How to Work with a Statistician

Sandra Taylor, Ph.D.
Clinical and Translational Science Center
University of California, Davis
26 April 2012
Overview

- Understand what statisticians can (and cannot) do for you
- Know your responsibilities as a researcher
- Highlight some “do’s” and “don’ts”
- Resources available through the CTSC
- Example consultations
Why and when to work with a statistician?

- **Planning the study**
  - Study design, randomization, sample size
  - Proposal preparation

- **Conducting the study**
  - Interim analyses, DSMB

- **Evaluating the results**
  - Conducting statistical analyses

- **Reporting the results**
  - Interpreting the results
  - Manuscript preparation
The Research Process
“Gee, I wonder if...”
Planning Stage Involvement
Planning Stage Involvement

Statistics can’t fix a poorly designed study!
Planning Stage Activities

**RESEARCHER**
- Develop specific aims
- Identify endpoints
- Formulate testable hypotheses
- Develop study protocol
- Identify confounding factors, potential biases

**STATISTICIAN**
- Formulate testable hypotheses
- Study design, randomization, matching
- Sample size needs
- Statistical analysis plan
- Identify confounding factors, potential biases
Before Meeting with a Statistician

- Clearly define your specific aims
- Identify measureable endpoints for each aim
- Know how you will collect the data
- Consider confounders and potential biases
- Review relevant literature
Why are these important?

- Protocol
- Confounders
- Literature
- Specific Aims
- Endpoints
- Study Design
- Methods
- Sample Size
Specific Aim 1: Determine if new treatment is better than standard care.

- What constitutes *better*?
- What measurable parameter reflects *better*?
  - Survival, number of events, mean value?
- What testable hypothesis addresses the specific aim?
  - $H_0$: Mean cholesterol under new treatment does not differ from standard care
  - $H_a$: Mean cholesterol under new treatment differs from standard care
How do I test my hypothesis?

- **Type of study**
  - Prospective or retrospective

- **Definition of cases and controls**
  - Exclusion and inclusion criteria

- **Matching/stratification procedures**

- **Randomization**
  - Subject selection, block randomization

- **Sample size requirements**
How many subjects do I need?

- Too few – insufficient power to detect differences
- Too many – unnecessary costs
- Statisticians need input from researchers to determine sample size requirements.
Biologist talks to Statistician

Please don’t do this.
Sample size determination comes at the end of a series of steps.

Specific Aim → Testable Hypothesis

Study Design → Sample Size Calculations

Statistical Methods → Study Design
How is sample size determined?

Depends on:
- Specific aim – primary hypothesis of the study
- Study design
  - These two influence the statistical test.
- Effect size to be detected
- Variability of the response variable
  - Researchers need to provide this information
Example: New medication study

- **Test:** $H_0: \mu_{\text{new}} = \mu_{\text{old}}$ vs. $H_a: \mu_{\text{new}} \neq \mu_{\text{old}}$
- **Design:** Randomized into each arm
- **Statistical method:** $t$-test

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2 \sigma^2}{\Delta^2}$$

- $\sigma^2 = \text{variance}$; $\Delta = \text{effect size to detect}$

Researchers need to provide this information.
- Published results
- Pilot data
- Clinically meaningful change
Sample size calculations may not be straight-forward

- More complex designs require more complex calculations
- **Examples:**
  - Longitudinal studies
  - Cross-over studies
  - Correlation of outcomes
- **Sometimes simulations are required**
Sample size calculations may not be straight-forward

- More complex designs require more complex calculations
- **Examples:**
  - Longitudinal studies
  - Cross-over studies
  - Correlation of outcomes
- Sometimes simulations are required

**THERE IS NO MAGIC NUMBER.**
Proposal Content

Proposal sections involving statistics

- Sample size justification
- Statistical analysis plan
  - Statistical methods for each aim

For clinical trials,

- Interim analyses/Early stopping rules

Engage a statistician to write or at least review these sections.
Common statistical problems in proposals

- Sample size justification absent or insufficiently justified
- Lack of statistical analysis plan for all aims, including secondary aims
- Inappropriate statistical analysis methods
Common statistical problems in proposals

- Sample size justification absent or insufficiently justified
- Lack of statistical analysis plan for all aims, including secondary aims
- Inappropriate statistical analysis methods

These issues can doom a proposal.
Data Collection and Compilation

- **Valid results require**
  - Collection of accurate data
  - Clear and accurate data compilation

