Data Management Considerations for Clinical Trials

Brad Pollock, M.P.H., Ph.D.
Department of Public Health Sciences

The UC Davis CTSC receives support from the NIH National Center for Advancing Translational Sciences (award TR001860).
Topics

- Data operations
- Databases
- Software
  - Spreadsheets
  - Database management systems
  - Clinical trials management systems
- Other considerations
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>Case Report Form</td>
</tr>
<tr>
<td>DB</td>
<td>Database</td>
</tr>
<tr>
<td>QC</td>
<td>Quality Control</td>
</tr>
<tr>
<td>DMP</td>
<td>Data Management Plan</td>
</tr>
<tr>
<td>CSR</td>
<td>Clinical Study Report</td>
</tr>
<tr>
<td>DCF</td>
<td>Data Clarification Form</td>
</tr>
</tbody>
</table>
Data Management Overview for Clinical Research
Data Collection Form Design → Data Repository Build → Site Data Entry / Export from EMR

Site Data Entry / Export from EMR → Data Cleaning/Data Quality Monitoring

Core DM Process

Data Exporting → Data Export

Data Export → Database Freeze

Data Export Archive

Statistical Analysis

Include policies and procedures
My Background

• Biostatistics and epidemiology: oncology, HIV, clinical & translational research

• Biostatistics core director:
  ▪ Cancer Center BSR
  ▪ CTSA BERD
  ▪ CTSA Informatics Cores

• National CTSA activities:
  ▪ Chair of the BERD Key Function Committee
  ▪ Co-Chair, Methods & Processes Domain Task Force

• PI, intervention trials and etiology studies
Biostatistics Core Functions

• Design studies
  ▪ Clarify hypotheses and objectives
  ▪ Define endpoints
  ▪ Select study/experimental design
  ▪ Sample size/power calculations
  ▪ Develop analytic plans

• Monitor studies
  ▪ Efficacy/futility
  ▪ Safety

• Analyze studies
  ▪ Statistical analysis
  ▪ Writing reports/manuscripts
Why Talk About Data Management in a Biostatistics Seminar Series?

• You have learned a lot about biostatistics, but for most statisticians, the drudgery and hard work is getting and preparing study data for statistical analysis.

• 90/10 Rule
Clinical and Translational Research

• Purpose of clinical and translational research is to discovery new ways to improve the health of individuals and populations

• We do this by conducting research studies:
  ▪ Hypothesis-generating studies
  ▪ Hypothesis-testing studies*

*includes clinical trials, intervention trials, etc.
Clinical and Translational Research (continued)

• Regardless of type of study, the most eloquently designed study is only as good as its data.

• Strength of evidence depends on complete and valid data:

  Data → Information → Knowledge
Clinical and Translational Research (continued)

• Data completeness and quality are critical for scientific discovery:
  ▪ Good data with a bad design are worthless
  ▪ Bad data with a good design is even worse

• Many investigators armed with an electronic spreadsheet think they have the tools to conduct reproducible clinical/translational research

  Wrong!
Clinical and Translational Research (continued)

- What’s sexier?
  - Statistical methods
  - Data management (DM)

- Data management is easily one of the most overlooked, underappreciated aspects of clinical and translational research

Note: For our discussion, a clinical trial is a specific study design within a range of clinical/translational research study types
Statistical competencies for medical research learners: What is fundamental?

Felicity T. Enders\textsuperscript{1}, Christopher J. Lindsell\textsuperscript{2}, Leah J. Welty\textsuperscript{3}, Emma K. T. Benn\textsuperscript{4}, Susan M. Perkins\textsuperscript{5}, Matthew S. Mayo\textsuperscript{6}, Mohammad H. Rahbar\textsuperscript{7}, Kelley M. Kidwell\textsuperscript{8}, Sally W. Thurston\textsuperscript{9}, Heidi Spratt\textsuperscript{10}, Steven C. Grambow\textsuperscript{11}, Joseph Larson\textsuperscript{1}, Rickey E. Carter\textsuperscript{1}, Brad H. Pollock\textsuperscript{12} and Robert A. Oster\textsuperscript{13}

