Working With Your Statistician:
How we can make each others’ jobs easier

Jeannie-Marie Leoutsakos, PhD MHS
Assistant Professor, Department of Psychiatry and Behavioral Sciences
Director, Psychiatry Data Core
Questions

- How many of you have a statistician working as part of your group?
- How many of you work with a statistician outside your group?
- Does the statistician become involved before or after the data are collected?
- How many of you also act as the statistician for your group?
- What questions are you hoping will be answered today?
Outline

* My Background
* Statisticians at Johns Hopkins
* Ideal and Non-Ideal Collaborations, things to keep in mind.
* Specific Recommendations
  * Data Coding
  * Data Documentation
  * Data Delivery
* Questions?
How I got here

- 1993-7 Pre-Med/CogSci at Homewood
- 1997-0 Started work at JHH (Research assistant, data manager, data analyst, network administrator)
- 2000-3 Biostat master’s at JHSPH
- 2003-7 Mental Health PhD at JHSPH
- 2007-9 Postdoc in Psychiatry
- 2009- Data-Core/Teaching/Methods Research
(Bio)statisticians at Hopkins

- 53 statistician/biostatistician
- 53 research data analysts
- 46 Biostatistics Faculty
- 100 Biostatistics Students
- 20 Research Data Manager
- 9 Database Specialists
- 100 Programmer Analysts
Collaborator: involvement throughout the project.
- Hypothesis Development/Grant writing
- Database setup
- Data Analysis
- Manuscript Preparation

Teacher:
- should be mutual and integrative

Non-Ideal Collaborations

* **Helper:** technician; responds to questions. Accountability problems.

* **Leader:** lack of substantive expertise.

* **Data-Blesser:** curb-side advice.

* **Archaeologist:** my other statistician stopped returning my e-mails…
Timeline for Collaboration

- Throughout the life of the project / end-product focused
- Assist PI with hypothesis development/study design
- Consult on database design with PI & DBM
  - Check that necessary variables are present, etc.
  - Check that unnecessary variables are not included
  - Statistician can be your advocate – stressing important of data integrity to PI
- Perform Interim analyses (if necessary)
- Perform Final analyses
- Assist in manuscript preparation
What Statisticians Know

• Some portion of statistics(!)

• May know little about databases, particularly your database software

• May have very circumscribed programming ability.

• May have little or no subject knowledge- don’t assume that they are familiar with certain variables or instruments/acronyms.
Specific Recommendations

- Database Software
- Variable Names/Value labels
- Data Documentation
- Datafile Version Control
- File Formats/Transmission of Data Files
* MS Excel – simple but limited, sorting problem, security
* MS Access, Filemaker Pro - labor intensive for DBMs
* Redcap – web-based, allows tracking, nice features
* CRMS – ?
* Statistician will likely convert what you give them to a statistical package (Stata/R/SAS, etc)
* May have memory issues: STATA/IC 2047 variables
* MAC/PC issues
Golden Rules

1. Will this be completely unambiguous to an outside person with little or no prior knowledge of the study?

2. Is this as consistent as possible? (both internally and externally)
Variable/Field Names

* Name Length Limits (should ask)
  * For SAS and STATA, now 32
  * Others: may be as low as 8
* Need to start with a letter, avoid CAPS and special characters (\#$&@+, esp *!)
* Use a consistent convention: e.g. Use first three characters to denote form (if you have multiple forms).
* For dichotomous variables, consider a category as the name: (e.g., instead of “sex” coded 0/1, use “male” coded as 0/1)
Be careful how you name variables and encode values that might be considered *sensitive*.

- Sex/gender/orientation
- Race/ethnicity
- Anthropometrics
Variable Formats

- May not matter if transformed to .txt or .csv file
- Numeric: byte, float, double
- Date: format should be explicit
- String/Text:
- Memo/extended text:
- **ALERT:** if database consists of multiple datafiles, ensure that variable names and formats of identifiers are consistent across all data files.
Extended Variable Name/Description

* Variable name: ham14
* Variable Label: “hamilton depression rating scale q. 14”

* Particularly useful with short variable name lengths
* Check to see if statistician’s software will read them
* Take note of label length limits (STATA: 80)
  * Use consistent convention
* Check to see if statistician’s software will accept them
* Use a convention, avoid CAPS
* Code functional values of dichotomous variables as 0/1
* Missing Data:
  * Can have multiple missing value codes: don’t know, refused, not applicable, etc
  * Value codes should be universal and sequential, and outside the possible range of non-missing data.
  * No fields should be intentionally left blank (except possibly due to skip patterns)
Data Documentation

