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

Ideal Collaborations

Collaborator: involvement throughout the project.

- Hypothesis Development/Grant writing
- Database setup
- Data Analysis
- Manuscript Preparation

Teacher:

- should be mutual and integrative

Kirk RE. (1991) Statistical consulting in a university: dealing with people and other challenges. *American Statistician 45(1):28-34.*

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

* thoughout the life of the project / end-product focused
* Assist PI with hypothesis development/study design 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.

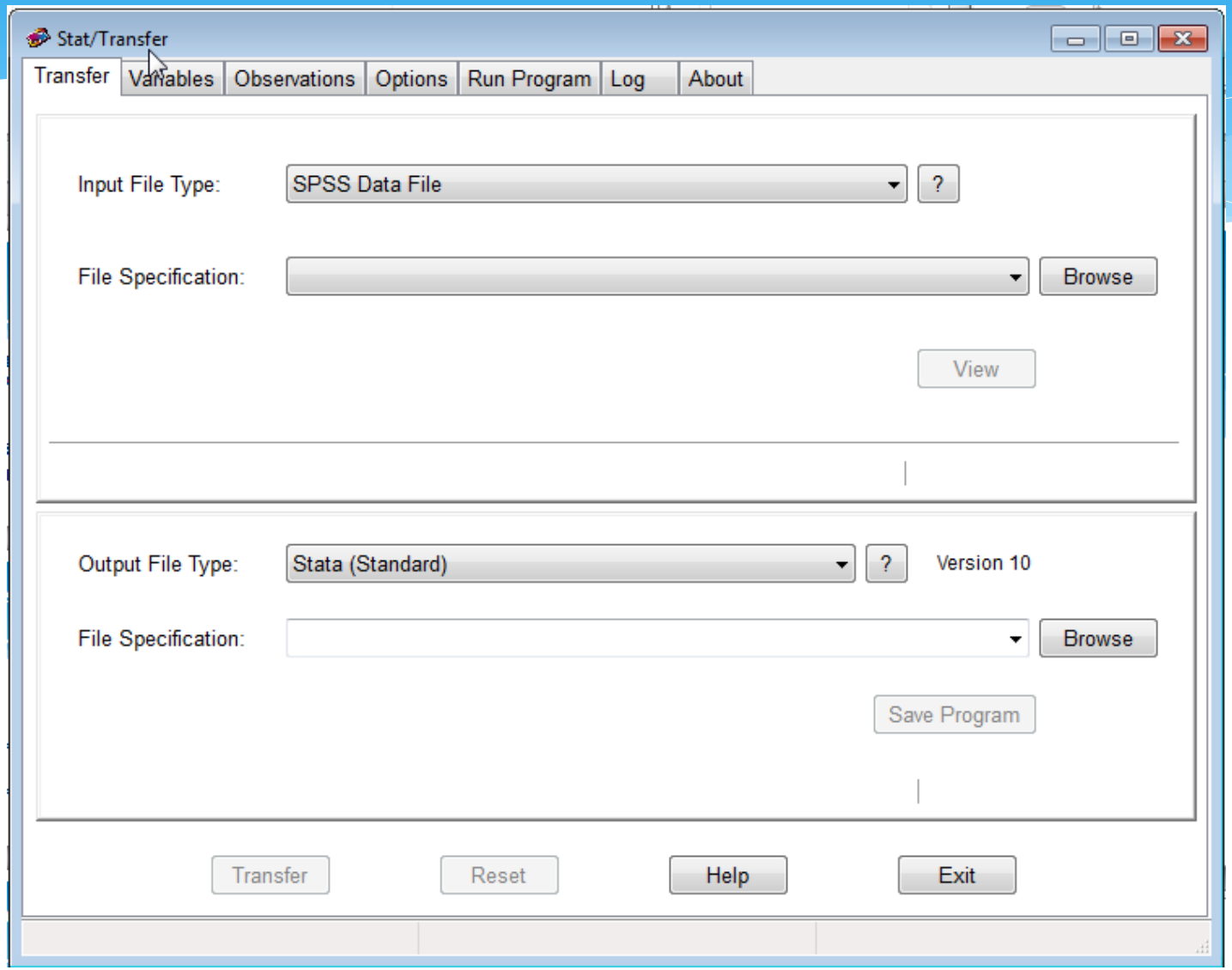
Specific Recommendations

* Database Software
* Variable Names/Value labels
* Data Documentation
* Datafile Version Control
* File Formats/Transmission of Data Files

Database Software

* 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

Stat/Transfer



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)

Pitfalls with Variable Names

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.

Variable Labels

* 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

Encoding/Value Labels

* 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

* Study Protocol/Data Operations Manual
* Codebook/Data Dictionary (ideally electronic and string searchable)
* Sample CRF (binder with data collection forms)
* Unresolved Queries/Issues
* Invalid Values
* Version Control

Codebooks/Data Dictionaries

* Range from v. elaborate to v. simple
* Variable Name
* Variable Description
* Variable Format (for dates, be careful and explicit as to 12/10/1975 vs 10/12/1975)
* Encoding (if any)
* Ranges, acceptable values
* Counts, Descriptives
* Value Labels
* Missing Data codes