\textsuperscript{1} Division of Biomedical Statistics & Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
\textsuperscript{2} Department of Emergency Medicine and Center for Clinical and Translational Science and Training, University of Cincinnati, Cincinnati, OH, USA
\textsuperscript{3} Department of Preventive Medicine, Northwestern University, Evanston, IL, USA
\textsuperscript{4} Department of Population Health Science and Policy, Center for Biostatistics, Icahn School of Medicine at Mount Sinai, New York, NY, USA
\textsuperscript{5} Department of Biostatistics, School of Medicine, Indiana University, Indianapolis, IN, USA
\textsuperscript{6} Department of Biostatistics, School of Medicine, University of Kansas Medical Center, Kansas City, KS, USA
\textsuperscript{7} Department of Internal Medicine, McGovern Medical School, Division of Clinical and Translational Sciences, University of Texas Health Science Center at Houston, Houston, TX, USA
\textsuperscript{8} Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
\textsuperscript{9} Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA
\textsuperscript{10} Department of Preventive Medicine and Community Health, The University of Texas Medical Branch, Galveston, TX, USA
\textsuperscript{11} Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA
\textsuperscript{12} Department of Public Health Sciences, University of California, Davis, Davis, CA, USA
\textsuperscript{13} Department of Medicine, Division of Preventive Medicine, University of Alabama at Birmingham, Birmingham, AL, USA

Statistical Competencies

6. Understand the value of **data quality and data management**.

7. Understand the reasons for performing **research that is reproducible from data collection** through publication of results.

9. Distinguish between **variable types** (e.g. continuous, binary, categorical) and understand the implications for selection of appropriate statistical methods. Extensively covered by required coursework.

12. Understand issues relating to generalizability of a study, including sampling methods and the amount and type of **missing data**.

16. Understand the need to address **loss to follow-up**.

21. Understand the purpose of **data and safety monitoring plans**.

DATA MANAGEMENT
What is Data Management?

- The development, execution and supervision of plans, policies, programs and practices that control, protect, deliver, and enhance the value of data and information assets*

*Data Management Association, Data Management Body of Knowledge (DAMA-DMBOK), 2008
Who is Involved in Data Management?

- Subjects
  - Participants
  - Patients

- Investigators
  - Clinicians
  - Research Staff
  - Clinical Staff

- Research IT
  - Analysts
  - Programmers
  - DBAs

- Central IT
  - CIO
  - CRIIO

End-to-End Process
Data Management within the Research Process

Protocol Development

IT Involvement

Data Management Process

Final Statistical Analysis
Data management considerations are beginning to influence the science. Storage and long-term utilization affect the data long after the protocol’s final analysis.
Data Management Elements

• Need to maintain functional, flexible, scalable, cost-efficient set of resources to handle a variety of data:
  - Demographic
  - Clinical/laboratory and -omics
  - Environmental

• Data quality and compliance with regulatory requirements
  - HIPAA, 21 CFR Part 11, FISMA

• Prospective planning for:
  - Long time horizons
    • Environmental Influences on Child Health Outcomes (ECHO)
  - Interoperability and federation
    • OnCore CTMS Enterprise Research with EPIC and REDCap
Database Management Functions

• Database design
  ▪ Data elements
  ▪ Relationships (data model)
  ▪ Access control/security/integrity

• Application development
  ▪ Data capture
  ▪ Data curation
  ▪ Querying
  ▪ Reporting
  ▪ Audit

• Database operations
How Data Are Handled?