- **Create workable and documented data sets**

- **QA/QC procedures**
  - Validation during data entry
  - Periodically audit the data
  - Conduct internal validation of final data
### Comparison of Drug A and Drug B

<table>
<thead>
<tr>
<th></th>
<th>Drug A</th>
<th>Age of Patient</th>
<th>Gender</th>
<th>Height (inches)</th>
<th>Weight (pound)</th>
<th>24hrct</th>
<th>blood pressure</th>
<th>tumor stage</th>
<th>Race</th>
<th>Date</th>
<th>complications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></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>Drug A</td>
<td>Age of Patient</td>
<td>Gender</td>
<td>Height (inches)</td>
<td>Weight (pound)</td>
<td>24hrct</td>
<td>blood pressure</td>
<td>tumor stage</td>
<td>Race</td>
<td>Date</td>
<td>complications</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>25</td>
<td>Male</td>
<td>61''</td>
<td>&gt;350</td>
<td>38%</td>
<td>120/80</td>
<td>2-3</td>
<td>Hispanic</td>
<td>1/15/99</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>65+</td>
<td>female</td>
<td>5'8''</td>
<td>161</td>
<td>32</td>
<td>140/90</td>
<td>II</td>
<td>White</td>
<td>2/05/1999</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>?</td>
<td>Male</td>
<td>120cm</td>
<td>12</td>
<td>&gt;160/110</td>
<td>IV</td>
<td></td>
<td>Black</td>
<td>Jan 98</td>
<td>yes, pneumonia</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>31</td>
<td>m</td>
<td>5'6''</td>
<td>obse</td>
<td>40</td>
<td>40 sys 105 dias</td>
<td></td>
<td>African-Arr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>42</td>
<td>f</td>
<td>&gt;6 ft</td>
<td>normal</td>
<td>39</td>
<td>missing =&gt;2</td>
<td></td>
<td>W</td>
<td>Feb 99</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>45</td>
<td>f</td>
<td>5.7</td>
<td>160</td>
<td>29</td>
<td>80/120</td>
<td>NA</td>
<td>B</td>
<td>last fall</td>
<td>n</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>unknown</td>
<td>?</td>
<td>6</td>
<td>145</td>
<td>35</td>
<td>normal</td>
<td>1</td>
<td>W</td>
<td>2/30/99</td>
<td>n</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>55</td>
<td>m</td>
<td>72</td>
<td>161.45</td>
<td>12/39</td>
<td>120/95</td>
<td>4</td>
<td>African-Arr</td>
<td>6-15-00</td>
<td>y</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>6 months</td>
<td>f</td>
<td>66</td>
<td>174</td>
<td>38</td>
<td>160/110</td>
<td>3</td>
<td>Asian</td>
<td>14/12/00</td>
<td>y</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>21</td>
<td>f</td>
<td>5'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></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></td>
<td>Drug B</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></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></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></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>4</td>
<td>44</td>
<td>na</td>
<td>65</td>
<td>?</td>
<td>40</td>
<td>120/80</td>
<td>2</td>
<td>?</td>
<td>9/14/99</td>
<td>n</td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>66</td>
<td>fem</td>
<td>71</td>
<td>0</td>
<td>41</td>
<td>140/90</td>
<td>4</td>
<td>w</td>
<td>Sep 14th</td>
<td>y, sepsis</td>
</tr>
<tr>
<td>22</td>
<td>6</td>
<td>unknown</td>
<td>?</td>
<td>172</td>
<td>199</td>
<td>38</td>
<td>&gt;160/110</td>
<td>3</td>
<td>b</td>
<td>unknown</td>
<td>y, died</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
<td>45</td>
<td>m</td>
<td>?</td>
<td>204</td>
<td>32</td>
<td>40 sys 105 dias</td>
<td>1</td>
<td>b</td>
<td>12/25/00</td>
<td>n</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>34</td>
<td>m</td>
<td>NA</td>
<td>145</td>
<td>36</td>
<td>130</td>
<td>3</td>
<td>w</td>
<td>July 97</td>
<td>n</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>13</td>
<td>m</td>
<td>66</td>
<td>161</td>
<td>39</td>
<td>166/115</td>
<td>2a</td>
<td>w</td>
<td>06/06/99</td>
<td>n</td>
</tr>
<tr>
<td>26</td>
<td>10</td>
<td>66</td>
<td>m</td>
<td>68</td>
<td>176</td>
<td>41</td>
<td>1120/80</td>
<td>3</td>
<td>w</td>
<td>01/21/58</td>
<td>n</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Average</td>
<td>45</td>
<td></td>
<td>65</td>
<td>155</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RedCAP is user-friendly alternative.

