Comparing Excel, Access and REDCap as Data Management Tools for Human Health Research Data

Sue Lowry (sklowry@umn.edu)
Database Programmer/Analyst
Biostatistical Design and Analysis Center
Clinical and Translational Science Institute
Overview

- Looking at some tools that can be used to manage human health research data
- Benefits of and problems with using Excel
- Benefits of and problems with using Access
- Benefits of and problems with using REDCap
Benefits of Using Excel

- Excel is easy to use and is available to most researchers
- Most researchers are familiar with Excel
- Excel is good for what it is designed for, which Microsoft describes as:
  - Calculations and simple descriptive statistics
  - Creating charts, graphics, and diagrams
  - Organizing lists
  - What-if scenarios
  - Automating repetitive tasks using macros
Problems With Using Excel for Human Health Research Data

➢ Makes it too easy to corrupt or lose your data
### Problems With Using Excel for Human Health Research Data

- **Makes it too easy to corrupt or lose your data**

#### Example 1:
If you accidentally sort only the enrollment date column, that data is no longer matched to the correct records.

<table>
<thead>
<tr>
<th>study_id</th>
<th>enroll</th>
<th>fname</th>
<th>lname</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>04/15/2012</td>
<td>Mary</td>
<td>Lowry</td>
</tr>
<tr>
<td>2</td>
<td>02/03/2012</td>
<td>John</td>
<td>Harris</td>
</tr>
<tr>
<td>3</td>
<td>01/05/2012</td>
<td>Peter</td>
<td>Smith</td>
</tr>
<tr>
<td>4</td>
<td>08/02/2012</td>
<td>Paula</td>
<td>Johnson</td>
</tr>
<tr>
<td>5</td>
<td>05/12/2012</td>
<td>Susan</td>
<td>Jones</td>
</tr>
</tbody>
</table>
Problems With Using Excel for Human Health Research Data

- Makes it too easy to corrupt or lose your data

**Example 2:**
If you have many columns of data, it’s easy to accidentally change the data on the wrong row.

<table>
<thead>
<tr>
<th>study_id</th>
<th>enroll</th>
<th>fname</th>
<th>lname</th>
<th>...</th>
<th>final_visit</th>
<th>flag22</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>04/15/2012</td>
<td>Mary</td>
<td>Lowry</td>
<td></td>
<td>02/15/1975</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>02/03/2012</td>
<td>John</td>
<td>Harris</td>
<td></td>
<td>05/03/1988</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>01/05/2012</td>
<td>Peter</td>
<td>Smith</td>
<td></td>
<td>09/05/1997</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>08/02/2012</td>
<td>Paula</td>
<td>Johnson</td>
<td></td>
<td>02/02/1955</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>05/12/2012</td>
<td>Susan</td>
<td>Jones</td>
<td></td>
<td>09/12/2002</td>
<td>1</td>
</tr>
</tbody>
</table>

**Example 3:**
It’s also too easy to clear data from one or more cells, rows or columns or accidentally change the data in a cell.
Problems With Using Excel for Human Health Research Data

- Makes it too easy to corrupt or lose your data
- Most people don’t use data validation in Excel
Problems With Using Excel for Human Health Research Data

- Most people don’t use data validation in Excel

**Example 1:**
Allows non-numeric characters in a numeric variable
Problems With Using Excel for Human Health Research Data

- Most people don’t use data validation in Excel

**Example 2:** Allows you to enter measurement unit in numeric variable

<table>
<thead>
<tr>
<th>E</th>
<th>F</th>
<th>G</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs_id</td>
<td>Age</td>
<td>Height</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>52</td>
<td>180cm</td>
<td></td>
</tr>
<tr>
<td>124</td>
<td>19</td>
<td>176cm</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>36</td>
<td>169cm</td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>-9</td>
<td>192cm</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>43</td>
<td>-9</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>-9</td>
<td>179cm</td>
<td></td>
</tr>
<tr>
<td>129</td>
<td>71</td>
<td>158cm</td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>29</td>
<td>164cm</td>
<td></td>
</tr>
</tbody>
</table>

**Bad**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs_id</td>
<td>Age</td>
<td>Height_cm</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-----------</td>
</tr>
<tr>
<td>123</td>
<td>52</td>
<td>180</td>
</tr>
<tr>
<td>124</td>
<td>19</td>
<td>176</td>
</tr>
<tr>
<td>125</td>
<td>36</td>
<td>169</td>
</tr>
<tr>
<td>126</td>
<td>-9</td>
<td>192</td>
</tr>
<tr>
<td>127</td>
<td>43</td>
<td>-9</td>
</tr>
<tr>
<td>128</td>
<td>-9</td>
<td>179</td>
</tr>
<tr>
<td>129</td>
<td>71</td>
<td>158</td>
</tr>
<tr>
<td>130</td>
<td>29</td>
<td>164</td>
</tr>
</tbody>
</table>

