Biostatistical Resource Guide

General Considerations for Database Development

- Know what you are collecting, why you are collecting it, and what you will use it for analytically. While you should be careful to include everything on the front-end that you might need, consider that the more you variables you have, the more data entry there is to be done.
- You should ensure that your database solution addresses issues related to data provenance (meaning the origin and history of the data as well as the steps used in the analysis or experiment). Be sure to include some combination of data quality assurance strategies when designing your database system (*see "how can I ensure the quality of my data" section below for more information).*
- Keep patient confidentiality and HIPAA regulations in mind when planning and creating your database. Ensure that your database is stored in a secure folder that only appropriate study personnel can access. Also remember that electronic files containing identifiable information need to be encrypted prior to e-mailing. Possible ways to encrypt your files include encryption options in WinZip, Lotus Notes (only works between Lotus Notes users), or a strong encryption program such as PGP (http://www.pgp.com/). When sending data files, it's best to include as little identifiable information as possible. You should only include variables needed for analysis.
- Make sure you backup your database on a regular schedule.

What Software Should I Use for My Database?

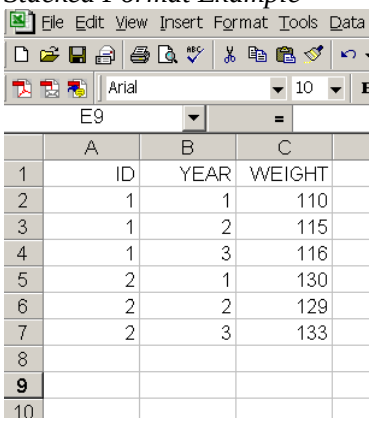
- Microsoft Excel versus relational database systems (Microsoft Access, FileMaker Pro, Oracle, Microsoft SQL, REDCap)
    - One advantage of relational database systems (such as Microsoft Access) is that they allow the inclusion of data validation rules that help minimize data entry errors.
        - One primary disadvantage of relational database systems is that because of their flexibility and sophistication, it generally requires more time investment/technical expertise to set up databases using them.
    - Microsoft Excel is easy to use but the default settings within Excel can cause downstream problems when it comes to data analysis. **This has important implications for data provenance (meaning the origin and history of the data as well as the steps used in the analysis or experiment).**
        - Not easy to set up data validation rules to prevent data entry errors.
        - Easy to set up data in Excel tables in a format that will be difficult to import into statistical analysis software later

Tips for Using Microsoft Excel (If you insist on using it despite its flaws…)

- A few do's and don'ts using Microsoft Excel
    - Keep your column headings short and descriptive.
    - Keep your Excel sheets clean... don't use **boldface** or other types of formatting such as decorative lines or shading. These can result in complications when importing data into statistical analysis software.
    - Don't include extraneous graphs or summary statistics in the same Excel sheet as your raw data. Do your graphing and calculations in a separate Excel sheet.
    - If you are collecting longitudinal data (e.g., repeated measurements on the same subject over time), there are downstream benefits to collecting this data in a "stacked format" instead of a "one record per subject format". One possibility is to collect all baseline data in one Excel sheet using the "one record per subject format". Keep your longitudinal data in a separate

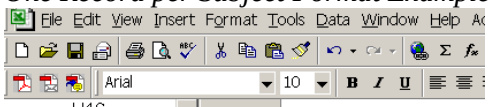
Excel sheet using the "stacked format". As long as you use the same unique identifier (in the sample below its ID) in each sheet, it's easy to link the data back together once it has been imported into a statistical analysis package).

Stacked Format Example



	A	B	C
1	ID	YEAR	WEIGHT
2	1	1	110
3	1	2	115
4	1	3	116
5	2	1	130
6	2	2	129
7	2	3	133
8			
9			
10			

One Record per Subject Format Example



	A	B	C	D	E
1	ID	WEIGHT1	WEIGHT2	WEIGHT3	
2	1	110	115	116	
3	2	130	129	133	
4					
5					

Creating a Data Dictionary for Your Study

- The Database Dictionary defines the common terms, codes, and conventions used in the database.