- Paper forms (CRFs) and keypunch
- Client-server DBMS and networked DBMS
- Web-front end DBMS
  - Pediatric Oncology Group replaced paper in 1998
    - Web front-end
    - Oracle back-end
- Clinical Trials Management System (CTMS)
Databases

• Data elements
Data Elements

- **Common Data Elements (CDE)**
  - Try to use standards with ontologies
    - Common Terminology Criteria for Adverse Events (CTCAE)
    - Patient-Reported Outcomes Measurement Information System (PROMIS)
    - International Classification of Diseases for Oncology (ICD-O)
  - Data dictionaries
  - Case Report Forms (eCRFs)
  - Map/link to other information systems (biorepository, EHR)

- **Specialized** (study-specific data elements)
Building and Adolescent and Young Adult Oncology Research Database

• Demo
Databases

- Data elements
- Database models
Database Model: Data Relationships

- Three types of relationships:
  - One to One
  - One to Many
  - Many to Many

- The relationships of the data reflect the rules of the system (your protocol) and not all potential possibilities
  - **NOTE:** One of the most expensive things to change once underway
Databases

- Data elements
- Database models
- Validation
  - Part of the data plan, multiple methods
- Curation
  - Goal is to maintain the value of the data over time
  - Organization, annotation, revisions/audit log
  - Reuse, future proofing
Software
“Database Management” Software

Microsoft Excel
Excel Characteristics

• Advantages
  ▪ Easy to work with → Quick start up, low costs
  ▪ Potentially you can force data types

• Disadvantages
  ▪ Easy to work with →
    No requirement to clearly define needs
  ▪ Will “interpret” data entries for you
    • Will not allow you to automatically override
### Examples of Good and Bad Variable Names

<table>
<thead>
<tr>
<th>good name</th>
<th>good alternative</th>
<th>avoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max_temp_C</td>
<td>MaxTemp</td>
<td>Maximum Temp (°C)</td>
</tr>
<tr>
<td>Precipitation_mm</td>
<td>Precipitation</td>
<td>precmm</td>
</tr>
<tr>
<td>Mean_year_growth</td>
<td>MeanYearGrowth</td>
<td>Mean growth/ycar</td>
</tr>
<tr>
<td>sex</td>
<td>sex</td>
<td>M/F</td>
</tr>
<tr>
<td>weight</td>
<td>weight</td>
<td>w.</td>
</tr>
<tr>
<td>cell_type</td>
<td>CellType</td>
<td>Cell type</td>
</tr>
<tr>
<td>Observation_01</td>
<td>first_observation</td>
<td>1st Obs.</td>
</tr>
</tbody>
</table>
A spreadsheet with inconsistent date formats

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Date</td>
<td>Assay date</td>
<td>Weight</td>
</tr>
<tr>
<td>2</td>
<td>12/9/05</td>
<td>54.9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>12/9/05</td>
<td>45.3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>12/6/2005</td>
<td>e</td>
<td>47</td>
</tr>
<tr>
<td>5</td>
<td>e</td>
<td>45.7</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>e</td>
<td>52.9</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1/11/2006</td>
<td>46.1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1/11/2006</td>
<td>38.6</td>
<td></td>
</tr>
</tbody>
</table>
Examples of spreadsheets that violate the “no empty cells” recommendation

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>id</td>
<td>date</td>
<td>glucose</td>
</tr>
<tr>
<td>2</td>
<td>101</td>
<td>2015–06–14</td>
<td>149.3</td>
</tr>
<tr>
<td>3</td>
<td>102</td>
<td></td>
<td>95.3</td>
</tr>
<tr>
<td>4</td>
<td>103</td>
<td>2015–06–18</td>
<td>97.5</td>
</tr>
<tr>
<td>5</td>
<td>104</td>
<td></td>
<td>117.0</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td></td>
<td>108.0</td>
</tr>
<tr>
<td>7</td>
<td>106</td>
<td>2015–06–20</td>
<td>149.0</td>
</tr>
<tr>
<td>8</td>
<td>107</td>
<td></td>
<td>169.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<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>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 min</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>strain</td>
<td>normal</td>
<td>mutant</td>
<td>normal</td>
<td>mutant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>147</td>
<td>139</td>
<td>166</td>
<td>179</td>
<td>334</td>
<td>354</td>
<td>451</td>
<td>474</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>246</td>
<td>240</td>
<td>178</td>
<td>172</td>
<td>514</td>
<td>611</td>
<td>412</td>
<td>447</td>
</tr>
</tbody>
</table>
A tidy version of the above data

