Data Management

Walter Leite, PhD
Assistant Professor
Research Evaluation & Methodology

Survey Data Analysis

Data Management Issues

- Database construction
- Data dictionary (code book)
- Weighting
- Imputation

Database construction

Depends on:

- Data types
- Amount of data
- Budget (expertise)
- Sources of data

Types of Data (medical)

- Participation data (includes enrollment, Inclusion/Exclusion criteria, loss to follow-up)
- Demographics data
- Medical History
- Coexisting diseases
- Medications
- Adverse Experiences
- Lab measurements
- Vital Signs
- Physical Exams, Neuropsych testing
- Specific survey data

Types of Data (education)

- Participation data (includes enrollment, Inclusion/Exclusion criteria, loss to follow-up)
- Demographics data
- Medical/neuropsych data
- School history
  - Achievement
  - Attendance
  - Disciplinary
  - Special Education
- Specific survey data

Options

Data Entry Mechanisms	Databases
Paper Scannable forms Web based Phone (CATI) Phone (voice)	Excel spreadsheet ACCESS database SQL database Oracle

Date

/ /

MRN

0

Modified UPDRS Unified Parkinson Disease Rating Scale

Visit Type:
- ☐ General Followup
- ☐ Baseline
- ☐ 4 month
- ☐ 6 month
- ☐ 12 month
- ☐ 24 month
- ☐ 36 month
- ☐ 48 month

Test Number:
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4

Home Treatment States

Dopa/Agonist Meds	☐ ON	☐ OFF
DBS Left	☐ ON	☐ OFF
DBS Right	☐ ON	☐ OFF

A. Mentation Behavior and Mood

1. Intellectual Impairment:

0 - None
1 - Mild. Consistent forgetfulness with partial recollection of events and no other difficulties
2 - Moderate memory loss, with disorientation and moderate difficulty handling complex problems. Mild but definite impairment of functions at home with need of occasional prompting
3 - Severe memory loss with disorientation for time and often to place, severe impairment in handling problems
4 - Severe memory loss with disorientation preserved to person only, unable to make judgements or solve problems, requires much help with personal care, can not be left alone at all.

☐

2. Thought Disorders: (due to dementia or drug intoxication)

0 - None
1 - Vivid dreaming
2 - "Benign" hallucinations with insight retained
3 - Occasional to frequent hallucinations of delusions without insight, could interfere with daily activities
4 - Persistent hallucinations of delusions or florid psychosis, not able to care for self

☐

3. Depression:

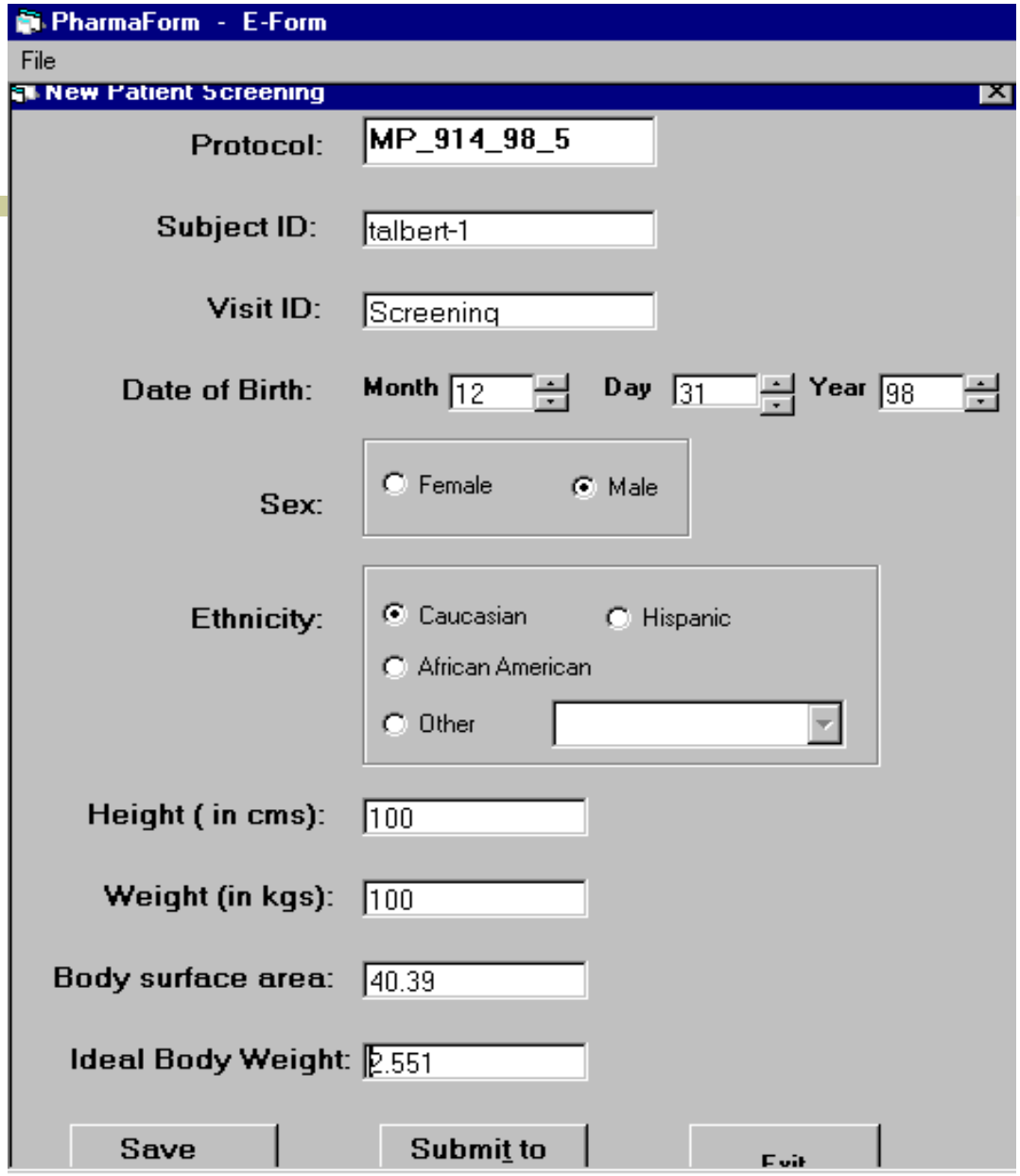
0 - None
1 - Periods of sadness or guilt greater than normal, never sustained for days or weeks
2 - Sustained depression (1 week or more)
3 - sustained depression with vegetative symptoms (insomnia, anorexia, weight loss, loss of interest)
4 - Sustained depression with vegetative symptoms and suicidal thoughts or intents