<table>
<thead>
<tr>
<th>Event Name: <strong>Baseline</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient Number</strong></td>
</tr>
<tr>
<td>(To rename this record, modify the value immediately below.)</td>
</tr>
<tr>
<td><strong>Patient Information</strong></td>
</tr>
<tr>
<td><strong>Patient Number</strong></td>
</tr>
<tr>
<td><em>must provide value</em></td>
</tr>
<tr>
<td>Identifying number assigned to the patient</td>
</tr>
<tr>
<td><strong>Patient Name</strong></td>
</tr>
<tr>
<td>Given name</td>
</tr>
<tr>
<td><strong>Date of birth</strong></td>
</tr>
<tr>
<td>MM-DD-YYYY</td>
</tr>
<tr>
<td><strong>MRN</strong></td>
</tr>
<tr>
<td>Assigned by UCDMC (some patients may not have this)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
</tr>
<tr>
<td>Ethnicity of patient</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
</tr>
<tr>
<td>Gender of patient</td>
</tr>
</tbody>
</table>
Leverage informatics and biostatistics expertise

- Medical informatics group can
  - Create forms for data collection
  - Extract information from EMR

- Use inter-disciplinary team to determine what information to collect and how
  - Investigator, practitioners, biostatistician, informatics specialist
  - Ensures information is collected and compiled in a manner that facilitates analysis
Data Do’s and Don’ts

• Use RedCAP where possible
• Assign unique ID to each subject and use consistently
• Remove all PHI prior submitting to statistician
• Unambiguously and consistently specify missing values
  – Avoid using “0” or blanks for missing values
• Avoid free text fields
Each data set needs a Code Book

Guidelines for Data Set Documentation
http://biostats.ucdavis.edu/DatasetDocumentation.php

- Name of the data file as it is stored on the computer
- Name of the code book's author, including contact information
- Date the code book was last updated.
- Number of records in the data file.
  - List of variables, including for each variable
  - Variable name
  - Location of variable, length of field
  - Allowable range for data
  - Missing data codes
  - Interpretation of values if not obvious (e.g. 1-male, 2-female.)
Go forth and collect data!
Statistician activities during the study

• Sometimes limited involvement by statistician

• Involvement can include
  – Conducting interim analyses
  – Serving on DSMB

• Some study designs entail periodic reassessments and statistician will necessarily be involved during the study
  – Two-stage, adaptive or sequential trials
Analysis and Reporting

(P.S. This is the fun part.)
Analyzing the data

- Conduct statistical analyses
  - Data validation
  - Run statistical tests
- Interpret the results
- Prepare tables and figures to illustrate findings
Working with a Statistician to Analyze your Data

Range of support provided

One-time
Advice Only

Select methods
Conduct all analyses
If statistician analyzes the data...

- Remove PHI
- Provide "clean" data set
- Provide data dictionary
- Allow sufficient time for analysis
  - Rule of Thumb is 4 to 6 weeks
- Provide references from similar studies if available
- Iterative and interactive process
Incorrect method results in incorrect conclusions.

Data with special statistical issues

- **Correlated values**
  - Multiple measurements on same subject
  - Clustering of subjects (e.g., same hospital, same litter)

- **Longitudinal data**
  - Repeated measurements over time on same subject

- **Missing data and drop outs**

- **High-dimensional data**
  - microarrays, proteomics
Report/Publication Preparation

- Craft statistical analysis section
- Contribute to results section
- Generate tables and figures
Resources at UC Davis

- Clinical and Translational Science Center
  - Biostatistics Workshop: 12-1 on Tuesdays
    sltaylor@ucdavis.edu
  - Biostatistics Core
    - Assist with study design, grant writing, data analysis and interpretation
    - Application for Resource Use (AFRU)
      http://www.ucdmc.ucdavis.edu/ctsc/

- Division of Biostatistics
Parting Tidbits
What we can’t do.

Fix a poorly designed study

Make something significant

Tell you what’s important

P value < 0.05
What we won’t do.

Find a significant result, i.e., “data snooping”

Clean up your data

Troubleshoot your code
Requests that Seriously Irritate Us

- “My abstract is due Friday. Can you analyze my data by tomorrow?”
- “I have a really simple question. It’ll only take 15 minutes.”
- “My paper got rejected because of the statistics. Can you fix it?
- “I keep getting error messages with my code. Can you look at it?”
What we love to do.