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>strain</td>
<td>genotype</td>
<td>min</td>
<td>replicate</td>
<td>response</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>normal</td>
<td>1</td>
<td>1</td>
<td>147</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>normal</td>
<td>1</td>
<td>2</td>
<td>139</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>normal</td>
<td>1</td>
<td>1</td>
<td>246</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>normal</td>
<td>1</td>
<td>2</td>
<td>240</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>mutant</td>
<td>1</td>
<td>1</td>
<td>166</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>mutant</td>
<td>1</td>
<td>2</td>
<td>179</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>mutant</td>
<td>1</td>
<td>1</td>
<td>178</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>mutant</td>
<td>1</td>
<td>2</td>
<td>172</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>normal</td>
<td>5</td>
<td>1</td>
<td>334</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>normal</td>
<td>5</td>
<td>2</td>
<td>354</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>normal</td>
<td>5</td>
<td>1</td>
<td>514</td>
</tr>
<tr>
<td>13</td>
<td>B</td>
<td>normal</td>
<td>5</td>
<td>2</td>
<td>611</td>
</tr>
<tr>
<td>14</td>
<td>A</td>
<td>mutant</td>
<td>5</td>
<td>1</td>
<td>451</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>mutant</td>
<td>5</td>
<td>2</td>
<td>474</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>mutant</td>
<td>5</td>
<td>1</td>
<td>412</td>
</tr>
<tr>
<td>17</td>
<td>B</td>
<td>mutant</td>
<td>5</td>
<td>2</td>
<td>447</td>
</tr>
</tbody>
</table>
Spreadsheets with nonrectangular layouts

<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>1</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td>104</td>
<td>105</td>
</tr>
<tr>
<td>2</td>
<td>sex</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Male</td>
</tr>
<tr>
<td>3</td>
<td>glucose</td>
<td>134.1</td>
<td>120.0</td>
<td>124.8</td>
<td>83.1</td>
</tr>
<tr>
<td>4</td>
<td>insulin</td>
<td>0.60</td>
<td>1.18</td>
<td>1.23</td>
<td>1.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1MIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>B6</td>
<td>146.6</td>
<td>138.6</td>
<td>155.6</td>
<td>166</td>
<td>179.3</td>
</tr>
<tr>
<td>3</td>
<td>BTBR</td>
<td>245.7</td>
<td>240</td>
<td>243.1</td>
<td>177.8</td>
<td>171.6</td>
</tr>
<tr>
<td>4</td>
<td>5MIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Normal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>B6</td>
<td>333.6</td>
<td>353.6</td>
<td>408.8</td>
<td>450.6</td>
<td>474.4</td>
</tr>
<tr>
<td>7</td>
<td>BTBR</td>
<td>514.4</td>
<td>610.6</td>
<td>597.9</td>
<td>412.1</td>
<td>447.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Date</td>
<td>11/3/14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Days on diet</td>
<td>126</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Mouse #</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>sex</td>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>experiment values</td>
<td>mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>control</td>
<td>0.186</td>
<td>0.191</td>
<td>1.081</td>
<td>0.49</td>
<td>0.52</td>
</tr>
<tr>
<td>7</td>
<td>treatment A</td>
<td>7.414</td>
<td>1.468</td>
<td>2.254</td>
<td>3.71</td>
<td>3.23</td>
</tr>
<tr>
<td>8</td>
<td>treatment B</td>
<td>9.811</td>
<td>9.259</td>
<td>11.296</td>
<td>10.12</td>
<td>1.05</td>
</tr>
<tr>
<td>9</td>
<td>fold change values</td>
<td>mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>treatment A</td>
<td>15.26</td>
<td>3.02</td>
<td>4.64</td>
<td>7.64</td>
<td>6.65</td>
</tr>
<tr>
<td>11</td>
<td>treatment B</td>
<td>20.19</td>
<td>19.05</td>
<td>23.24</td>
<td>20.83</td>
<td>2.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GTT date</td>
<td>321</td>
<td>2/9/15</td>
<td>24.5</td>
<td>0</td>
<td>99.2</td>
</tr>
<tr>
<td>2</td>
<td>GTT weight</td>
<td>349.3</td>
<td>0.205</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>time</td>
<td>15</td>
<td>296.1</td>
<td>0.129</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>glucose mg/dl</td>
<td>30</td>
<td>312</td>
<td>0.175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>insulin ng/ml</td>
<td>60</td>
<td>99.9</td>
<td>0.122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>treatment A</td>
<td>120</td>
<td>217.9</td>
<td>lo off curve</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>treatment B</td>
<td>322</td>
<td>2/9/15</td>
<td>18.9</td>
<td>0</td>
<td>186.8</td>
</tr>
<tr>
<td>8</td>
<td>treatment A</td>
<td>5</td>
<td>297.4</td>
<td>2.228</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>treatment B</td>
<td>10</td>
<td>439</td>
<td>2.078</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>treatment A</td>
<td>30</td>
<td>382.3</td>
<td>0.775</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>treatment B</td>
<td>60</td>
<td>232.7</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>fold change values</td>
<td>mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>treatment A</td>
<td>120</td>
<td>260.7</td>
<td>0.523</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>treatment B</td>
<td>323</td>
<td>2/9/15</td>
<td>24.7</td>
<td>0</td>
<td>198.5</td>
</tr>
<tr>
<td>15</td>
<td>fold change values</td>
<td>mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>treatment A</td>
<td>5</td>
<td>530.6</td>
<td>off curve</td>
<td>lo</td>
<td></td>
</tr>
</tbody>
</table>
spreadsheet with a rectangular layout

