Everything You Never Thought You Wanted to Know About Thought You Wanted to Know About Biostatistics & Study Design Services

Janet Foote, PhD, Assistant Professor, Public Health

Chengcheng Hu, PhD, Associate Professor, Public Health
Biostatistician
- Chengcheng Hu

Epidemiologist
- Janet Foote

M.S.: Paul Kang
GRA: David Margraf / Shafquat “Shaf” Saif

Additional Faculty from the Department of Biostatistics & Epidemiology as needed.
On the Research page, select **Resources**

Under the guidance of our institutional leadership and the affiliations and **partnerships** with our community affiliates, we are on an exciting course to strengthen health care for the Arizona community. Our research prowess is based in collaboration among institutions, as we are strongly committed to cultivating and expanding such interactions.

As the interdisciplinary research on our campus and those of our partners becomes seamless, federal and state government, biomedical industry, foundations, institutions of higher learning, and forward-thinking and committed individuals will become evermore engaged in empowering excellence and outcomes. The return on research investment can be measured in many ways, from local economic benefits to world-changing advancements.

Research at the College is carried out in these key **Research Departments**:

- Department of Basic Medical Sciences.
- Department of Child Health.
- Center for Applied NanoBioscience & Medicine.
- Center for Toxicology and Pharmacology Education and Research.
- Arizona Emergency Medicine Research Center.
Under Resources, select Biostatistics & Study Design Services.
Biostatistics & Study Design Services

Why do you need biostatistical support?

Biostatistics applies statistical reasoning and methods to biomedical and public health research. Our support is vital at every stage of a research project, from study design to publication. We are valuable members of interdisciplinary research teams — ensuring appropriate data collection, management and analysis. We also develop statistical methods or modify existing methods to address your study’s problems, if or when the standard approaches do not work well.

Additional information:

- Statistical Services:
  - Available Services.
  - Initial Contact Procedures.
  - User Responsibilities.
- Education — Independent Learning Modules (ILMs).
- Staff, Location and Availability.
Separate **Intake Form** link for Faculty & Students

**INITIAL CONTACT PROCEDURES FOR EACH PROJECT**

An intake form is required before consultation starts for each project. For faculty members, the intake form will first be reviewed by the UA College of Medicine – Phoenix Research Administration, and, if approved, will be forwarded to the **Biostatistics Services Team**.

Any student working on the Scholarly Project needs to talk to his/her mentor and Dr. Matthew McEchon first.

**Intake Forms:**

- [Request for Statistical Assistance for faculty members](#)
- [Request for Statistical Assistance for medical students](#) (related to Scholarly Project)

Contact information: [Biostatistical team](#)
When should you contact Biostat support?

- As you are starting to design your study protocol
- Prior to submitting proposals for projects, routing
- Before you pilot your survey / start data collection
- As you are preparing to analyze data

Many roads.....
We are here to help!
What information do you need to have?

- What is your overall research question?
- What are your specific aims to answer your question?
- Who is your target population?
- What approach do you think you’ll be using (i.e. measurements, surveys, observations/chart reviews, etc)?
- How much difference / change / variation is important?
- When do you need this?
Study Design 101
In science, we take an organized, methodical approach to examine a question. The first thing we need to do for research, is ask a question.
Once we have a question, we need to complete background research. By reading what has been published about a topic, we often find out new information and sometimes change our question a bit because we are better informed about the topic.
The next important step in the scientific method is constructing a hypothesis.
A hypothesis is an ‘educated idea’ about how things will work.

There is a set way one to state the hypothesis.

If ________(I do this), then ________(this)_______ will happen.

Ex: If I test the blood sugar of 100 adults, more than 20% will be at risk* for diabetes.

If I add methyltrexate to HL1 cancer cells, the cells will stop multiplying, so the cancer cannot progress.

The hypothesis should be measurable, and do able.

* Risk is defined as......
Testing a hypothesis is not a “one and done” procedure. In science, results must be shown to be repeatable and consistent. Statistics helps us understand the odds that the results we see are ‘real’ based on our study design.
Statistics:

- Help understand the odds that results are real
- Dependent on type and characteristics of data
- Cannot **fix** design / data / recruitment problems
Data: values, pieces of information
Types

- Categorical
- Quantity
- Nominal
- Ordinal
- Binary
- Discrete and continuous data.
- Interval and ratio variables

- Qualitative
- Quantitative

Characteristics of data
Categorical Data

- The objects being studied are grouped into categories.
- Categories are usually based on a qualitative trait.
- These data are merely labels or categories.
- May or may not have any underlying order.
Nominal Data

Categorical data in which objects fall into \textit{unordered} categories.

Examples:

• Type of Bicycle
  – Mountain bike, road bike, chopper, folding, BMX.

• Ethnicity
  – Asian, Pacific Islander, African American, Caucasian, Latino, Native American (note problems with these categories).

• Smoking status
  – smoker, non-smoker, former smoker
Ordinal Data

• Categorical data in which order is important.
• Highest Education level – elementary, high school, college graduate
• Degree of illness - none, mild, moderate, acute, chronic.
• Opinion of students about stats classes - Very unhappy, unhappy, neutral, happy, ecstatic!
**Binary Data**

- Special type of categorical data in which there are *only two categories*.
- Binary data can either be nominal or ordinal.
- Current smoking status: smoker, non-smoker
- Attendance: present, absent
- Class mark: pass, fail.
- Status of student: undergraduate, postgraduate.
Categorical data classified as Nominal, Ordinal, and/or Binary.