- Codebook/Data Dictionary (ideally electronic and string searchable)
- Sample CRF (binder with data collection forms)
- Unresolved Queries/Issues
- Invalid Values
- Version Control
<table>
<thead>
<tr>
<th>Codebooks/Data Dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Range from v. elaborate to v. simple</td>
</tr>
<tr>
<td>* Variable Name</td>
</tr>
<tr>
<td>* Variable Description</td>
</tr>
<tr>
<td>* Variable Format (for dates, be careful and explicit as to 12/10/1975 vs 10/12/1975)</td>
</tr>
<tr>
<td>* Encoding (if any)</td>
</tr>
<tr>
<td>* Ranges, acceptable values</td>
</tr>
<tr>
<td>* Counts, Descriptives</td>
</tr>
<tr>
<td>* Value Labels</td>
</tr>
<tr>
<td>* Missing Data codes</td>
</tr>
</tbody>
</table>
Over 100 PDF files corresponding to each separate datafile
Study also collected data on participants’ spouses and caregivers
<table>
<thead>
<tr>
<th>SPSS Variable Label</th>
<th>Description</th>
<th>Coding</th>
</tr>
</thead>
</table>
| c1csdt              | Item: DATE  
                   Case Staffing: DATE OF 
                   SECOND INCIDENCE CLINICAL 
                   ASSESSMENT CASE 
                   STAFFING          | Date     |
| c1cc01              | Item: 1  
                   Case Staffing: Caseness 
                   Checklist: EVIDENCE OF 
                   MEMORY IMPAIRMENT FOR 
                   NEW EVENTS.         | 0 No     |
|                     |             | 1 Yes   |
|                     |             | 9 Missing |
| c1cc02chk           | Item: 2  
                   Case Staffing: Caseness 
                   Checklist: AT LEAST ONE OF 
                   THE FOLLOWING IS 
                   IMPAIRED           | 0 No     |
|                     |             | 1 Yes   |
|                     |             | 9 Missing |
| c1cc02              | Item:2_1  
                   Case Staffing: Caseness 
                   Checklist: AT LEAST ONE OF 
                   THE FOLLOWING IS 
                   IMPAIRED: ABSTRACT 
                   THINKING           | 0 No     |
|                     |             | 1 Yes   |
|                     |             | 9 Missing |
vname ;vlen;req;vtype;lo;hi;vlab
clinic ;3 ;r ; c ; ; ; 1 Field site ID
adapt ;5 ;r ; c ; ; ; 2 Participant ID
namecd ;5 ;r ; c ; ; ; 3 Participant name code
formdate ;7 ;r ; d ; ; ; 4 Date of contact
visit ;3 ;r ; c ; ; ; 5 Visit ID code
form ;3 ;r ; c ; ; ; 6 Form and revision
db1007 ;5 ;r ; c ; ; ; 7 Collateral respondent name code
db1008 ;7 ; ; d ; ; ; 8 Date of Telephone Assessment Contact
db1009a ;1 ; ; n ; 1; 1;9a Triggered by Telephone Assessment Battery (TAB)
db1009b ;1 ; ; n ; 1; 1;9b Self-report by participant
db1009c ;1 ; ; n ; 1; 1;9c Report by collateral respondent
db1009d ;1 ; ; n ; 1; 1;9d Decline in cognitive scores
db1009e ;1 ; ; n ; 1; 1;9e Field site staff noticed decline
db1009f ;1 ; ; n ; 1; 1;9f Other reason
db1009fs ;60 ; ; c ; ; ; 9f Specify other reason
db1010 ;1 ; ; n ; 1; 2;10 Has participant undergone cognitive testing
db1011 ;2 ; ; n ; 0; 60;11 Approximately how long ago did the testing occur
db1012 ;1 ; ; n ; 1; 2;12 Has participant been told that he/she needs a hearing aid
db1013 ;1 ; ; n ; 1; 2;13 Is participant wearing a hearing aid
db1014 ;1 ; ; n ; 1; 2;14 Did participant use the audio amplifier during testing
db1015 ;1 ; ; n ; 1; 2;15 Does participant wear corrective lenses
db1016 ;1 ; ; n ; 1; 2;16 Was participant wearing lenses during testing
db1017 ;1 ; ; c ; ; ; 17 What was the smallest line read by participant
db1018a ;4 ; ; n ;0100; 1259;18a Time DEB started for participant
db1018ap ;1 ; ; n ; 1; 2;18a am/pm time DEB started for participant
db1018b ;4 ; ; n ;0100; 1259;18b Time DEB completed for participant
db1018bp ;1 ; ; n ; 1; 2;18b am/pm time DEB completed for participant
db1019 ;3 ; ; n ; 001; 999;19 Total time for DEB for participant
ADAPT-FS

Dementia Evaluation Battery Results
(DB-1)

Purpose: Record the results of the Dementia Evaluation Battery (DEB) from the DEB booklet and DEB supplement.

When: Dementia Evaluation Visit (DEV).

By whom: Study psychometrician or neuropsychologist who administered the DEB.

Instructions: Calculate scores from the DEB and record below. Information for section B is obtained from the TB form and the DEB booklet. Information for sections C, D, and E is obtained from the DEB booklet. Information for section F is obtained from the DEB Supplement. See instructions in the ADAPT-FS handbook about how to assign the collateral respondent name code in section A. Refer to the ADAPT-FS Neuropsychology Manual for details on test scoring.

