DATA PREPARATION FOR YOUR BIOSTATISTICIAN

ROCS Biostats Team
OVERVIEW

1. ROCS Biostatistics Services
2. Rules of Collaboration
3. Excel Data Cleaning
4. Developing a Data Dictionary
5. Preparing to Discuss Your Analysis
6. Data Sharing Guidelines
Suhong Tong, MS
Research Senior Instructor, Dept of Pediatrics

Areas of expertise:
• Large data
• Longitudinal data
• Time series
• Survey analysis with population complex design
• Structural equation modeling
• Quality improvement analysis
• Survival analysis

Kaci Pickett, MS
Research Instructor, Dept of Pediatrics

Areas of expertise:
• Survival analysis
• Dynamic prediction
• Statistical consulting

Emily Cooper, MS
Research Instructor, Dept of Pediatrics

Area of expertise:
• Large Data
• Statistical consulting

Samantha Bothwell, MS
Research Instructor, Dept of Pediatrics

Area of Expertise
• Time Series
• Longitudinal Data
• Clustering Analysis
• Spatial Statistics
• Statistical Consulting
The biostatistician can collaborate in all parts of protocol development and implementation, not just the statistics!

- Sample size calculation
- Analysis plan and protocol development
- Big data analysis (PHIS, TQIP)
- Retrospective data consulting
- Prospective data consulting
- Randomization schedules
- Database structure and review (REDCap)
- Clinical trial and general study design
- Abstract and poster development
- Manuscript and grant preparation
RULES OF COLLABORATION WITH ROCS BIOSTATISTICIANS

1. Involve your biostatistician early and communicate often!
2. Data should be stored in REDCap!
   • If your data is already in Excel, the data should be cleaned and include a data dictionary
3. Be considerate of the time we need to complete your analysis!
   • For simple analyses, we need at least 4 weeks from receiving clean data
4. Your biostatistician should be considered a co-investigator
   • Considered co-authors on manuscripts, usually 2nd author
5. Notify us of any co-authored submissions and acceptance, even for abstracts
**RULES OF COLLABORATION WITH ROCS BIOSTATISTICIANS**

Hours of work doesn’t equal weeks of work!

<table>
<thead>
<tr>
<th>Request</th>
<th>Minimum Time Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power/Sample size calculation</td>
<td>10-20 hours</td>
</tr>
<tr>
<td>Protocol and analysis plan development</td>
<td>20-40 hours</td>
</tr>
<tr>
<td>Data analysis (depending on complexity)</td>
<td>40-120 hours +</td>
</tr>
<tr>
<td>Abstract assistance</td>
<td>10-40 hours</td>
</tr>
</tbody>
</table>
EXCEL DATA CLEANING

Eight Tips for Creating Clean Data in Excel:

1. Create concise variable names
   • Ideally 4-20 characters long
   • Should not include special characters or spaces

2. Leave cells blank to indicate missing values
   • May also use a standard value like 9999 or NA

3. For categorical variables, use shorthand notation to label categories
   • Use numbers (1, 2) or individual letters (M, F)
   • Be consistent in using uppercase or lowercase!
EXCEL DATA CLEANING

Eight Tips for Creating Clean Data in Excel:

4. Ensure dates use identical formatting, such as MM-DD-YYYY or MM/DD/YYYY
   - Use Excel date formatting by using Home > Number > Format Cells > Date

5. For select all that apply questions, each option should be a separate yes/no variable
EXCEL DATA CLEANING

Eight Tips for Creating Clean Data in Excel:

6. Free text should only be used for notes that won’t be used in analysis!
7. Don’t highlight or color-code your data!
   • Instead, add variables to indicate this
8. Always keep an original copy of your data, even if it’s messy!
How to identify data entry errors:

• For numeric variables, check the minimum and maximum values using =MIN() and =MAX()

• For categorical variables, select a column and use Data > Sort & Filter > Filter and click the down arrow
Data cleaning can take a lot of time!

It is usually an iterative process between you and your biostatistician. Ask your biostatistician for a meeting to discuss data cleaning expectations before you start!
Your data dictionary is usually a separate Excel document or sheet!
It should be up-to-date and include the same information as a REDCap codebook!

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Variable Label</th>
<th>Variable Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>record_id</td>
<td>Unique record identifier</td>
<td></td>
</tr>
<tr>
<td>mrn</td>
<td>Medical record number</td>
<td></td>
</tr>
<tr>
<td>dob</td>
<td>Date of birth</td>
<td>MM-DD-YYYY</td>
</tr>
<tr>
<td>sex</td>
<td>Patient sex</td>
<td>1=Female, 2=Male</td>
</tr>
<tr>
<td>ethnicity</td>
<td>Ethnicity</td>
<td>H=Hispanic or Latino, N=Not Hispanic or Latino, U=Unknown</td>
</tr>
<tr>
<td>height</td>
<td>Height (cm)</td>
<td></td>
</tr>
<tr>
<td>iss</td>
<td>Injury severity score</td>
<td>0-75</td>
</tr>
</tbody>
</table>
Main types of data analysis we see in ROCS:

1. Descriptive statistics
2. Hypothesis testing
3. Correlations
4. Regression analyses
5. Survival analyses
6. Publication-ready figures

If you don’t know what type of analysis you’d like, we’re here to help!
Things to discuss with your biostatistician:

• What is your primary aim?
  • When possible, provide your IRB protocol to your biostatistician!

• Is there similar published literature?
  • Consider looking at adult studies!
  • Helpful for identifying covariates/potential confounders and information on standard statistical tests

• What have previous researchers found?
PREPARING TO DISCUSS YOUR ANALYSIS

Things to discuss with your biostatistician:

• Will some patients have repeat events, admissions, or measurements?
• What is your primary outcome?
  • If your outcome is continuous, what is a clinically meaningful difference?
• What are your inclusion and exclusion criteria?
PREPARING TO DISCUSS YOUR ANALYSIS

Things to discuss with your biostatistician:

• Should the analysis be stratified? Should the analyses be performed separately within groups?
  • Are there known cofounders?
• Do you want to perform a sub-analysis? Do you want to repeat the analyses in a smaller group?
  • Is there a subset of your study population that may respond better to your intervention?
Regardless of what you find in similar literature, understand that a specific analysis may not be feasible for your study!

Your biostatistician has the expertise to know when and how to appropriately implement analyses! We will consider:

• Sample size
• Study type
• Statistical assumptions
DATA SHARING GUIDELINES

Data containing PHI should not be shared with ROCS biostatisticians via email!

Some alternative methods:

• OneDrive
• Microsoft Teams
• REDCap Send-It

Reach out to ARC for specific information on data sharing best practices!
Thank you!

Questions?