- Nominal data
  - Binary
  - Not binary
- Ordinal data
  - Binary
  - Not binary
Quantity Data

• Whatever is under study is being ‘measured’ based on some quantitative trait.

• Data are set of numbers.

Examples

• Pulse rate
• Height
• Age
• Exam marks
• Size of bicycle frame
• Time to complete a statistics test
• Number of cigarettes smoked
Quantity data can be classified as **Discrete** or **Continuous**
Discrete Data

Only certain values are possible (there are gaps between the possible values). Implies counting.

Continuous Data

Theoretically, with a fine enough measuring device, no gaps.
Discrete Data

- Number of children in a family
- Number of students passing a stats exam
- Number of crimes reported to the police
- Number of bicycles sold in a day.

*Generally, discrete data are counts.*

*We would not expect to find 2.2 children in a family or 88.5 students passing an exam or 127.2 crimes being reported to the police or half a bicycle being sold in one day.*
Continuous data

• Size of bicycle frame
• Height
• Time to run 500 metres
• Age

‘Generally, continuous data come from measurements.
(any value within an interval is possible with a fine enough measuring device’- (Rowntree 2000)).
Discrete data -- Gaps between possible values - count

Continuous data -- Theoretically, no gaps between possible values - measure
The type of data collected in a study determines the type of statistical analysis used.
A database is a method of organizing and analyzing information.
Why use a database?

• **Organize & analyze** information in different ways
  - Sorting
  - Grouping
  - Querying
  - Reporting
  - Exporting for statistical analysis

• **Computerized database**
  - Speed
  - Quality control
  - Precision
  - Automate repetitive tasks
Databases versus Excel

• Excel has some **limited** capabilities to sort data but its **primary function is to create financial spreadsheets**
  – Can create “what if” scenarios to determine financial consequences
  – **Can be used for small /limited research data sets & simple lists**
  – Not multi-user such that only one person can work on the file at a time

• Databases: **designed to collect, sort, & manipulate data**
  – Databases can process large amounts of data; usually limited by hardware constraints
  – Structure is in the same format for each member record of a table
  – Data quality control features ensure that valid data is entered
  – A relational database allows for linking of an unlimited number of tables
  – Databases are multi-user because the data can reside on a server and multiple people can have access at the same time
  – Many databases offer web interfaces thereby eliminating the need for each user to have a copy of the program on their computer
Many databases offer audit functions required by certain regulatory agencies

- Tracks date record created and modified
- Tracks original and changed values
- Requires user to give reason for the change

Databases are more suitable for importing data from multiple sources

- More robust in connecting to different data sources
- Imports of different data types into different tables can be linked via common identifiers such as subject ID
- Merging multiple data sources into Excel so that the rows line up properly in a flat file format can be a challenge
U of A resources

- Qualtrics™ – Survey Monkey on steroids
- Redcap™ – Clinical database
How is a database organized?

• One or more tables
• Tables store records
  ▪ Patient identifiers
  ▪ Demographics and history
  ▪ Test results
  ▪ Etc.....
• A record is a collection of fields
  – Patient identifiers
    • Name, DOB, address, ..... are stored in separate fields
# Records and Fields

## Fields

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Gender</th>
<th>Group</th>
<th>Race</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>3001</td>
<td>50.48</td>
<td>Male</td>
<td>Combined</td>
<td>CC</td>
<td>0</td>
</tr>
<tr>
<td>3002</td>
<td>65.55</td>
<td>Male</td>
<td>Diet</td>
<td>AA</td>
<td>0</td>
</tr>
<tr>
<td>3003</td>
<td>63.59</td>
<td>Female</td>
<td>Diet</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3005</td>
<td>50.07</td>
<td>Female</td>
<td>Combined</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3010</td>
<td>60.28</td>
<td>Male</td>
<td>Diet</td>
<td>CC</td>
<td>0</td>
</tr>
<tr>
<td>3011</td>
<td>56.43</td>
<td>Female</td>
<td>Diet</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3012</td>
<td>45.80</td>
<td>Female</td>
<td>Combined</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3013</td>
<td>56.05</td>
<td>Female</td>
<td>Combined</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3014</td>
<td>65.48</td>
<td>Female</td>
<td>Diet</td>
<td>AA</td>
<td>1</td>
</tr>
<tr>
<td>3015</td>
<td>58.21</td>
<td>Female</td>
<td>Diet</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3016</td>
<td>57.30</td>
<td>Female</td>
<td>Combined</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3017</td>
<td>53.93</td>
<td>Female</td>
<td>Combined</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3018</td>
<td>50.12</td>
<td>Female</td>
<td>Diet</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3019</td>
<td>57.36</td>
<td>Female</td>
<td>Combined</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3020</td>
<td>51.05</td>
<td>Male</td>
<td>Diet</td>
<td>CC</td>
<td>0</td>
</tr>
<tr>
<td>3021</td>
<td>66.11</td>
<td>Female</td>
<td>Diet</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3024</td>
<td>54.90</td>
<td>Female</td>
<td>Diet</td>
<td>AA</td>
<td>1</td>
</tr>
<tr>
<td>3025</td>
<td>65.62</td>
<td>Female</td>
<td>Combined</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3027</td>
<td>45.91</td>
<td>Female</td>
<td>Diet</td>
<td>AA</td>
<td>1</td>
</tr>
<tr>
<td>3029</td>
<td>58.42</td>
<td>Female</td>
<td>Combined</td>
<td>CC</td>
<td>1</td>
</tr>
<tr>
<td>3032</td>
<td>53.50</td>
<td>Male</td>
<td>Diet</td>
<td>CC</td>
<td>0</td>
</tr>
<tr>
<td>3033</td>
<td>48.83</td>
<td>Female</td>
<td>Combined</td>
<td>AA</td>
<td>1</td>
</tr>
</tbody>
</table>
How is data displayed?