Sample Below

Field name	Description of field and values	Normative Ranges
Group	0 = Control 1 = Dose Level 1 2 = Dose Level 2 3 = Does Level 3	
DayOfWeek	The day of week 1 = Monday 2 = Tuesday 3 = Wednesday 4 = Thursday 5 = Friday 6 = Saturday 7 = Sunday	
Weekday	0 = {Saturday, Sunday }, 1 = {Monday ... Friday}	

Height	Subject height in inches	
Weight	Subject weight in kilograms	
Gender	0 = Female 1 = Male	
Race	0 = White 1 = Non-white	

- Include information in the data dictionary about normative ranges for continuous variables (this information will be used for data cleaning purposes during statistical programming).
- It's important to distinguish between raw variables and derived variables. Generally, it's a good idea to describe the derived variables in a separate section of the data dictionary. Derived variables should include explicit definitions of how the variables are constructed.
- Consistency is crucial for data entry. For example, if the variable is 'DayOfWeek', and you enter Monday as 'M' part of the time and 'Mo' other times, there will be downstream problems. It's best to use shorter values and always be consistent. Follow the coding conventions laid out in your data dictionary.
- From an analytic standpoint, using numbers for categories can be advantageous as it can make certain calculations within statistical software easier to accomplish. For example, for a dichotomous variable, one could use '0' for NO and '1' for YES.
- It's a good idea to code missing values with a unique code such as '.' or '-999' so it's clear exactly which values are true missing values.
- Free-form text fields are difficult to analyze quantitatively.

How Can I Ensure the Quality of My Data?

- The use of data quality assurance strategies is an important component of database development. A variety of strategies can be employed:
    - Validation rules (range checks, required input fields, allowable values) and consistency checks (checking for logically inconsistent patterns of response) can be designed (1) to operate at the time of data entry or (2) to clean the data after data entry has occurred.
        - *At time of entry:*
          Many database programs have the capability to incorporate validation rules and consistency checks into the data entry forms that operate in real-time as data is entered.
        - *After entry has completed:*
          A written list of validation rules and consistency checks can be given to the statistical programmer along with the dataset to be used for data cleaning.
    - Double data entry is a widely used strategy to enhance the quality of study data. This may be used for all variables in the study or only for a small subset of critical variables. Some database programs provide built-in capability for double data entry and discrepancy resolution. If using simple database solutions, such as Microsoft Excel, simply enter the data twice in separate Excel sheets. A statistical programmer can compare the sheets using statistical procedures which will produce a discrepancy report.

When It Comes Time to Transfer the Database to the Statistician, the More Things on This List That You Can Provide (in Addition to the Actual Data Files), the More Efficient and Productive the Collaboration Will Be

- Data dictionary
- Summary of hypotheses to be tested, which should include:
    - Overview of study design (e.g., methods section of grant)
    - Primary outcomes of interest
    - When primary outcomes were observed (how many time points)
    - Other relevant variables (potential covariates)
    - Assessment of the magnitude of difference between groups on the outcome that translates into a meaningful clinical difference?
    - Proposed table shells for manuscript

Another resource to use is the DTMI Biostatistics Core:
(https://www.dtmi.duke.edu/for-researchers/quantitative-resources/biostatistics-core)

You can find their Data Submission Guidelines document here:
https://www.dtmi.duke.edu/for-researchers/quantitative-resources/biostatistics-core/DataSubmissionGuidelines.pdf/view

What Statisticians Do And Why You Need One

Outline

- Why you need a statistician on your clinical research project
- A detailed description of what statisticians do throughout the research process
  - Delineates between PhD statistician roles and Master's level statistician roles
- A chart providing budgetary guidance for statistical grant support for different types of NIH funding mechanisms

Why you need a statistician

- Grant Proposals
  - Most grant funding agencies (public & private) have regular statistician members on their study section rosters:
    - It is almost a certainty that your grant will be reviewed by someone with expertise in statistics.