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>id</td>
<td>sex</td>
<td>glucose</td>
<td>insulin</td>
<td>triglyc</td>
</tr>
<tr>
<td>2</td>
<td>101</td>
<td>Male</td>
<td>134.1</td>
<td>0.60</td>
<td>273.4</td>
</tr>
<tr>
<td>3</td>
<td>102</td>
<td>Female</td>
<td>120.0</td>
<td>1.18</td>
<td>243.6</td>
</tr>
<tr>
<td>4</td>
<td>103</td>
<td>Male</td>
<td>124.8</td>
<td>1.23</td>
<td>297.6</td>
</tr>
<tr>
<td>5</td>
<td>104</td>
<td>Male</td>
<td>83.1</td>
<td>1.16</td>
<td>142.4</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>Male</td>
<td>105.2</td>
<td>0.73</td>
<td>215.7</td>
</tr>
</tbody>
</table>
**Experiment 2:**

**Objective:**
Examine if drugs enhance Antibody and PBMC (ADCC) inhibitory effect on co-injected neuroblastoma cells in NOD/SCID mice.

**Material and Methods:**
2. Day 0: i.v. injection.
3. Day 1: injection of 1x10^6 Cell line XMuco and 0.25x10^6 PBMC in 200ul 25%matrigel medium into both flanks in group 2,3,4.
4. Day 1: injection of 1x10^6 Cell line XMuco and 0.25x10^6 PBMC (NK depleted) in 200ul 25%matrigel medium into both flanks in group 5,6.

**Data Analysis:**

<table>
<thead>
<tr>
<th>Group</th>
<th>Time Points</th>
<th>Left Flank</th>
<th>Right Flank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Colline XMuco alone</td>
<td>Colline XMuco alone</td>
</tr>
<tr>
<td></td>
<td>day 11/4/2010</td>
<td>Total Flux (x10^3)</td>
<td>Total Flux (x10^3)</td>
</tr>
<tr>
<td></td>
<td>day 8 11/10/2010</td>
<td>Total Flux (x10^3)</td>
<td>Total Flux (x10^3)</td>
</tr>
<tr>
<td></td>
<td>day 15 11/16/2010</td>
<td>Total Flux (x10^3)</td>
<td>Total Flux (x10^3)</td>
</tr>
</tbody>
</table>

**Remarks:**
- Complicated structure due to mix of one-time and serial data.
- Data mixed with headers.
- Cannot determine qualifiers for every data point.