- Fields are displayed on layouts
  - Forms
  - Web
  - Reports
- Data can be from a single table or many tables if using a relational database
## Relational Database Example

<table>
<thead>
<tr>
<th>Subject Info</th>
<th>Anthropometrics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Id</strong></td>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>10</td>
<td>Smith</td>
</tr>
<tr>
<td>11</td>
<td>Jones</td>
</tr>
<tr>
<td>12</td>
<td>Doe</td>
</tr>
</tbody>
</table>

**Anthropometrics**

<table>
<thead>
<tr>
<th><strong>Id</strong></th>
<th><strong>KCAL</strong></th>
<th><strong>KCAL/kg</strong></th>
<th><strong>V02</strong></th>
<th><strong>V02/kg</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2400</td>
<td>23.1</td>
<td>2.8</td>
<td>26.7</td>
</tr>
<tr>
<td>11</td>
<td>2652</td>
<td>27.5</td>
<td>3.2</td>
<td>33.1</td>
</tr>
<tr>
<td>12</td>
<td>2350</td>
<td>25.9</td>
<td>2.1</td>
<td>23.2</td>
</tr>
</tbody>
</table>
Differences between a clinical & research database

- **Clinical database**
  - Form or report oriented so data is displayed for clinical decision making
  - Emphasis on displaying or reporting of individual data rather than accumulating multiple records

- **Research database**
  - Table oriented so that data is accumulated for eventual export to a statistical package for data analysis and reporting
  - Less emphasis on individual records
Clinical data:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject ID</td>
<td>2002</td>
</tr>
<tr>
<td>Sequence</td>
<td>Baseline</td>
</tr>
<tr>
<td>Visit Date</td>
<td>06/16/2004</td>
</tr>
<tr>
<td>Weight (lb)</td>
<td>185.0</td>
</tr>
<tr>
<td>HR Rest (bpm)</td>
<td>67.</td>
</tr>
<tr>
<td>Time (max sec)</td>
<td>534</td>
</tr>
<tr>
<td>Max Speed (mph)</td>
<td>3.0</td>
</tr>
<tr>
<td>Max Elev (%)</td>
<td>5.0</td>
</tr>
<tr>
<td>Max VO2/kg</td>
<td>21.1</td>
</tr>
<tr>
<td>Max VO2 (L/min)</td>
<td>1778</td>
</tr>
<tr>
<td>Max VCO2 (L/min)</td>
<td>1992</td>
</tr>
<tr>
<td>Max ReR</td>
<td>1.12</td>
</tr>
<tr>
<td>Maximal HR (bpm)</td>
<td>134</td>
</tr>
<tr>
<td>Max VO2/HR</td>
<td>13</td>
</tr>
<tr>
<td>Max RR</td>
<td>48</td>
</tr>
<tr>
<td>Max VE (L/min)</td>
<td>83.7</td>
</tr>
<tr>
<td>Max RPE</td>
<td>9.</td>
</tr>
<tr>
<td>Max SBP (mmHg)</td>
<td>220</td>
</tr>
<tr>
<td>Max DBP (mmHg)</td>
<td>88.</td>
</tr>
</tbody>
</table>

**AT**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>% VO2 Max</td>
<td>61.00</td>
</tr>
<tr>
<td>HR (bpm)</td>
<td>112</td>
</tr>
<tr>
<td>VO2/kg (mL/kg/min)</td>
<td>13.00</td>
</tr>
</tbody>
</table>
## Research data:

