2022 Research Boot Camp Series: Data Preparation for your Biostatistician

Presented by:

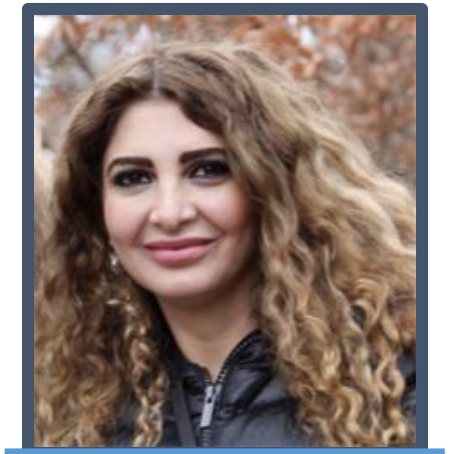
Research in Outcomes for Children's Surgery (ROCS)

Center for Children's Surgery

Data Preparation for your Biostatistician

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

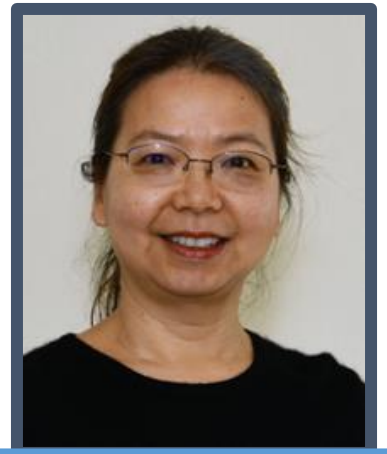


Souha Fares, PhD

Assistant Professor, Biostatistics & Informatics, School of Public Health

Areas of expertise:

- Research study design
- Protocol development
- Time series analysis
- Psychometric analysis
- Statistical consulting



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:

- Omics
- Large Data
- Statistical consulting

Jared Rieck, BS

Student Research Assistant, Biostatistics & Informatics, School of Public Health

Area of expertise:

- Statistical consulting

1: ROCS Biostatistics Services

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

2: 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 as 2nd author.

2: Rules of Collaboration with ROCS Biostatisticians

Hours of work doesn't equal weeks of work!

Request	Minimum Time Required
Power/Sample size calculation	10-20 hours
Protocol and analysis plan development	20-40 hours
Data analysis (depending on complexity)	40-120 hours +
Abstract assistance	10-40 hours

3: Excel Data Cleaning

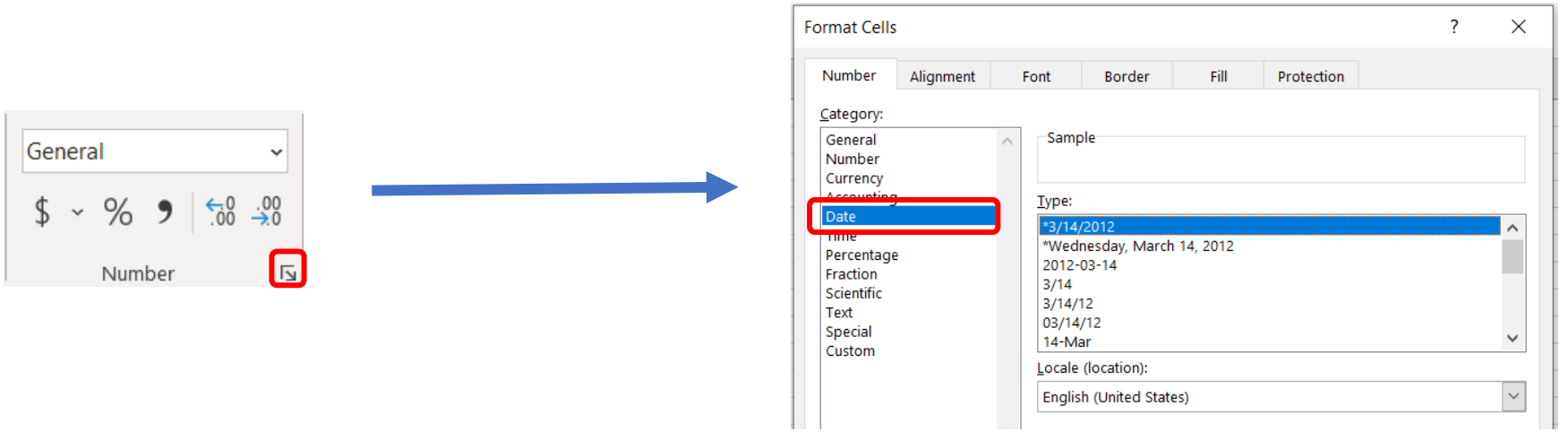
Eight Tips for Creating Clean Data in Excel:

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!

