap_event_name</td>
<td>meds</td>
<td>nummeds</td>
<td>med1</td>
<td>dose1</td>
<td>medreason1</td>
<td>med2</td>
<td>dose2</td>
<td>medreason2</td>
<td>med3</td>
<td>dose3</td>
</tr>
<tr>
<td>1</td>
<td>interview_arm_1</td>
<td>1</td>
<td>2</td>
<td>Motrin</td>
<td>200 mg every 4-6 hours</td>
<td>abdominal pain</td>
<td>Lipitor</td>
<td>20 mg</td>
<td>high cholesterol</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>visit_1_arm_1</td>
<td>1</td>
<td>2</td>
<td>amoxicillin</td>
<td>250 mg every 8 hours</td>
<td>sinus infection</td>
<td>Sudafed</td>
<td>30 mg every 6 hours</td>
<td>congestion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>visit_2_arm_1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>interview_arm_1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>visit_1_arm_1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>visit_2_arm_1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>interview_arm_1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Several Issues:
- The more medications, the more columns
  - If demographics were included, sheet would be extremely wide
- Saves columns even when no data (medication 3)
- Very difficult to run SQL against this to aggregate (create dataset)
- There is more information here than requested from REDCap
  - Because study consists of scheduled visits, REDCap exports data with lines saved for those visits
Scheduled Events

• Predictable
  – Therefore can be programmed
  – A tool can be used to create a dashboard
REDCap Version: Weigh In Project
**REDCap Version: Weigh In Project**

**weightcontrol**

<table>
<thead>
<tr>
<th>bloodwork</th>
<th>Share this instrument</th>
</tr>
</thead>
</table>

**Editing existing Record ID 1**

**Event Name:** First Blood Draw

<table>
<thead>
<tr>
<th>Record ID</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>date of blood draw</td>
<td>03-06-2013</td>
</tr>
<tr>
<td>Arsenic</td>
<td>33</td>
</tr>
<tr>
<td>Lead</td>
<td>44</td>
</tr>
<tr>
<td>Selenium</td>
<td>31</td>
</tr>
</tbody>
</table>

**Form Status**

**Complete?** Complete

**Lock this record for this form?**

If locked, no user will be able to edit this record on this form until someone with Lock/Unlock privileges unlocks it.

[Save Record]

[Save and Continue]
Dashboard

Represents 3 programmed forms:
- Demographic Information
- Weigh In Information
- Blood test results

This project scheduled:
- 1 demographic record
- 3 visits
- 2 blood draws
Export: Requesting Specific Variables

### Form: `demographicInfo`
- **Record ID** (`record_id`) (✓)
- **Date of Birth** (`dateOfBirth`) (✓)
- **Gender** (`gender`) (✓)
- **Do you have a dog?** (`dog`) (☐)
- **Weight at registration** (`regWeight`) (✓)

### Form: `Weightin`
- **Date of weigh in** (`dateWeight`) (✓)
- **Age at weigh in** (`weight_age`) (☐)
- **Weight** (`weight`) (✓)
- **Pounds Lost** (`lbslost`) (☐)
- **Waist Circum (inches)** (`waistcircum`) (✓)

### Form: `bloodwork`
- **Date of blood draw** (`blooddate`) (☐)
- **Arsenic** (`arsenic`) (☐)
- **Lead** (`lead`) (☐)
- **Selenium** (`selenium`) (☐)

### Form Status
- **Complete** (`demographicinfo_complete`) (☐)
- **Complete** (`weightin_complete`) (☐)
- **Complete?** (`bloodwork_complete`) (☐)
### Downloaded Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>record_id</td>
<td>date_of_birth</td>
<td>reg_weight</td>
<td>date_of_weight</td>
<td>weight</td>
</tr>
<tr>
<td>1</td>
<td>register_arm_1</td>
<td>5/12/1980</td>
<td>195</td>
<td>3/1/2013</td>
<td>189</td>
</tr>
<tr>
<td>1</td>
<td>first_weigh_in_arm_1</td>
<td>3/1/2013</td>
<td>185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>first_blood_draw_arm_1</td>
<td>4/1/2013</td>
<td>175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>second_weigh_in_arm_1</td>
<td>4/1/2013</td>
<td>175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>second_blood_draw_arm_1</td>
<td>5/1/2013</td>
<td>162</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>third_weigh_in_arm_1</td>
<td>5/1/2013</td>
<td>162</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>register_arm_1</td>
<td>5/13/1986</td>
<td>153</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>first_weigh_in_arm_1</td>
<td>5/22/2013</td>
<td>153</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>second_weigh_in_arm_1</td>
<td>5/24/2013</td>
<td>123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>register_arm_1</td>
<td>1/7/2013</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>first_weigh_in_arm_1</td>
<td>5/1/2013</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>register_arm_1</td>
<td>2/11/1999</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>first_weigh_in_arm_1</td>
<td>5/1/2013</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>register_arm_1</td>
<td>1/7/2013</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>first_weigh_in_arm_1</td>
<td>5/5/2013</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>register_arm_1</td>
<td>1/15/1982</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>register_arm_1</td>
<td>2/20/2013</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>register_arm_1</td>
<td>3/22/2013</td>
<td>195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>register_arm_1</td>
<td>3/22/2013</td>
<td>195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>register_arm_1</td>
<td>5/1/2013</td>
<td>175</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Demographic Data