Help investigators achieve goals

Collaborate to bring new insights

Contribute to societal improvements
Example Project #1

Traumatic Brain Injury Outcomes
How do TBI patient outcomes differ with Glasgow Coma Scores at admission?

- What is your outcome?
How do TBI patient outcomes differ with Glasgow Coma Scores at admission?

- What is your outcome? Survival
- Survival to when?
  - 60-day?
  - 30-day?
  - Discharge?
How do TBI patient outcomes differ with Glasgow Coma Scores at admission?

- What is your outcome? Survival
- Survival to when?
  - 60-day?
  - 30-day?
  - Discharge
- What specifically do we want to test?
Suppose main interest is survival differences between moderate and good GCS groups.
Specific Aim 1: Determine if survival to discharge differs between patients with moderate (9-12) and good (13-15) Glasgow Coma Scores.

Endpoint: Survival to discharge

Testable Hypothesis:
$H_0: p_{mod} = p_{good}$ versus $H_a: p_{mod} \neq p_{good}$

Statistical Test: Two-sample Proportion Test
How do we test this hypothesis?

- How about a prospective study?
  - How could you do this?
  - Follow patients that come in with TBI and GCS of 9 or greater

- What issues/questions arise with a prospective approach?
  - How many patients necessary?
  - How long to recruit this number?
  - What are the confounders? Can we control for them?
How do we test this hypothesis?

- How about a retrospective study?
  - How could you do this?
  - Extract EMR information for patients with TBI and a GCS of 9 or greater

- What issues/questions arise with a retrospective approach?
  - Do we have enough patients in the EMR?
  - What are the confounders? How can we account for them?
Prospective vs. Retrospective

**Prospective**
- Ensure ALL desired data collected
- Current information
- Randomization
- Matching or stratification

**Retrospective**
- Large sample size
- Data available now
- Easier logistically
- Confounding factors
Prospective vs. Retrospective

Prospective
- Ensure ALL desired data collected
- Current information
- Randomization
- Matching or stratification

Retrospective
- Large sample size
- Data available now
- Easier logistically
- Confounding factors

For this particular study, a retrospective approach makes the most sense.
Sample Size Requirements

- Prospective studies – how many to recruit
- Retrospective studies – do we have enough
- Information needs
  - Survival of patients with moderate GCS
  - Difference in survival you want to detect
Inclusion/Exclusion Criteria

**Include**
- TBI with GCS 9+
- 16+ years old
- Arrival at Level 1 trauma center, including transfers

**Exclude**
- No TBI or GCS < 9
- Under 16 years old
- Penetrating injury
- Burn injury
- Others?
What are other potential confounding factors?

- Injury severity
- Age
- Gender
- Comorbidities
What are potential confounding factors?

- Injury severity
- Age
- Gender
- Comorbidities

Statistics can’t help a woefully unbalanced study
Next Steps

- Develop statistical analysis plan
- Work with Informatics and Biostatistics
  - Identify data to extract from the EMR
  - Develop data dictionary and supporting data forms
- Conduct data validation
- Statistically analyze the data
Example Project #2

Toradol Use for Rib Fractures
Does Toradol reduce the incidence of pneumonia in patients with rib fractures?

- What is the endpoint of interest?
- Prospective or retrospective?
Does Toradol reduce the incidence of pneumonia in patients with rib fractures?

- What is the endpoint of interest?
- Prospective or retrospective?
  - Prospective preferred (Why?)
Does Toradol reduce the incidence of pneumonia in patients with rib fractures?

- What is the endpoint of interest?
- Prospective or retrospective?
  - Prospective preferred (Why?)
- What issues arise for a retrospective study?
  - Comparability of groups
  - Age, injury severity, comorbidities
What if the data look like this?

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>50</td>
<td>15</td>
</tr>
<tr>
<td>60</td>
<td>20</td>
</tr>
<tr>
<td>70</td>
<td>15</td>
</tr>
<tr>
<td>80</td>
<td>10</td>
</tr>
</tbody>
</table>
Take Home Messages

- **Statistics matter**
  - Study design, sample size, appropriate analytical methods

- **Consult earlier rather than later**
  - Greatest value from consulting early
  - Can’t fix a poorly designed study

- **View statisticians as collaborators**
  - Authorship, funding support