<table>
<thead>
<tr>
<th>ID1</th>
<th>Seq</th>
<th>VisitDate</th>
<th>WeightLb</th>
<th>HRRest</th>
<th>Time</th>
<th>Speed</th>
<th>Elev</th>
<th>VO2kg</th>
<th>VO2</th>
<th>VC02</th>
<th>REF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>Baseline</td>
<td>6 mo</td>
<td>06/16/2004</td>
<td>185.0</td>
<td>67.</td>
<td>534</td>
<td>3.0</td>
<td>5.0</td>
<td>21.1</td>
<td>1773</td>
<td>1992</td>
</tr>
<tr>
<td>2001</td>
<td>Baseline</td>
<td>6 mo</td>
<td>06/09/2004</td>
<td>195.0</td>
<td>48.</td>
<td>746</td>
<td>03.0</td>
<td>10.0</td>
<td>18.5</td>
<td>1637</td>
<td>1763</td>
</tr>
<tr>
<td>2004</td>
<td>Baseline</td>
<td>6 mo</td>
<td>06/10/2004</td>
<td>239.0</td>
<td>91.</td>
<td>854</td>
<td>03.0</td>
<td>10.0</td>
<td>19.2</td>
<td>2081</td>
<td>2051</td>
</tr>
<tr>
<td>2005</td>
<td>Baseline</td>
<td>6 mo</td>
<td>06/18/2004</td>
<td>280.0</td>
<td>97.</td>
<td>514</td>
<td>3.0</td>
<td>5.0</td>
<td>20.6</td>
<td>2629</td>
<td>2732</td>
</tr>
<tr>
<td>2006</td>
<td>Baseline</td>
<td>6 mo</td>
<td>07/15/2004</td>
<td>204.0</td>
<td>98.</td>
<td>660</td>
<td>3.0</td>
<td>7.5</td>
<td>18.3</td>
<td>1697</td>
<td>1819</td>
</tr>
<tr>
<td>2011</td>
<td>Baseline</td>
<td>6 mo</td>
<td>07/15/2004</td>
<td>169.0</td>
<td>75.</td>
<td>557</td>
<td>3.0</td>
<td>7.0</td>
<td>11.7</td>
<td>0898</td>
<td>0908</td>
</tr>
<tr>
<td>2020</td>
<td>Baseline</td>
<td>6 mo</td>
<td>09/01/2004</td>
<td>243.4</td>
<td>63.</td>
<td>444</td>
<td>3.0</td>
<td>5.0</td>
<td>13.7</td>
<td>1512</td>
<td>1691</td>
</tr>
<tr>
<td>2025</td>
<td>Baseline</td>
<td>6 mo</td>
<td>08/24/2004</td>
<td>200.0</td>
<td>72.</td>
<td>820</td>
<td>3.0</td>
<td>10.0</td>
<td>22.9</td>
<td>2085</td>
<td>2357</td>
</tr>
<tr>
<td>2019</td>
<td>Baseline</td>
<td>6 mo</td>
<td>09/07/2004</td>
<td>133.9</td>
<td>63.</td>
<td>857</td>
<td>3.0</td>
<td>10.0</td>
<td>21.5</td>
<td>1307</td>
<td>1466</td>
</tr>
<tr>
<td>2028</td>
<td>Baseline</td>
<td>6 mo</td>
<td>10/07/2004</td>
<td>222.0</td>
<td>85.</td>
<td>1134</td>
<td>3.0</td>
<td>15.0</td>
<td>28.7</td>
<td>2895</td>
<td>3035</td>
</tr>
<tr>
<td>2017</td>
<td>Baseline</td>
<td>6 mo</td>
<td>10/15/2004</td>
<td>211.0</td>
<td>68.</td>
<td>540</td>
<td>3.0</td>
<td>5.0</td>
<td>17.9</td>
<td>1715</td>
<td>1809</td>
</tr>
<tr>
<td>2041</td>
<td>Baseline</td>
<td>6 mo</td>
<td>10/25/2004</td>
<td>216.0</td>
<td>83.</td>
<td>629</td>
<td>3.0</td>
<td>7.5</td>
<td>18.4</td>
<td>1803</td>
<td>1783</td>
</tr>
<tr>
<td>2035</td>
<td>Baseline</td>
<td>6 mo</td>
<td>11/04/2004</td>
<td>209.0</td>
<td>92.</td>
<td>783</td>
<td>3.0</td>
<td>10.0</td>
<td>23.4</td>
<td>2233</td>
<td>2260</td>
</tr>
<tr>
<td>2040</td>
<td>Baseline</td>
<td>6 mo</td>
<td>11/04/2004</td>
<td>284.0</td>
<td>88.</td>
<td>360</td>
<td>3.0</td>
<td>2.5</td>
<td>20.0</td>
<td>2585</td>
<td>2613</td>
</tr>
<tr>
<td>2029</td>
<td>Baseline</td>
<td>6 mo</td>
<td>09/16/2004</td>
<td>168.0</td>
<td>85.</td>
<td>777</td>
<td>3.0</td>
<td>10.0</td>
<td>18.8</td>
<td>1433</td>
<td>1467</td>
</tr>
<tr>
<td>2016</td>
<td>Baseline</td>
<td>6 mo</td>
<td>08/02/2004</td>
<td>183.0</td>
<td>69.</td>
<td>869</td>
<td>3.0</td>
<td>10.0</td>
<td>23.1</td>
<td>1918</td>
<td>2084</td>
</tr>
<tr>
<td>2034</td>
<td>Baseline</td>
<td>6 mo</td>
<td>11/18/2004</td>
<td>196.0</td>
<td>85.</td>
<td>1168</td>
<td>3.0</td>
<td>15.0</td>
<td>31.1</td>
<td>2773</td>
<td>3260</td>
</tr>
<tr>
<td>2045</td>
<td>Baseline</td>
<td>6 mo</td>
<td>11/23/2004</td>
<td>245.9</td>
<td>69.</td>
<td>813</td>
<td>3.0</td>
<td>10.0</td>
<td>23.1</td>
<td>2577</td>
<td>2700</td>
</tr>
<tr>
<td>2039</td>
<td>Baseline</td>
<td>6 mo</td>
<td>12/03/2004</td>
<td>250.0</td>
<td>73.</td>
<td>474</td>
<td>3.0</td>
<td>5.0</td>
<td>14.8</td>
<td>1683</td>
<td>1906</td>
</tr>
<tr>
<td>2051</td>
<td>Baseline</td>
<td>6 mo</td>
<td>12/06/2004</td>
<td>215.0</td>
<td>68.</td>
<td>670</td>
<td>3.0</td>
<td>7.5</td>
<td>17.7</td>
<td>1728</td>
<td>1740</td>
</tr>
<tr>
<td>2005</td>
<td>Baseline</td>
<td>6 mo</td>
<td>12/28/2004</td>
<td>280.0</td>
<td>97.</td>
<td>638</td>
<td>3.0</td>
<td>7.5</td>
<td>22.0</td>
<td>2806</td>
<td>2813</td>
</tr>
<tr>
<td>2056</td>
<td>Baseline</td>
<td>6 mo</td>
<td>01/18/2005</td>
<td>140.5</td>
<td>59.</td>
<td>744</td>
<td>2.0</td>
<td>10.0</td>
<td>14.6</td>
<td>0931</td>
<td>0924</td>
</tr>
<tr>
<td>2025</td>
<td>Baseline</td>
<td>6 mo</td>
<td>01/31/2005</td>
<td>200.0</td>
<td>85.</td>
<td>694</td>
<td>3.0</td>
<td>7.5</td>
<td>21.0</td>
<td>1912</td>
<td>2115</td>
</tr>
<tr>
<td>2053</td>
<td>Baseline</td>
<td>6 mo</td>
<td>02/08/2005</td>
<td>204.5</td>
<td>91.</td>
<td>473</td>
<td>3.0</td>
<td>5.0</td>
<td>16.5</td>
<td>1536</td>
<td>1566</td>
</tr>
<tr>
<td>2016</td>
<td>Baseline</td>
<td>6 mo</td>
<td>02/04/2005</td>
<td>180.0</td>
<td>73.</td>
<td>814</td>
<td>3.0</td>
<td>10.0</td>
<td>24.0</td>
<td>1964</td>
<td>2140</td>
</tr>
<tr>
<td>2004</td>
<td>Baseline</td>
<td>6 mo</td>
<td>02/14/2005</td>
<td>252.0</td>
<td>72.</td>
<td>885</td>
<td>3.0</td>
<td>10.0</td>
<td>21.4</td>
<td>2445</td>
<td>2515</td>
</tr>
<tr>
<td>2001</td>
<td>Baseline</td>
<td>6 mo</td>
<td>02/14/2005</td>
<td>195.0</td>
<td>39.</td>
<td>812</td>
<td>3.0</td>
<td>10.0</td>
<td>17.8</td>
<td>1580</td>
<td>1733</td>
</tr>
<tr>
<td>2006</td>
<td>Baseline</td>
<td>6 mo</td>
<td>01/31/2005</td>
<td>202.0</td>
<td>111.</td>
<td>649</td>
<td>3.0</td>
<td>7.5</td>
<td>18.3</td>
<td>1679</td>
<td>1700</td>
</tr>
<tr>
<td>2061</td>
<td>Baseline</td>
<td>6 mo</td>
<td>02/16/2005</td>
<td>226.0</td>
<td>88.</td>
<td>913</td>
<td>3.0</td>
<td>10.0</td>
<td>25.3</td>
<td>2603</td>
<td>2603</td>
</tr>
<tr>
<td>2052</td>
<td>Baseline</td>
<td>6 mo</td>
<td>03/04/2005</td>
<td>203.0</td>
<td>72.</td>
<td>1066</td>
<td>3.0</td>
<td>12.5</td>
<td>27.2</td>
<td>2506</td>
<td>2687</td>
</tr>
<tr>
<td>2066</td>
<td>Baseline</td>
<td>6 mo</td>
<td>03/16/2005</td>
<td>257.0</td>
<td>87.</td>
<td>982</td>
<td>3.0</td>
<td>12.5</td>
<td>24.1</td>
<td>2816</td>
<td>3063</td>
</tr>
<tr>
<td>2069</td>
<td>Baseline</td>
<td>6 mo</td>
<td>03/16/2005</td>
<td>200.2</td>
<td>55.</td>
<td>616</td>
<td>3.0</td>
<td>7.5</td>
<td>18.1</td>
<td>1653</td>
<td>1733</td>
</tr>
<tr>
<td>2063</td>
<td>Baseline</td>
<td>6 mo</td>
<td>03/23/2005</td>
<td>360.0</td>
<td>59.</td>
<td>583</td>
<td>3.0</td>
<td>7.5</td>
<td>16.9</td>
<td>2002</td>
<td>2076</td>
</tr>
</tbody>
</table>
Advantages of a database