☐

4. Motivation / Initiative:

0 - Normal
1 - Less assertive than usual; more passive
2 - Loss of initiative or disinterest in elective (non-routine) activities
3 - Loss of initiative or disinterest in day-to-day (routine) activities
4 - Withdrawn, complete loss of motivation

☐


PharmaForm - E-Form
File
New Patient Screening
Protocol:
MP_914_98_5
Subject ID:
talbert-1
Visit ID:
Screening
Date of Birth:
Month
12
Day
31
Year
98
Sex:
Female
Male
Ethnicity:
Caucasian
Hispanic
African American
Other
Height (in cms):
100
Weight (in kgs):
100
Body surface area:
40.39
Ideal Body Weight:
2.551
Save
Submit to
Exit

Side	Distance (cm)	Subject ID	Onset Latency (ms)	Peak Latency (ms)	Velocity (m/s)	Age	Sex	Height (inches)	Weight (lbs)
R	3	5	2.9	3.8	36.8	40	M	70.5	197
R	4	5	2.9	3.8	36.8	40	M	70.5	197
R	3	6	2.8	3.5	40.0	29	M	69	165
R	4	6	2.7	3.4	41.2	29	M	69	165
L	3	6	2.6	3.2	43.8	29	M	69	165
L	4	6	2.7	3.2	43.8	29	M	69	165
R	3	4	2.9	3.6	38.9	33	M	68.5	155
R	4	4	3	3.7	37.8	33	M	68.5	155
L	3	4	2.9	3.5	40.0	33	M	68.5	155
L	4	4	3	3.7	37.8	33	M	68.5	155
R	3	7	3.2	3.9	35.9	33	F	68	145
R	4	7	3.2	4	35.0	33	F	68	145
L	3	7	2.9	3.6	38.9	33	F	68	145
L	4	7	2.9	3.7	37.8	33	F	68	145
R	3	9	2.9	3.5	40.0		M	70	170
R	4	9	3.1	3.8	36.8		M	70	170
L	3	9	2.6	3.4	41.2		M	70	170
L	4	9	2.7	3.4	41.2		M	70	170
R	3	11	2.9	3.6	38.9	32	M	66	130
R	4	11	3	3.7	37.8	32	M	66	130
L	3	11	3	3.7	37.8	32	M	66	130
L	4	11	2.8	3.8	36.8	32	M	66	130
R	3	10	3	3.8	36.8	31	M	70	170
R	4	10	3	3.7	37.8	31	M	70	170
L	3	10	3	3.8	36.8	31	M	70	170
L	4	10	2.9	3.6	38.9	31	M	70	170
R	3	12	2.8	3.4	41.2	29	F	63	130
R	4	12	2.9	3.5	40.0	29	F	63	130
L	3	12	2.5	3.1	45.2	29	F	63	130
L	4	12	2.6	3.3	42.4	29	F	63	130
R	3	13	2.5	3.2	43.8	33	M	70	160
R	4	13	2.5	3.2	43.8	33	M	70	160
L	3	13	2.6	3.2	43.8	33	M	70	160
L	4	13	2.7	3.3	42.4	33	M	70	160
R	3	14	3.2	3.9	35.9	29	F	63	108
R	4	14	3.1	3.9	35.9	29	F	63	108
L	3	14	3.2	4	35.0	29	F	63	108
L	4	14	3.2	3.9	35.9	29	F	63	108
R	3	15	2.5	3.2	43.8	23	F	60	100
R	4	15	2.5	3.3	42.4	23	F	60	100
L	3	15	2.7	3.4	41.2	23	F	60	100
L	4	15	2.6	3.3	42.4	23	F	60	100
R	3	16	2.6	3.1	45.2	30	M	72	180
R	4	16	2.7	3.2	43.8	30	M	72	180