**Conclusion:**
Further analysis is required to determine the inhibitory effect of the drugs.
Data Organization in Spreadsheets

Karl W. Broman & Kara H. Woo

To cite this article: Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989
Dangers of Spreadsheets

• The dangers are real
  ▪ European Spreadsheet Risks Interest Group keeps a public archive of spreadsheet “horror stories” (http://www.eusprig.org/horror-stories.htm).

• Many researchers have examined error rates in spreadsheets
  – Panko* (2008) reported that in 13 audits of real-world spreadsheets, an average of 88% contained errors.

• Popular spreadsheet programs also make certain types of errors easy to commit and difficult to rectify.
  ▪ Excel converts some gene names to dates and stores dates differently between operating systems, causing problems in downstream analyses (Zeeberg et al. 2004; Woo 2014).

*http://panko.shidler.hawaii.edu/SSR/Mypapers/whatknow.htm
• Researchers who use spreadsheets should be aware of these common errors and design spreadsheets that are tidy, consistent, and as resistant to mistakes as possible.
“Database Management”
Software
<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>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Variable / Field Name</td>
<td>Form Name</td>
<td>Section Header</td>
<td>Field Type</td>
<td>Field Label</td>
<td>Choices, Calculations, OR Slider Labels</td>
<td>Field Note</td>
<td>Text Validation Type OR Show Slider Number</td>
<td>Text Validation Min</td>
<td>Text Validation Max</td>
</tr>
<tr>
<td>2</td>
<td>record_id</td>
<td>header</td>
<td>text</td>
<td>text</td>
<td>Record ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>siteid</td>
<td>header</td>
<td>text</td>
<td>text</td>
<td>Site ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>pid</td>
<td>header</td>
<td>text</td>
<td>text</td>
<td>Participant ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>first_initial</td>
<td>header</td>
<td>text</td>
<td>text</td>
<td>First Initial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>middle_initial</td>
<td>header</td>
<td>text</td>
<td>text</td>
<td>Middle Initial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>last_initial</td>
<td>header</td>
<td>text</td>
<td>text</td>
<td>Last Initial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>stratum</td>
<td>eligibility</td>
<td>Eligibility</td>
<td>radio</td>
<td>IMPACT Stratum</td>
<td></td>
<td>0, Helen DeVos Children’s Hospital at Spectrum Health</td>
<td>1, BI-LO Charities Children’s Cancer Center</td>
<td>2, Blank Children’s Hospital</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>dx_primary</td>
<td>eligibility</td>
<td>text</td>
<td>text</td>
<td>Primary cancer diagnosis ICD-O code</td>
<td></td>
<td>ICD-O Morphology</td>
<td>Data entry check: Patient must be 0-25.99 years at enrollment</td>
<td>DATE_MDY</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>dx_date</td>
<td>eligibility</td>
<td>date</td>
<td>Date of diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>dx_age_calc</td>
<td>eligibility</td>
<td>calc</td>
<td>Age (years) at diagnosis</td>
<td></td>
<td>round(datediff(dob,dx_date),y),0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>dx_age_confirm</td>
<td>eligibility</td>
<td>radio</td>
<td>Was the patient in the indicated age range at time of diagnosis?</td>
<td></td>
<td>0, Yes</td>
<td>1, No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To be eligible, Patient must have a new diagnosis of cancer or relapsed cancer with an intent to administer chemotherapy and must be within 30 days of starting chemotherapy. (The first day that chemotherapy was administered will be day one. Patients will be eligible for enrollment during the 30 calendar days following day zero).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To be eligible, The patient or a parent/guardian must have receptive and expressive language skills in English or Spanish since the assessment instruments are available in these languages only.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This page allows you to build and customize your data collection instruments one field at a time. You may add new fields or edit existing ones. New fields may be added by clicking the Add Field Here button. You can begin editing an existing field by clicking on the Edit icon. If you decide that you do not want to keep a field, you can simply delete it by clicking on the Delete icon. To reorder the fields, simply drag and drop a field to a different position within the form below. NOTE: While in development status, all field changes will take effect immediately in real time.