**Good**
Problems With Using Excel for Human Health Research Data

- Most people don’t use data validation in Excel

**Example 3:**
Allows you to enter the same data in an inconsistent way

<table>
<thead>
<tr>
<th>Obs_id</th>
<th>Age</th>
<th>City</th>
<th>State</th>
<th>Proc_completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>52</td>
<td>Minneapolis</td>
<td>MN</td>
<td>Yes</td>
</tr>
<tr>
<td>124</td>
<td>19 1/2</td>
<td>St. Paul</td>
<td>MN</td>
<td>No</td>
</tr>
<tr>
<td>125</td>
<td>36</td>
<td>Hudson</td>
<td>WI</td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>-9</td>
<td>st paul</td>
<td>MN</td>
<td>Not compl.</td>
</tr>
<tr>
<td>127</td>
<td>5/1/1965</td>
<td>Saint Paul</td>
<td>Minnesota</td>
<td>Aug 15</td>
</tr>
<tr>
<td>128</td>
<td>-9</td>
<td>ST. PAUL</td>
<td>mn</td>
<td>Not</td>
</tr>
<tr>
<td>129</td>
<td>71</td>
<td>St. Paul</td>
<td>Minn.</td>
<td>NO</td>
</tr>
<tr>
<td>130</td>
<td>29</td>
<td>Mpls</td>
<td>MN</td>
<td>Did not complete</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs_id</th>
<th>Age</th>
<th>City</th>
<th>State</th>
<th>Proc_completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>123</td>
<td>52 Minneapolis</td>
<td>MN</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>124</td>
<td>19 St. Paul</td>
<td>MN</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>125</td>
<td>36 Hudson</td>
<td>WI</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>126</td>
<td>-9 St. Paul</td>
<td>MN</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>127</td>
<td>43 St. Paul</td>
<td>MN</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>128</td>
<td>-9 St. Paul</td>
<td>MN</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>129</td>
<td>71 St. Paul</td>
<td>MN</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>130</td>
<td>29 Minneapolis</td>
<td>MN</td>
<td>No</td>
</tr>
</tbody>
</table>

**Bad**

**Good**
Problems With Using Excel for Human Health Research Data

Most people don’t use data validation in Excel

Example 4:
Allows dates to be entered with inconsistent formats
Problems With Using Excel for Human Health Research Data

- Most people don’t use data validation in Excel

Example 5:
People sometimes leave a cell blank or put a quote to indicate that the value is the same as in the row above

![Excel Tables Example](image)
Problems With Using Excel for Human Health Research Data

- Most people don’t use data validation in Excel

**Example 6:**
Missing values are indicated in inconsistent ways (They should be blank or a consistent value outside the valid numeric range)
Problems With Using Excel for Human Health Research Data

- Makes it too easy to corrupt or lose your data
- Most people don’t use data validation in Excel
- Allows you to layout the data any way you want
Problems With Using Excel for Human Health Research Data

- Allows you to layout the data any way you want

**Example 1:**
More than one observation per row

![Excel example](image)
Problems With Using Excel for Human Health Research Data

- Allows you to layout the data any way you want

**Example 2:**
Using color coding to flag certain cases instead of using flag variables
Problems With Using Excel for Human Health Research Data

- Allows you to layout the data any way you want

**Example 3:**
Indicating grouping by placement in spreadsheet instead of using a group variable
Problems With Using Excel for Human Health Research Data

- Makes it too easy to corrupt or lose your data
- Most people don’t use data validation in Excel
- Allows you to layout the data any way you want
- Excel 2003 and earlier versions, which many people are still using, have a limit of 256 columns and 65,536 rows for storing your data.
- It doesn’t sufficiently meet HIPAA requirements for handling protected health information (PHI)
HIPAA requirements for data management

- Secure location
  - PHI data reside on a secure server
- Encryption
  - Data is encrypted
- Authentication
  - Logins and passwords
- Authorization
  - Role-based security
- Audit trail
  - Created by event logging
HIPAA Compliance: Excel

- ✓ Secure location
- ✓ Encryption (manual)
- X Authentication - No
- X Authorization – Not “out of the box”
- X Audit trail - No
Benefits of Using Access

- Many researchers are familiar with Access
- Access is fairly easy to use and is available to many researchers
- Access can validate data as you’re entering it
- Access does not allow you to layout the data any way you want (e.g. multiple observations per row, color coding for flagging)
Problems With Using Access for Human Health Research Data

- It’s easy to accidentally change data without meaning to, which Access automatically saves when you leave a form or record.
- In Access, you can set your own rules for missing values, like adding a “data not collected” checkbox field or indicating in a field label how missing values should be entered, but it doesn’t have a built-in way of indicating missing values.
- Access doesn’t sufficiently meet HIPAA requirements for handling protected health information (PHI).
HIPAA Compliance: Access

- Secure location
- Encryption (manual)
- Authentication - No
- Authorization – Not “out of the box”
- Audit trail - No
Benefits of Using REDCap
(Research Electronic Data Capture)

A secure,
HIPAA-compliant,
web-based application
designed specifically to support data capture for research studies.
Benefits of Using REDCap
(Research Electronic Data Capture)