- Collection of data in a centralized location
- Controls redundant data
- Data stored so as to appear to users in one location
  - Data can be stored in multiple tables and come from multiple sources
  - A relational database brings it all together
Sharing and Exchanging Data

- Multiple users can access the same database via a network
  - Can be local or over the internet
  - Best done when the data are stored on a database server
    - Access via a client application
    - Access via a web interface
  - Server allows remote access over the internet from anywhere
    - Should be behind a firewall for security with access via VPN and password protection
Database Design Considerations

• **What to collect**
  – What questions are to be answered?
  – Think of the data tables in your future publications
    • **Focus on the key data elements rather than collect as much as possible**
  
• **What statistical package will be used?**
  – Format of the data file to which the data will be exported
    • Allowable characters
    • Format for certain analyses
      – For example, gender can be recorded in the database as M or F but statistical package may require 0 and 1
    • Length of data field labels
    • Long or wide format
Long versus Wide Format

Long: each year is represented as its own observation in a record

<table>
<thead>
<tr>
<th>famid</th>
<th>year</th>
<th>faminc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>96</td>
<td>40000</td>
</tr>
<tr>
<td>2.</td>
<td>97</td>
<td>40500</td>
</tr>
<tr>
<td>3.</td>
<td>98</td>
<td>41000</td>
</tr>
<tr>
<td>4.</td>
<td>96</td>
<td>45000</td>
</tr>
<tr>
<td>5.</td>
<td>97</td>
<td>45400</td>
</tr>
<tr>
<td>6.</td>
<td>98</td>
<td>45800</td>
</tr>
<tr>
<td>7.</td>
<td>96</td>
<td>75000</td>
</tr>
<tr>
<td>8.</td>
<td>97</td>
<td>76000</td>
</tr>
<tr>
<td>9.</td>
<td>98</td>
<td>77000</td>
</tr>
</tbody>
</table>

Wide: each family is a record and each year is a field with that record

<table>
<thead>
<tr>
<th>famid</th>
<th>faminc96</th>
<th>faminc97</th>
<th>faminc98</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>3</td>
<td>75000</td>
<td>76000</td>
</tr>
<tr>
<td>2.</td>
<td>1</td>
<td>40000</td>
<td>40500</td>
</tr>
<tr>
<td>3.</td>
<td>2</td>
<td>45000</td>
<td>45400</td>
</tr>
</tbody>
</table>
Selected Elements of Data Management Planning
Quality Control of Data Before Study

- Collect only needed variables
- Select appropriate computer hardware & software
- Plan analyses with dummy tabulations
- Develop study forms
  - Precode responses
  - Format boxes for data entry
  - Label each page with date, time, ID
  - Consider scan technology
What needs to be in the research database?