Current instrument: Demographics

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Required</th>
<th>Label</th>
<th>Value</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study ID</td>
<td>Text</td>
<td>X</td>
<td>study_id</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Name</td>
<td>Text</td>
<td></td>
<td>first_name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last Name</td>
<td>Text</td>
<td></td>
<td>last_name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date of Birth</td>
<td>Date</td>
<td></td>
<td>dob</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Select</td>
<td></td>
<td>sex</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Video: How to use this page]

RETURN TO PREVIOUS PAGE
**REDCap Features**

- **Good Points**
  - Easy to set up, not resource intensive
  - Requires a real data dictionary
  - Central server engine (security & data integrity)
  - Easy access through web front-end

- **Not so Great Points**
  - Display interface not very customizable
    - Layout, limited skip patterns, etc.
  - Each application is a separate instance
  - Adverse events monitoring difficult
  - Not truly relational
  - **No data curation**, electronic data collection only
REDCap (Research Electronic Data Capture)

- Online or offline use
- Regulatory compliance
  - HIPAA, 21 CFR Part 11, FISMA
- Features:
  - Customizable
  - Automated export procedures, built-in project calendar, scheduling module
  - Audit trails
  - Ad hoc reporting tools
  - Branching logic, file uploading, and calculated fields
“Database Management”
Software
Clinical Trials Management Systems (CTMS)

IMPACT® CTMS

Uses:
- Planning, preparation, monitoring and reporting of clinical trials
- Administrative/financial/portfolio management capabilities
- Electronic case report forms (eCRFs)
- ± Interoperate with other systems
Other Considerations for Data Operations

- Standard Operations Procedures (SOPs)
- Disaster recovery
- Version control (Surround SCM)
- Audit
- Separation of duties
  - DBAs, analysts, statisticians
- Electronic Sign-offs (Editor → Monitor → PI)
- Honest broker role (PHI-related)
How important are research IT/informatics solutions for novel clinical trial designs?
The I-SPY 2 Trial

The Clinical Trial, Re-Imagined

The ground-breaking I-SPY 2 trial of neoadjuvant treatment for locally advanced breast cancer established a new benchmark for efficiency of phase II clinical trials. Widely regarded as a pioneer of the ‘platform’ trial, I-SPY 2’s success continues to be a major influence on the development of next-generation trial designs in oncology and beyond.
I-SPY 2 TRIAL
(Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging And molecular Analysis 2)

I-SPY 2 is a clinical trial for women with newly diagnosed locally advanced breast cancer (neoadjuvant)
New trial design
Uses genetic profiles to highlight ‘biomarker’ differences among patients and to match drugs to patients with biomarkers that predict a benefit.

PHASE II
Randomized or non-randomized trials: about 60 patients are put in two groups: one drug and the other serves as a control group. About 40 patients receive the experimental drug, and 20 in the control group receive the placebo.

PHASE III
If a drug graduates to phase III, it typically takes 3,000 patients and about three years to determine if it is safe and effective enough for approval.

HISTORIC SUCCESS RATE
30 to 40%

PHASE II
Patients are placed in groups based on genetic profiles and are randomly assigned to either standard therapy or one of five different drugs plus standard care.

PHASE III
Researchers expect that drugs graduating from phase II to phase III can be tested with 300 patients selected according to genetic profiles found to respond to the drug in phase II. It is hoped that this will shorten the time to approval.

PROBABILITY OF SUCCESS
85%

Source: Donald Berry, M.D., Anderson Cancer Center
A New Rx for Medicine

Fed up with slow drug trials, cancer patients and doctors are testing a fast track to personalized treatments.