(A secure, HIPAA-compliant, web-based application designed specifically to support data capture for research studies)

- Easy to use
- Used by many researchers around the world
- Available **free of charge** to UMN researchers
- Makes it easy to validate data as you’re entering it
Benefits of Using REDCap

- Makes it easy to validate data as you’re entering it

**Example 1:** Using field type validation

![Image of form with validation error message]

![Image of date picker]

This value you provided is not an integer. Please try again.
Benefits of Using REDCap

- Makes it easy to validate data as you’re entering it

**Example 2: Using data range validation**
Benefits of Using REDCap

- Makes it easy to validate data as you’re entering it

**Example 3: Using multiple choice**

<table>
<thead>
<tr>
<th>1. Gestational Age Group</th>
<th>0. 28 0/7 - 31 6/7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. 32 0/7 - 35 6/7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Group Assignment</th>
<th>0. Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. LMA Group</td>
</tr>
</tbody>
</table>

**What was the cause of hospitalization?**

- Vascular access related events
- CVD events
- Other

**Date of hospital admission**

- [ ] Lexapro
- [ ] Celexa
- [ ] Prozac
- [ ] Paxil
- [ ] Zoloft

(check all that apply)
Benefits of Using REDCap
(Research Electronic Data Capture)

- Easy to use
- Available free to UMN researchers
- Makes it easy to validate data as you’re entering it
- Does not allow you to layout the data any way you want
- Web-based, remotely accessible
- REDCap allows participants to fill out surveys online
- Study personnel from multiple sites or functions can enter the data for which they are responsible
- All study personnel, regardless of geographic location, have access to all of the study data to which they have been granted access as it is collected
- Can export data to SAS, SPSS, R, Stata, Excel or PDF
- Includes features to support HIPAA compliance
### HIPAA Compliance: REDCap

- Secure location
- Encryption (automatic)
- Authentication
- Authorization
- Audit trail

#### Comprehensive User Rights View

<table>
<thead>
<tr>
<th>User name</th>
<th>Expiration</th>
<th>Calendar</th>
<th>Data Export Tool</th>
<th>Data Import Tool</th>
<th>Data Comparison Tool</th>
<th>Logging</th>
<th>File Repository</th>
<th>User Rights</th>
<th>Data Access Groups</th>
<th>Graphical Data View &amp; Store</th>
<th>Data Quality (create/edit values)</th>
<th>Data Quality (concrete values)</th>
<th>Reports &amp; Report Builder</th>
<th>Role Content</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>never</td>
<td></td>
<td>✓ De-identified</td>
<td>✗</td>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>never</td>
<td></td>
<td>✓</td>
<td>✗</td>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>never</td>
<td></td>
<td>✓</td>
<td>✗</td>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>never</td>
<td></td>
<td>✓</td>
<td>✗</td>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>never</td>
<td></td>
<td>✓</td>
<td>✗</td>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

#### Time / Date | User name | Action | List of Data Changes OR Fields Exported
--- | --- | --- | ---
1:35pm 06/21/2012 | GHJHGH324 | Updated Record 3106 | cs_fno_total = '486'
1:33pm 06/21/2012 | GHJHGH324 | Updated Record 3106 | ew_subject_id = '3106', ew_date = '2012-06-21', ew_cw = '0', ew_cby = '2012-06-21', ew_review(1) = checked, early_withdrawal_complete = '1'
1:32pm 06/21/2012 | GHJHGH324 | Updated Record 3106 | cc_subject_id = '3106', cc_any = 0, cc_review(1) = checked, adverse_event_complete = '1'
1:31pm 06/21/2012 | GHJHGH324 | Updated Record 3106 | pd_subject_id = '3106', pd_any = 0, pd_01_eby = '2012-06-21', pd_review(1) = checked, protocol_deviation_complete = '1'
People don’t always set up the field validation for numbers, dates, and so on. You can, in fact, create an entire REDCap database that is composed of only unvalidated text fields.

In REDCap, you can set your own rules for missing values, like adding a “data not collected” checkbox field or indicating in a field label how missing values should be entered, but REDCap does not yet have a built-in way of indicating missing values.

REDCap’s report builder is pretty limited.
For More Information About REDCap

- Read information on the CTSI website:
  
  http://www.ctsi.umn.edu/research/tools-software/  
  - Register for a REDCap demonstration  
    - First Thursday of the month 2:00-3:30pm  
    - Third Wednesday of the month 10:00-11:30am  
    - Special sessions can be arranged for groups of 6 or more

- Watch videos and read the FAQ page on our REDCap website:
  
  https://redcap.ahc.umn.edu/

- Watch videos and read about REDCap on the project website:
  
  http://www.project-redcap.org/
Acknowledgements:
Some of the information and examples for this presentation were taken from information on the University of Colorado’s website at connect.ucdenver.edu, with permission. Other examples were taken from “DataTips.xls” from Philippe Gaillard in the CTSI’s BDAC group.