- Research variables directly related to the hypotheses being tested- **YES**
- Clinical measures used for screening- **MAYBE**
  - Blood work, ECG, medical history
- Administrative data- **NO**
  - Contact information
  - Scheduling
Where Are the Original Data?

In the source documents
What is a Source Document?

- It is the First Recording

- What does it tell?
  1. It is the data that document the trial
  2. Study was carried out according to protocol
Source Documents

- Original Lab reports
- Pathology reports
- Surgical reports
- Physician Progress Notes
- Nurses Notes
- Medical Record
- Letters from referring physicians
- Original radiological films
- Tumor measurements
- Patient Diary/patient interview
Common Data Elements

- **Standardized**, unique terms and phrases that delineate discrete pieces of information used to collect data in a clinical trial
- **Uniform** representation of demographics and data points to consistently track trends
- Elements **define** study parameters and **endpoints**
Designing the questions

- Granular primary data
  - No observer conclusions, synthesis, coding
- Categorical/ordinal data when possible—statistical power. Re-slice at analysis
- Use validated scales/instruments
  - Don’t build your own unless unavoidable
- Collect key variables with >1 question
- Avoid measurements that cluster at one end of scale
  - Distribution problems, Likert scales
**Influenza-Associated Pediatric Deaths Case Report Form**

**State Use Only – Do not send information in this section to CDC**

<table>
<thead>
<tr>
<th>Last Name:</th>
<th>First Name:</th>
<th>County:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td>City:</td>
<td>State, Zip:</td>
</tr>
</tbody>
</table>

**Patient Demographics**

1. State: 
2. County: 
3. State ID: 
4. CDC ID: 

5. Age: [ ] Days [ ] Months [ ] Years
6. Date of birth: [ ] [ ] [ ] [ ] [ ]
7. Sex: [ ] Male [ ] Female
8. Ethnicity: [ ] Hispanic or Latino [ ] Not Hispanic or Latino [ ] Unknown
9. Race: [ ] White [ ] Black [ ] Asian [ ] Native Hawaiian or Other Pacific Islander [ ] American Indian or Alaska Native

**Death Information**

10. Date of llness onset: [ ] [ ] [ ] [ ]
11. Date of death: [ ] [ ] [ ] [ ]
12. Was an autopsy performed? [ ] Yes [ ] No
13. Location of death: [ ] Home [ ] Emergency Dept (ER) [ ] Inpatient ward [ ] ICU [ ] Other (specify): 

**Influenza Testing (check all that were used)**

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Result</th>
<th>Specimen Collection Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ] Commercial rapid diagnostic test</td>
<td>[ ] Influenza A [ ] Influenza B [ ] Negative</td>
<td><em><strong>/</strong></em>/___</td>
</tr>
<tr>
<td>[ ] Influenza A/B (Not Distinguishable)</td>
<td>[ ] Influenza A [ ] Influenza B [ ] Negative</td>
<td><em><strong>/</strong></em>/___</td>
</tr>
<tr>
<td>[ ] Viral culture</td>
<td>[ ] Influenza A (Subtyping Not Done) [ ] Influenza B [ ] Negative</td>
<td><em><strong>/</strong></em>/___</td>
</tr>
<tr>
<td>[ ] Influenza A (Unuable Te Subtype) [ ] Influenza A (H1) [ ] Influenza A (H3)</td>
<td>[ ] Influenza A [ ] Influenza B [ ] Negative</td>
<td><em><strong>/</strong></em>/___</td>
</tr>
<tr>
<td>[ ] Direct fluorescent antibody (DFA)</td>
<td>[ ] Influenza A [ ] Influenza A/B</td>
<td>[ ] Influenza B [ ] Negative</td>
</tr>
<tr>
<td>[ ] Indirect fluorescent antibody (IFA)</td>
<td>[ ] Influenza A [ ] Influenza A/B</td>
<td>[ ] Influenza B [ ] Negative</td>
</tr>
<tr>
<td>[ ] Enzyme immunoassay (EIA)</td>
<td>[ ] Influenza A (Subtyping Not Done)</td>
<td>[ ] Influenza B [ ] Negative</td>
</tr>
<tr>
<td>[ ] Influenza A (Unuable Te Subtype) [ ] Influenza A (H1) [ ] Influenza A (H3)</td>
<td>[ ] Influenza B [ ] Negative</td>
<td><em><strong>/</strong></em>/___</td>
</tr>
<tr>
<td>[ ] ET-PCR</td>
<td>[ ] Influenza A (Subtyping Not Done)</td>
<td>[ ] Influenza B [ ] Negative</td>
</tr>
<tr>
<td>[ ] Influenza A (Unuable Te Subtype)</td>
<td>[ ] Influenza A (H1) [ ] Influenza A (H3)</td>
<td>[ ] Influenza B [ ] Negative</td>
</tr>
<tr>
<td>[ ] Immunohistochemistry (IHC)</td>
<td>[ ] Influenza A</td>
<td>[ ] Influenza B [ ] Negative</td>
</tr>
</tbody>
</table>