3: 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.

3: Excel Data Cleaning

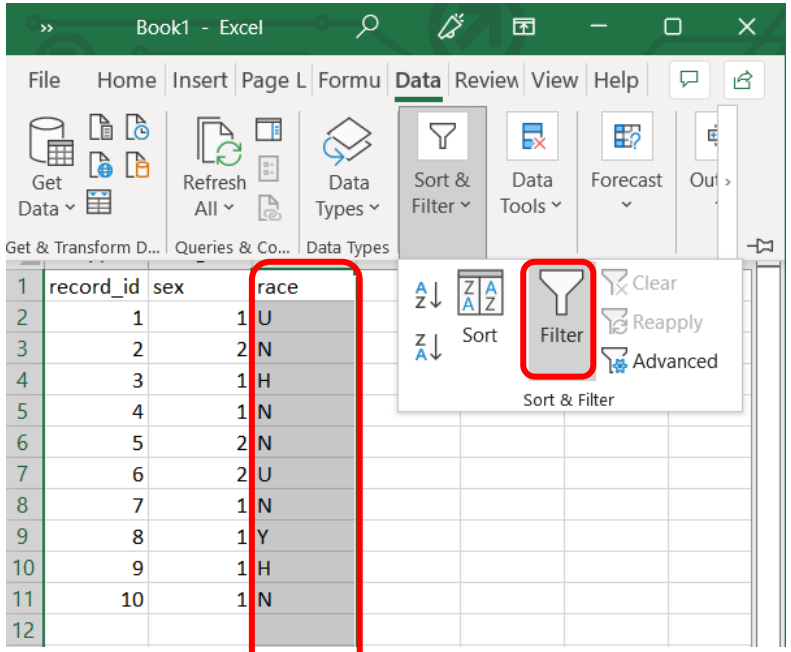
Eight Tips for Creating Clean Data in Excel:

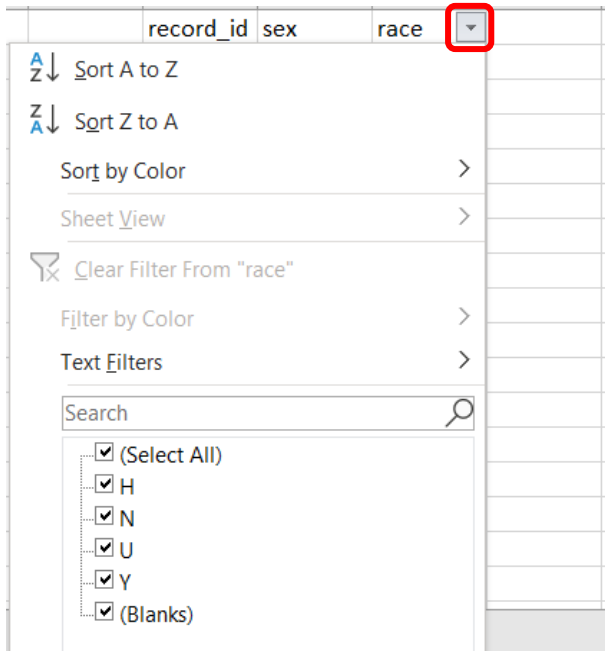
6. Free text should only be used for notes that won't be used in analysis!
7. Don't highlight or color-code your data!
    - Instead, add variables to indicate this
8. Always keep an original copy of your data, even if it's messy!

3: Excel Data Cleaning

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 





3: Excel Data Cleaning

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!

4: Developing a Data Dictionary

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!

Variable Name	Variable Label	Variable Attributes
record_id	Unique record identifier	
mrn	Medical record number	
dob	Date of birth	MM-DD-YYYY
sex	Patient sex	1=Female, 2=Male
ethnicity	Ethnicity	H=Hispanic or Latino, N=Not Hispanic or Latino, U=Unknown
height	Height (cm)	
iss	Injury severity score	0-75

5: Preparing to Discuss your Analysis

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!

5: Preparing to Discuss your Analysis

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?

5: Preparing to Discuss your Analysis

Things to discuss with your biostatistician:

- Will some patients have repeated 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?

5: 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 confounders?
- 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?

5: Preparing to Discuss your Analysis

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

6: 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!

Questions?