By RON WINSLOW

PERSONALIZED MEDICINE | How redesigning a clinical trial can speed drug development

Traditional clinical trial
Taking essentially all patients with a disease being studied and is typically intended to eliminate differences in patient characteristics that could bias measures of drug effectiveness.

New trial design
Uses genetic profiles to highlight ‘biomarker’ differences among patients and to match drugs to patients with biomarkers that predict a benefit.

PHASE II
Randomized or non-randomized trial: In a randomized trial, about 60 patients are put in two groups. One receives the experimental drug and the other serves as a control group. In a non-randomized trial, about 40 patients receive the experimental drug.

Less successful drugs are eliminated.

PHASE III
If a drug graduates to phase III, it typically takes 3,000 patients and about three years to determine if it is safe and effective enough for approval.

HISTORIC SUCCESS RATE
30 TO 40%

PHASE II
Patients are placed in groups based on genetic profiles and are randomly assigned to either standard therapy or one of five different drugs plus standard care.

Early results increase chances that patients entering the trial later will be assigned to a drug showing benefit against tumors with their genetic profile.

It will take up to 120 patients for each drug to determine which ones graduate to phase III studies.

PHASE III
Researchers expect that drugs graduating from 1-Spy 2 to phase III can be tested with 300 patients selected according to genetic profiles found to respond to the drug in phase I. It is hoped that this will shorten the time to approval.

PROBABILITY OF SUCCESS
85%

Source: Donald Berry, M.D. Anderson Cancer Center
I-SPY2 TRIAL

Population of patients adaptively randomize

Experimental arm 1
Experimental arm 2
Experimental arm 3
Experimental arm 4
Experimental arm 5
Standard therapy

Outcome: Complete response at surgery
I-SPY2 TRIAL

Population of patients

Randomize

Arm 2 graduates to small focused Phase 3 trial

Experimental arm 1
Experimental arm 2
Experimental arm 3
Experimental arm 4
Experimental arm 5
Standard therapy

Outcome: Complete response at surgery
I-SPY2 TRIAL

Outcome: Complete response at surgery

Population of patients

Randomize

Arms:
- Experimental arm 1
- Experimental arm 3
- Experimental arm 4
- Standard therapy

Arm 3 drops for futility
I-SPY2 TRIAL

Outcome: Complete response at surgery

Arm 5 graduates to small focused Phase 3 trial

Population of patients randomly assigned

Experimental arm 1
Experimental arm 4
Experimental arm 5
Standard therapy
I-SPY2 TRIAL

Outcome: Complete response at surgery

Population of patients

Randomly

Adaptively

Experimental arm 1
Experimental arm 4
Standard therapy

Arm 6 is added to the mix

Experimental arm 6
Infrastructure Considerations for Biomarker-Based Trials

- Adaptive randomization is highly dependent on near instantaneous synchronized data

- Research IT
  - Significant IT infrastructure is required to support biology-based risk-stratified or adaptive designs
  - Expensive, but there may be some economies of scale by establishing a single center to coordinate

- Repurposing
  - Design facilitates repurposing data and supporting future CER
Summary

• Do not underestimate the 90/10 rule
• You never want to visit a biostatistician for the first time with an already collected set of data
  ▪ Same thing here, plan out your data requirements and plan BEFORE you start your study
  ▪ Multidisciplinary team:
    • Biostatistician/epidemiologist
    • Research IT / informatician
    • Data management personnel
    • Regulatory personnel
• Comprehensive and thoughtful database design is key
Summary

• Comprehensive and thoughtful database design is key:
  ▪ Database content and documentation
  ▪ Software
  ▪ Hardware

• Consider capability as well as sustainability over the long-haul in how you develop your data management plan:
  ▪ Future proof as much as possible
  ▪ Stick to industry standards as much as possible
  ▪ Consider future regulatory issues
Summary (continued)

• I think that informatics/research IT should be core competencies in clinical and translational research.

• Computational technologies for managing data are changing faster than technologies for analysis.

• Good data management → High quality data

• High quality data → Analytic quality