A. Field site, participant, collateral, and visit identification

1. Field site ID code: _____ _____ _____

2. Participant ID: _____ _____ _____ _____

3. Participant name code: ____________________________

10. Has the participant undergone cognitive testing other than in ADAPT-FS in the last 60 days:

   Yes (____)  No (____)

11. Approximately how long ago did the testing occur:

   _____ _____ days

12. [Blank]
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable / Field Name</td>
<td>Form Name</td>
<td>Field Units</td>
<td>Section</td>
<td>Field Type</td>
<td>Field Label</td>
<td>Choices OR Calculations</td>
</tr>
<tr>
<td>id</td>
<td>inclusion_exclusion_criteria</td>
<td>text</td>
<td>text</td>
<td>Participant ID Number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>datenpi</td>
<td>inclusion_exclusion_criteria</td>
<td>text</td>
<td>text</td>
<td>Date of Visit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>incl_1</td>
<td>inclusion_exclusion_criteria</td>
<td>Inclusion</td>
<td>radio</td>
<td>Diagnosis of AD by NINCDS/ADRDA criteria (47)</td>
<td>1, No</td>
<td>2, Yes</td>
</tr>
<tr>
<td>incl_2</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Age &gt;60. This excludes early-onset AD cases which</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>incl_3</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Mini-Mental State Exam (MMSE) 16-26. This range</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>incl_4</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Clinical Dementia Rating (CDR) &lt;= 1 (mild demer)</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>incl_5</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Patients will be allowed to remain on current FD</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>incl_6</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Patients will be allowed to remain on antidepressants</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>incl_7</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Knowledgeable informant available for all study</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>excl_1</td>
<td>inclusion_exclusion_criteria</td>
<td>Exclusion</td>
<td>radio</td>
<td>Evidence of non-AD dementias including Hunting</td>
<td>1, No</td>
<td>2, Yes</td>
</tr>
<tr>
<td>excl_2</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Current DSM-IV Axis I diagnoses other than dem</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>excl_3</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Any clinically significant medical condition that</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>excl_4</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Current use of Beta-blocking agents</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>excl_5</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Contraindications to use of Beta-blocking agent</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>excl_6</td>
<td>inclusion_exclusion_criteria</td>
<td>radio</td>
<td>Clinically significant hepatic or renal insufficienc</td>
<td>1, No</td>
<td>2, Yes</td>
<td></td>
</tr>
<tr>
<td>examiner</td>
<td>inclusion_exclusion_criteria</td>
<td>text</td>
<td>text</td>
<td>Examiner ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>idmm</td>
<td>mini_mental_state_examination</td>
<td>text</td>
<td>text</td>
<td>Participant ID Number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmse_date</td>
<td>mini_mental_state_examination</td>
<td>text</td>
<td>text</td>
<td>Date</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmse_01</td>
<td>mini_mental_state_examination</td>
<td>text</td>
<td>text</td>
<td>Time orientation (0-5):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmse_02</td>
<td>mini_mental_state_examination</td>
<td>text</td>
<td>text</td>
<td>Place orientation (0-5):</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Considerations for Longitudinal Datasets

Wide: 1 line per patient
Long: 1 line per visit

Visit indicator needs to be at the end of the var name stub.
Resolution of discrepancies between double data-entered files (if applicable)

Resolutions of missing data or aberrant values

Valid Data Indicators (e.g., lab values that are known to be erroneous – recommend second variable which contains an indicator as to whether that target variable value is legitimate/to be included in analyses)

Statisticians shouldn’t clean data
  * Inefficient
  * We don’t have enough knowledge about the data
There are likely things like totals, data calculations, etc that are calculated based on the entered data, rather than being entered.

Discuss with statistician – depending on which software you are both using, there may be things that are a lot easier for them to do later, or vice versa – e.g, Long/wide

Documentation should include exactly how these were calculated.
It is likely that there will be multiple versions of the dataset (e.g., interim, after cleaning)

A log of all generated versions should be kept, and dataset names should include the date.

Try to distribute only finalized versions of datasets
Be careful about HIPAA!

PMI includes dates and ages if >90

It may be necessary to create “days from baseline variable”

A dataset containing PMI cannot be e-mailed unless it is encrypted

Best bet: only distribute de-identified datasets
  * Redcap will create one for you automatically

If someone e-mails me an unencrypted dataset with PMI, I am obligated to report them.

Consider Jshare or Sharepoint for file distribution
Main Points

* Encourage your PI to develop a collaboration early.
* You should be involved in that collaboration
* You and the statistician can save each other time
* Useful data is well-documented data
Questions?

* How do you find a statistician?
* Anybody having a problem with a statistician right now?
* Interpersonal aspect of working with a statistician.
* Data Scientist career paths
* Statistical software packages