**Culture confirmation of INVASIVE bacterial pathogens**

14. Was an INVASIVE bacterial infection confirmed by culturing an organism from a specimen collected from a normally sterile site (e.g., blood, cerebrospinal fluid [CSF], tissue, or pleural fluid)? [ ] Yes [ ] No

- [ ] Streptococcus pneumoniae [ ] Staphylococcus aureus, methicillin sensitive
- [ ] Neisseria meningitidis (serogroup, if known):
- [ ] Haemophilus influenzae type b [ ] Staphylococcus aureus, methicillin resistant (MRSA)
- [ ] Group A streptococcus
- [ ] Other invasive bacteria:

Public reporting burden of this collection of information is estimated to average 20 minutes per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. An agency may not conduct or sponsor, and a person is not required to respond to a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to CDC/ATSDR/Reports/Clerical Officer: 1600 Clifton Road NE, MS E-11, Atlanta, Georgia 30333; ATTN: IRA (9200.0007).
Operations Manual

- Defines entire study protocol, sequence
- Form-specific annotation, guidance
- Documents all post-hoc validity checks, edit checks, data curation criteria
- Evolving document with periodic updates
  - Preferably on-line
- Use for training, quality control, process planning
Data Dictionary - Operational

• For every form/table, lists:
  – Variable name (database field)
  – Variable description (plain English)
  – Variable type (string, integer, numeric, etc.)
  – Variable length (or precision)
  – Nullability (missing or no value indicator)
  – Range checks, allowable values
  – Coding conventions, with definitions
<table>
<thead>
<tr>
<th>Variable name</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANYSKCA</td>
<td>1=yes; 0=no</td>
<td>Any NMSC post-randomization?</td>
</tr>
<tr>
<td>ANYSKCA6</td>
<td>1=yes; 0=no</td>
<td>Any NMSC after 6 months post-randomization?</td>
</tr>
<tr>
<td>BCCOCC</td>
<td>1=yes; 0=no</td>
<td>Any BCC occurrence post-randomization?</td>
</tr>
<tr>
<td>BCCOCC6</td>
<td>1=yes; 0=no</td>
<td>Any BCC after 6 months post-randomization?</td>
</tr>
<tr>
<td>SCCOCC</td>
<td>1=yes; 0=no</td>
<td>Any SCC occurrence post-randomization?</td>
</tr>
<tr>
<td>SCCOCC6</td>
<td>1=yes; 0=no</td>
<td>Any SCC after 6 months post-randomization?</td>
</tr>
<tr>
<td>ALLSKCA</td>
<td>number</td>
<td>Total number of NMSC which occurred post-randomization</td>
</tr>
<tr>
<td>ALLSKCA6</td>
<td>number</td>
<td>Total number of NMSC after 6 months post-randomization</td>
</tr>
<tr>
<td>ALLBCC</td>
<td>number</td>
<td>Total number of BCC which occurred post-randomization</td>
</tr>
<tr>
<td>ALLBCC6</td>
<td>number</td>
<td>Total number of BCC after 6 months post-randomization</td>
</tr>
<tr>
<td>ALLSCC</td>
<td>number</td>
<td>Total number of SCC which occurred post-randomization</td>
</tr>
<tr>
<td>ALLSCC6</td>
<td>number</td>
<td>Total number of SCC after 6 months post-randomization</td>
</tr>
<tr>
<td>ANYMOS</td>
<td>number</td>
<td>Total number of months before any NMSC occurrence</td>
</tr>
<tr>
<td>ANYMOS6</td>
<td>number</td>
<td>Months (after 6 mos post-randomization) before NMSC</td>
</tr>
<tr>
<td>BCCMOS</td>
<td>number</td>
<td>Total number of months before first BCC occurrence</td>
</tr>
<tr>
<td>BCCMOS6</td>
<td>number</td>
<td>Months (after 6 mos post-randomization) before BCC</td>
</tr>
<tr>
<td>SCCMOS</td>
<td>number</td>
<td>Total number of months before first SCC occurrence</td>
</tr>
<tr>
<td>SCCMOS6</td>
<td>number</td>
<td>Months (after 6 mos post-randomization) before SCC</td>
</tr>
</tbody>
</table>
Why code:

- Forces analyzable data structure, format
- Vastly simplifies analysis
- Speeds data input/transcription
- Vastly simplifies data analysis/reporting
Example of the need for data coding

What is the subject’s sex?

- male
- Male
- M
- m
- Man
- Boy
- 0
- 1
- Gentleman
- Tarzan
- female
- Female
- F
- f
- Woman
- Girl
- 1
- 2
- Lady
- Jane
What do you mean & how will you record it?