- Manuscripts
  - Most prominent medical journals require a statistical review of all submitted manuscripts to evaluate the appropriateness of the statistical methodology used.
  - The CONSORT Statement is an evidence-based, minimum set of recommendations for reporting randomized controlled trials (RCTs) that is endorsed by prominent general medical journals, many specialty medical journals, and leading editorial organizations (see www.consort-statement.org).
  - Below are several statistically related items required by the CONSORT checklist:
    - How sample size was determined
    - Statistical methods used to compare groups for primary outcome(s)
    - Method used to generate the random allocation sequence
    - Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory.

What Statisticians Do

Stages of Research Protocol	Role	Activities
Grant development/writing	PhD	•Help formulate and refine an appropriate study design and provides guidance on handling of special design issues (e.g., clustering, randomization, treatment blinding, attrition) •Help generate/refine testable research hypotheses •Develop and write statistical analysis section of grant, including proposed analytic procedures for main study hypotheses and handling of special issues (missing data, bias, noncompliance, intention to treat) •Perform sample size calculation and writes sample size section of grant
	MS	•Analyze pilot data for grant under supervision of PhD
Study planning	PhD	•Provide input on database development •Supervise and/or implement randomization procedures •Provide input on other study start up procedures/issues
	MS	•Work with database programmers to ensure appropriate quality control procedures are implemented •Provide input on variable definition, naming, and coding
Study execution/monitoring	PhD	•Attend regular study and DSMB meetings •Consult with and provide needed reports to DSMB for study monitoring •Provide input on handling protocol deviations •Provide input on need for and appropriateness of study modifications •Help with analysis/preparation of abstracts •Oversee activities of MS statistician •Perform administrative functions and/or supervise study team activities beyond MS statistician
	MS	•Attend regular study meetings •Monitor study enrollment for quality assurance •Prepare interim study reports describing accrual numbers and completeness of data collection •Help with analysis/preparation of abstracts

What Statisticians Do

Stages of Research Protocol	Role	Activities
Statistical analysis of data during and at end of study	PhD	•Perform any planned interim analyses •Supervise/oversee programming/analysis of the MS statistician personnel •Supervise/perform main study analyses •Identify opportunities for improved statistical methodology •Develop new statistical methods needed for proper analysis and interpretation of data
	MS	•Responsible for importing completed database into statistical software in preparation for programming and analysis •Perform statistical programming (data checking for accuracy, data manipulation, variable creation, etc) to prepare data for statistical analyses •Perform statistical analyses under supervision of PhD statistician
Manuscript writing	PhD	•Serve as co-author on primary and secondary manuscripts arising from study •Write statistical methods section of manuscripts •Help write results section of manuscript, including specification of needed tables, figures, and graphs •Review and edit manuscript •Participate in the manuscript review and resubmission process, particularly in response to reviewers
	MS	•Serve as co-author on primary and secondary manuscripts arising from study •Generate needed statistical reports, tables and figures for manuscripts •Verify numbers in the manuscript are accurate

Budgeting for Statistical Support on a Research Proposal

Type of Proposal	Examples	Approximate Percent Effort*
Career development awards	•Mentored Clinical Scientist Development Program Award (K12)	•Statistical mentoring usually donated •Analytic support ~5-10% PhD
Single site study	• Exploratory/Pilot study (R03, R21) • RCT with 2-3 treatment arms (R01) • Prospective cohort observational study (R01) • Secondary database study (R03, R21, R01)	10-20% PhD 20-25% MS
•Multi-site study •Complex study design	• R01 with special issues requiring nonstandard/complex analyses, including: –Clustered data –Longitudinal outcomes –Time-varying covariates –Causal modeling of nonrandomized/observational data –Extensive missing data –New methodology required to properly analyze data	20-25% PhD 35-50% MS
Center grant with biostatistics core	P01 (integrated, multi-project involving multiple independent investigators who share common resources)	~25% (or more) PhD ~50% (or more) MS

*** These represent approximate/rough percent effort by type of proposal. Actual percent effort is calculated on a case-by-case basis, depending on the specific needs of the project.**