CACHE COUNTY MEMORY STUDY

Johns Hopkins University
Utah State University
Duke University

SECOND INCIDENCE WAVE

RISK FACTOR - V3,	CLINICAL ASSESSMENT- C3, CP3	LABS AND MRI'S L3	MD VISITS M3, MP3	INCIDENCE FOLLOWUP CA F3, FP3
Cognitive Screening Interim Risk Factor Interview	Clinical Assessment Case Staffing ADL/Social Support	Blood Work, MRI Urinalysis Physician Review	Physician Visits to Subjects	Clinical Assessment Case Staffing

THIRD INCIDENCE WAVE

You are here

RISK FACTOR – V4,	CLINICAL ASSESSMENT- C1, CP4	LABS AND MRI'S L4	MD VISITS M4, MP4	INCIDENCE FOLLOWUP CA C1, FP4
Cognitive Screening Interim Risk Factor Interview	Clinical Assessment Case Staffing ADL/Social Support	Blood Work, MRI Urinalysis Physician Review	Physician Visits to Subjects	Clinical Assessment Case Staffing

Over 100 PDF files corresponding to each separate datafile

Unique Subject Identification SPSS Dataset C1_CASESTAFF

SPSS Variable/ Label	Description	Coding
Id Unique Subject Identification	Unique Subject Identification	N=902

Common Variables SPSS Dataset C1_CASESTAFF

SPSS Variable/ Label	Description	Coding
gender Gender	Gender	1 Male 2 Female 7 Refused 8 Don't Know 9 Missing
dob Date of Birth	Date of Birth	Date

Study also collected data on participants' spouses and caregivers

SPSS Variable Label	Description	Coding
c1csdt C1:CS: Case Staffing Date	Item: DATE Case Staffing: DATE OF SECOND INCIDENCE CLNICAL ASSESSMENT CASE STAFFING	Date
c1cc01 C1:CS: Memory Impairment for new events	Item: 1 Case Staffing: Caseness Checklist: EVIDENCE OF MEMORY IMPAIRMENT FOR NEW EVENTS.	0 No 1 Yes 9 Missing
c1cc02chk C1:CS: At Least one of the following is impaired	Item: 2 Case Staffing: Caseness Checklist: AT LEAST ONE OF THE FOLLOWING IS IMPAIRED	0 No 1 Yes 9 Missing
c1cc02 C1:CS: Abstract Thinking	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
```

Keyed: ( )

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 (1) No (2)

12. ↵

11. Approximately how long ago did the testing occur: ____ ____ days

	A	B	C	D	E	F	G	H
1	Variable / Field Name	Form Name	Field Units	Section H	Field Type	Field Label	Choices OR Calculations	T
2	id	inclusion_exclusion_criteria			text	Participant ID Number		
3	datenpi	inclusion_exclusion_criteria			text	Date of Visit		
4	incl_1	inclusion_exclusion_cri		Inclusion	radio	Diagnosis of AD by NINCDS/ADRDA criteria (47)	1, No | 2, Yes	
5	incl_2	inclusion_exclusion_criteria			radio	Age >60. This excludes early-onset AD cases whic	1, No | 2, Yes	
6	incl_3	inclusion_exclusion_criteria			radio	Mini-Mental State Exam (MMSE) 16-26. This range	1, No | 2, Yes	
7	incl_4	inclusion_exclusion_criteria			radio	Clinical Dementia Rating (CDR) < = 1 (mild demer	1, No | 2, Yes	
8	incl_5	inclusion_exclusion_criteria			radio	Patients will be allowed to remain on current FD/	1, No | 2, Yes	
9	incl_6	inclusion_exclusion_criteria			radio	Patients will be allowed to remain on antidepres	1, No | 2, Yes	
10	incl_7	inclusion_exclusion_criteria			radio	Knowledgeable informant available for all study	1, No | 2, Yes	
11	excl_1	inclusion_exclusion_cri		Exclusion	radio	Evidence of non-AD dementias including Hunting	1, No | 2, Yes	
12	excl_2	inclusion_exclusion_criteria			radio	Current DSM-IV Axis I diagnoses other than deme	1, No | 2, Yes	
13	excl_3	inclusion_exclusion_criteria			radio	Any clinically significant medical condition that	1, No | 2, Yes	
14	excl_4	inclusion_exclusion_criteria			radio	Current use of Beta-blocking agents	1, No | 2, Yes	
15	excl_5	inclusion_exclusion_criteria			radio	Contradindications to use of Beta-blocking agent	1, No | 2, Yes	
16	excl_6	inclusion_exclusion_criteria			radio	Clinically significant hepatic or renal insufficienc	1, No | 2, Yes	
17	examiner	inclusion_exclusion_criteria			text	Examiner ID		
18	idmm	mini_mental_sate_examination_			text	Participant ID Number		
19	mmse_date	mini_mental_sate_examination_			text	Date		
20	mmse_01	mini_mental_sate_examination_			text	Time orientation (0-5):		
21	mmse_02	mini_mental_sate_examination			text	Place orientation (0-5):		

Considerations for Longitudinal Datasets

Wide: 1 line per patient

ptid	weightlb1	weightlb2	weightlb3
1	150	145	140
2	160	165	163
3	170	172	167
4	180	189	195

Long: 1 line per visit

ptid	visit	weightlb
1	1	150
1	2	145
1	3	140
2	1	160
2	2	165
2	3	163
3	1	170
3	2	172
3	3	167
4	1	180
4	2	189
4	3	195

Visit indicator needs to be at the end of the var name stub.

Dataset Cleaning

* 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

Calculated Variables/Data Programming

* 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.

Dataset Version Control

* 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

Dataset Distribution

* 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