- HEADACHE
  - Headache
  - Pain in the head
- ACHE:
  - Ache:Head
  - Head Pain
  - HP

Unless there is a standard code for the use of terms, data retrieval becomes difficult
Rules for Data Entry

- Each variable has a field in the dataset
- Categorical and nominal values require a number or string code
- Continuous values are entered directly
- Missing values must be different values from a real response
  - Common formats are “99” or bullets “·”
  - Don’t know is a response—do not leave blank
  - “0” is not the same as missing
- Coding instructions should be on form
- Avoid open-ended questions
Avoid open-ended questions

Enter the subject’s gender:___________________

Enter the subject's level of education:____________
Close Ended Question

What is the subject’s sex? *Check one*

- Male
- Female
Use pre-coded responses where possible

Subject ID  1001

Gender  ○ Male  ○ Female

Age  56

Education  ○ 6th grade or less  ○ 2 or 3 years of college
           ○ 7th, 8th, or 9th grade  ○ 4 years of college
           ○ 10th or 11th grade    ○ 5 or more years of college
           ○ 12th grade
## Data in Spreadsheet

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>Male</td>
<td>52</td>
</tr>
<tr>
<td>1002</td>
<td>Male</td>
<td>54</td>
</tr>
<tr>
<td>103</td>
<td>Mael</td>
<td>65</td>
</tr>
<tr>
<td>1004</td>
<td>Female</td>
<td>54</td>
</tr>
<tr>
<td>5</td>
<td>Female</td>
<td>52</td>
</tr>
<tr>
<td>1006</td>
<td>Female</td>
<td>52</td>
</tr>
<tr>
<td>1007</td>
<td>Femele</td>
<td>75</td>
</tr>
<tr>
<td>1008</td>
<td>Male</td>
<td>48</td>
</tr>
<tr>
<td>1009</td>
<td>M</td>
<td>37</td>
</tr>
<tr>
<td>1010</td>
<td>Female</td>
<td>73</td>
</tr>
<tr>
<td>11</td>
<td>F</td>
<td>54</td>
</tr>
</tbody>
</table>
Data Validation
Subject ID: 101
Gender: Male
Age: 56
Education:
- 6th grade or less
- 7th, 8th, or 9th grade
- 10th or 11th grade
- 12th grade
- 2 or 3 years of college
- 4 years of college
- 5 or more years of college
This field is defined to contain a value in the range from “1000” through “1500”. Allow this value which isn’t in the range?

- Revert Field
- No
- Yes

Options:
- 10th or 11th grade
- 5 or more years of college
- 12th grade
Types of Edit Checks

- Patient identification and record linkage
  - ID #’s, name spelling, ID#’s on all pages
- Legibility
- Correct form for examination
- Missing data
- Consistency
- Range and inadmissible codes
Backup

- Data must be backed up on a regular basis to protect against:
  - Theft, fire, floods, hurricanes,
  - Equipment failure

- Computer backup
  - Mirrored drives
  - Digital tapes
  - Store backup tapes off-site
Putting it All Together: Research Data Management

- An artful selection of physical & electronic management methods
  - Signed informed consent documents
  - Paper forms
  - Regulatory & project management binders
  - Data models and databases
  - Data acquisition and display technologies
  - Communications technologies for project management as well as data management
Attributes of Successful Data Management

- Attention to detail
- Explicit structure and process
- Robust designs
  - Anticipate failures, lapses and mistakes
  - Design systems that identify and correct them
- Mechanisms for verification
- Well documented
Quality

Fast is fine, but accuracy is everything.

(Wyatt Earp)
<table>
<thead>
<tr>
<th>Study Design</th>
<th>Study Design Decision Short-hand</th>
<th>Measure of Association</th>
<th>Interpretation of Measure of Association</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case control</td>
<td>Starts with cases</td>
<td>Odds ratio (OR)</td>
<td>Those with the disease were X times more/less likely to have been exposed</td>
</tr>
<tr>
<td>Cohort</td>
<td>Starts with a designated group or exposure; follows through time</td>
<td>Relative risk (RR)</td>
<td>Those with the exposure were X times more/less likely to have the disease</td>
</tr>
<tr>
<td>Clinical Trial</td>
<td>Investigator decides exposures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-sectional</td>
<td>None of the above</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Want to be thorough.....

...but as minimally-invasive as possible
SORRY SHOULDN'T TAKE MUCH LONGER...COULD YOU JUST CONFIRM THE MAIDEN NAME OF YOUR MOTHER'S UNCLE'S SECOND COUSIN?
Finding more information online:

The University of Arizona Cancer Center at Dignity Health St. Joseph’s Hospital and Medical Center Officially Unveiled

The New Center Will Bring Comprehensive Cancer Care to Downtown Phoenix
On the Research page, select **Resources**

Under the guidance of our institutional leadership and the affiliations and **partnerships** with our community affiliates, we are on an exciting course to strengthen health care for the Arizona community. Our research prowess is based in collaboration among institutions, as we are strongly committed to cultivating and expanding such interactions.

As the interdisciplinary research on our campus and those of our partners becomes seamless, federal and state government, biomedical industry, foundations, institutions of higher learning, and forward-thinking and committed individuals will become evermore engaged in empowering excellence and outcomes. The return on research investment can be measured in many ways, from local economic benefits to world-changing advancements.

Research at the College is carried out in these key Research Departments:

- [Department of Basic Medical Sciences](#)
- [Department of Child Health](#)
- [Center for Applied NanoBioscience & Medicine](#)
- [Center for Toxicology and Pharmacology Education and Research](#)
- [Arizona Emergency Medicine Research Center](#)
Under Resources, select

Biostatistics & Study Design Services
Biostatistics applies statistical reasoning and methods to biomedical and public health research. Our support is vital at every stage of a research project, from study design to publication. We are valuable members of interdisciplinary research teams — ensuring appropriate data collection, management and analysis. We also develop statistical methods or modify existing methods to address your study’s problems, if or when the standard approaches do not work well.

**Additional Information:**

- **Statistical Services:**
  - Available Services.
  - Initial Contact Procedures.
  - User Responsibilities.
- **Education** — Independent Learning Modules (ILMs).
- **Staff, Location and Availability.**