#### Weigh In Data

#### Space saved for blood data

---

Columbia University Mailman School of Public Health

The Department of Biostatistics
## Request of All Variables

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>record_id</td>
<td>date_of_birth</td>
<td>gender</td>
<td>age</td>
<td>date_of_death</td>
<td>weight</td>
<td>waist_circumference</td>
<td>blood_date</td>
<td>arsenic</td>
<td>lead</td>
<td>selenium</td>
</tr>
<tr>
<td>1</td>
<td>first_weigh_in_arm_1</td>
<td>3/1/2013</td>
<td>185</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>first_blood_draw_arm_1</td>
<td>3/6/2013</td>
<td>33</td>
<td>44</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>second_weigh_in_arm_1</td>
<td>4/1/2013</td>
<td>175</td>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>second_blood_draw_arm_1</td>
<td>4/2/2013</td>
<td>32</td>
<td>41</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>third_weigh_in_arm_1</td>
<td>5/1/2013</td>
<td>152</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>register_arm_1</td>
<td>5/13/1985</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1/1/2012</td>
<td>0.1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>first_weigh_in_arm_1</td>
<td>5/22/2013</td>
<td>153</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>second_weigh_in_arm_1</td>
<td>5/24/2013</td>
<td>123</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>register_arm_1</td>
<td>1/7/2013</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1/2/2012</td>
<td>0.1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>first_weigh_in_arm_1</td>
<td>5/1/2013</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>register_arm_1</td>
<td>2/11/1999</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1/3/2012</td>
<td>0.1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>first_weigh_in_arm_1</td>
<td>5/1/2013</td>
<td>12</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>register_arm_1</td>
<td>1/7/2013</td>
<td>1</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>1/4/2012</td>
<td>0.1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>first_weigh_in_arm_1</td>
<td>5/5/2013</td>
<td>12</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>register_arm_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>register_arm_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>register_arm_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>register_arm_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>register_arm_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>register_arm_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>register_arm_1</td>
<td>1/15/1982</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>first_weigh_in_arm_1</td>
<td>2/20/2013</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>first_blood_draw_arm_1</td>
<td>3/19/2013</td>
<td>22</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>second_weigh_in_arm_1</td>
<td>3/22/2013</td>
<td>195</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>second_blood_draw_arm_1</td>
<td>4/10/2013</td>
<td>12</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>third_weigh_in_arm_1</td>
<td>5/1/2013</td>
<td>175</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
REDCap: Positives

• Excellent for surveys and simple questionnaires
• Set up to be multi-user
  – On the web
• Economical
• Secure
  – Handles multilevel security
• Good reporting tools
• Multiple formats available for downloading
Negatives

• Can’t accommodate unscheduled events
  – Adverse events
  – Unscheduled visits
  – Complex protocols
    • If there’s a death, fill out form x

• Exported data require manipulation for analysis
  – Exported datasets cannot be stored in the system for later retrieval

• More advanced programming requirements necessitate programming expertise
  – Example: lookup drop down boxes
<table>
<thead>
<tr>
<th>Feature</th>
<th>Text File</th>
<th>Spreadsheet</th>
<th>Statistical Software</th>
<th>REUSE®</th>
<th>File-based Databases</th>
<th>Server-based Databases</th>
<th>Critical Trail Management Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of simultaneous users</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>No</td>
<td>~20</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Differential access by user/site</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Web access</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Other remote access (app, ftp, etc.)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Programming expertise required</td>
<td>No</td>
<td>No</td>
<td>Often</td>
<td>Often</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Calendar/scheduling</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Built-in</td>
<td>Custom</td>
<td>Custom</td>
<td>Built-in</td>
</tr>
<tr>
<td>Project management</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Custom</td>
<td>Custom</td>
<td>Custom</td>
<td>Built-in</td>
</tr>
<tr>
<td>Tracks modifications to data (audit trail)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Custom</td>
<td>Custom</td>
<td>Custom</td>
<td>Built-in</td>
</tr>
<tr>
<td>Data stored natively in tables</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Enforces data types</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Enforces value lists</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Enforces value ranges</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Enforces complex validation</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Parent-child/one-to-many structures</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Intra-record logic checks</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Intra/inter-table logic checks</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Simple data querying</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Complex data querying</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Unexpected events (adverse events, unscheduled visits, etc.)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Exports to multiple formats</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Imports from multiple formats</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Security</td>
<td>0+</td>
<td>++</td>
<td>++</td>
<td>++++</td>
<td>++++</td>
<td>++++</td>
<td>++++</td>
</tr>
<tr>
<td>Ease of data entry</td>
<td>0+</td>
<td>++</td>
<td>+</td>
<td>++++</td>
<td>++++</td>
<td>++++</td>
<td>++++</td>
</tr>
</tbody>
</table>
Thank You!

• Please email if you would like a copy of the slides and/or a summary article written by my colleague Richard Buchsbaum

• Diane.Levy@columbia.edu