Fundamental Questions

- How will data be secured?
- How will data be checked?
- How will data be retrieved?
- How will data be documented?

Data Dictionary or Codebook

Name	Type	Values	Description
age	Num		Age at registration
ptrdob	Char	(dd/mm/yyyy)	Date of birth
agegroup	Char	1 = 18 - < 40 2 = 40 < 60 3 = 60 plus	Age group
race	Char	B = black O = other W = white	Race of subject
ptrgdr	Char	M = male F = female	Gender
RopEduc	Char	ED01=less than high school ED02 = high school or GED ED03=Advanced education attended	Education classification
RopEdy	Num		Years of education
RopMar	Char	MA01= single (Not married, or widowed) MA02= married (Married or living together as married)	Marital Status

After Data Collection

- After data are collected there is a processing step:
- Measurements are scored (with checking)
- Scores are normed (with checking)
- Data are checked for logical values
- Data are checked for missingness
- Data are checked for outliers

Weighting

Reasons for weighting

1. Adjust for complex survey design
2. Adjust for non-response
3. Make adjustment to known population data

Example of weighting

Gender	Florida population*	Unweighted sample	Survey weight	Weighted sample
Male	49.1%	30%	(.491/.3)	49.1%
Female	50.9%	70%	(.509/.7)	50.9%

* Taken from 2000 census, http://www.census.gov/

What are sampling weights?

- Sampling weights are the number of individuals in the population each respondent in the sample is representing.
- A sample weight is the inverse of the probability of selection.
- For example, if my simple random sample is one tenth of the population size (i.e. my sampling fraction is 1/10), then each respondent in the sample is representing 10 people in the population.

Weights compensate for:

- Unequal probability of selection
- Unequal response rates
- Post-stratification (adjust the sample distribution for key variables of interest such as age, ethnicity, sex, to make it conform to a known population distribution)

How do weights work?

Score	Weight
4	1
2	2
1	4
5	1
2	2

Simple mean:

$$\frac{(4+2+1+5+2)}{5} = 2.8$$

Weighted mean:

$$\frac{(4x1)+(2x2)+(1x4)+(5x1)+(2x2)}{10} = 2.1$$

Weights are frequencies of each observation in the population.

Types of Weights

- Raw weights
- Relative or normalized weights
- Design effect adjusted weights
- Non-response adjusted weights

Raw weights or base weights

- Raw weights sum to the population size. They are the inverse of the probability of selection:

$$w_i = \frac{1}{p_i}$$

- For example, if the probability of selection of a unit is 1/50, its raw weight is 50.

Problems of using raw weights

- Estimates of means, proportions and standard errors obtained using raw weights will be based on the population size, not the sample size. The means and proportion estimates will be correct, but the test statistics will have too much power.

- Solution: Convert raw weights to normalized weights.

Normalized or relative weights

- Normalized weights sum to the sample size.
- With normalized weights in the analyses, the estimates of means, and proportions are correct. The estimates of standard errors are correct given a simple random sample or stratified sample.
- When a cluster or multi-stage sample is used, the estimation of standard errors will not be correct using only case weights. Special procedures such as Taylor-series approximation, bootstrapping or design effects need to be used to obtain correct standard errors.

Converting a raw weight to a normalized weight

- There are two ways of converting a raw weight to a normalized way:

1. Dividing the raw weights by the mean of the raw weights:

$$w_N = \frac{w_i}{\bar{w}}$$

2. Multiplying the raw weight by the overall sampling fraction:

$$w_N = w_i \frac{n}{N}$$

Standard Errors

- When a using a complex survey design, the estimation of standard errors will not be correct.
- Special procedures such as Taylor-series approximation, bootstrapping or design effects need to be used to obtain correct standard errors.

Imputation

- Mean value
- Regression
- Hot deck
- Multiple imputation

Planning

- What are your needs?
- What are your resources?
- Do you have the right expertises?
- Do you have a plan for training?
- Do you have a plan for pilot testing?
- Do you have a plan for imputation?
- How can you facilitate good communication?