Biostatistics, Epidemiology & Research Design (BERD)

Howard Cabral, PhD, MPH
Christine Chaisson, MPH

Seminar Series: CTSI Presents…

December 15, 2016

BOSTON UNIVERSITY

CTSI Presents…

Data Management: What you should know and why you should care

Christine Chaisson
Director, Data Coordinating Center
Assistant Research Professor, Biostatistics
Boston University School of Public Health

BU

Boston University School of Public Health
Data Coordinating Center

Things that can go wrong with data

- Crucial data elements may be missing
- Data may be incorrect due to errors in:
  - Data collection
  - Data entry
- Data may be not have common identifier
  - Cannot be merged
  - May be merged incorrectly
- Data may not be saved or backed up
- Data files may be lost or corrupted

BOSTON UNIVERSITY

Real World Examples

- A few illustrations of common data problems from the popular news sources

BOSTON UNIVERSITY

The Inquirer DAILY NEWS

philly.com

August 3, 2012

Forbes: Bad data hurt Haverford in college rankings

"Forbes' annual list is out, and Haverford plummeted from No. 7 to No. 27 - for no obvious reason. A College spokesman explained that the error was based on single figure:
A zero was incorrectly entered in database instead of 108 for the graduation rate of white women who enrolled in 2004.
…But no revision is planned, since the magazine and the online list has already been published."

BOSTON UNIVERSITY

Data Entry Error

PharmaTimes ONLINE

May 6, 2012

Vertex stock slides over cystic fibrosis data mistake

"Shares in Vertex Pharmaceuticals have taken a hit after the company had to take the rather embarrassing step of correcting previously-announced interim mid-stage results of a combination cystic fibrosis treatment.
…the result of a misinterpretation [of the denominator of the treatment group] between the firm and its outside statistical ve

BOSTON UNIVERSITY

Data Mismanaged

Oops: Excel Error Calls Into Question…

IEEE SPECTRUM

Posted 22 Apr 2013

- '*This Time It's Different*' *a* 2009 book by Harvard researchers (Reinhard & Rogoff) contained "Serious errors that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies in the post-war period." Investigators at UMass (among others) were unable to replicate work which led to accusations of intentional fudging of the data or
- The Authors admitted they failed to include five rows of data from an Excel file (Australia, Austria, Belgium, Canada, and Denmark) —a "coding error" which they said was "a significant lapse on our part."

excluded key data

BOSTON UNIVERSITY

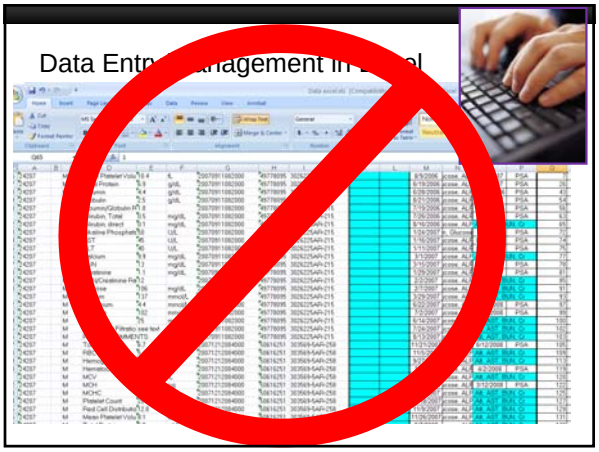
The New York Times July 7, 2011

How Bright Promise in Cancer Testing Fell Apart

- Duke Cancer Center's gene-based tests proved worthless, research behind them was discredited
- Statisticians from MD Anderson discovered errors such as columns moved over in a spread-sheet; Duke team "shrugged them off" as "clerical errors."
- Four papers were retracted
- Duke shut down three cancer trials
- Center leaders resigned or were removed
- People died and their relatives sued Duke

BOSTON UNIVERSITY

Data Entry/Management in Excel



Goal: Convert Data into Electronic Format as Quickly as Possible

BOSTON UNIVERSITY

How should data be managed?

- No single "right" way to collect or manage data
- Consider:
  - Environment/location
  - Resources ($)
  - Regulations
- Be sure to ***plan*** prior to study start
- Do what works for the study at hand

BOSTON UNIVERSITY

Where to start?

BOSTON UNIVERSITY

Data management plan

From Wikipedia, the free encyclopedia

A **data management plan** or **DMP** is a formal document that outlines how you will handle your data both during your research, and after the project is completed.[1] The goal of a data management plan is to consider the many aspects of data management, metadata generation, data preservation, and analysis before the project begins; this ensures that data are well-managed in the present, and prepared for preservation in the future.

DMP Purpose: To help you collect, manage and share your data; meet funder requirements. General elements include:

- Project or study description
- Documentation, organization, storage
- Access and sharing
- Archiving

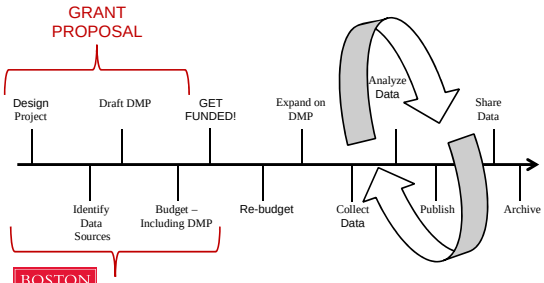
BOSTON UNIVERSITY

DMP: Basic Elements

- Study design, data types and sources
- Storage format and location
- Naming conventions, documentation
- Software used for manipulation
- Project Staff – who has permission to what
- Identifiers (if applicable)
- Back ups, security
- Archiving
- Sharing

BOSTON UNIVERSITY

Timeline



BOSTON UNIVERSITY

Beginning: Identify Key Data Elements

- Review hypotheses
- What are primary, secondary outcomes?
- What covariates and confounders must be collected?
- What are the data sources?
  - Questionnaires
  - Labs, imaging
  - Medical record review
  - other external sources (e.g., lab results, medical records, death certificates)

BOSTON UNIVERSITY

Other Data Elements

- Regulatory data:
  - IRB requirements
  - Safety (DSMB)
  - FDA (e.g., 21 CFR, part 11)
  - VA
  - Other?
- Tracking/Study management data:
  - Tracking participants
  - Tracking data elements by time-points
- Harmonization and sharing
  - NIH
  - Other

BOSTON UNIVERSITY

Visit Protocol: Data by Time-point

- Determine visit Schedule and data collected at each visit
  - Questionnaires
  - Labs
  - Other?
- Consider data not be connected to visits
  - Adverse events, serious adverse events
  - Hospitalization
  - Death
  - Medical records

BOSTON UNIVERSITY

Sample data/visit grid

	Screening Period	Baseline	Week 4	Week 8	Week 12	Week 16	Week 20	Week 24	Follow-up visit week 28
	1-2 visits between days -28 to 1	Day 1	Day 28	Day 56	Day 84	Day 112	Day 140	Day 168	One month off
Medical history & demographics/ mean Hgb, transfusion frequency prior or	X								
Physical examination (PE) (1)	X								
Pregnancy Test (Females) (9)	X	X	X	X	X	X	X	X	X
Inclusion/ exclusion criteria evaluation	X								
Informed consent / assent	X								
Randomization		X							
Targeted physical exam (8)		X	X	X	X	X	X	X	X
Vital signs with pulse oximetry (O_2 sat) (9)	X	X	X	X	X	X	X	X	X
CBC and Differential (9)	X	X	X	X	X	X	X	X	X
Reticulocyte count (9)	X	X	X	X	X	X	X	X	X
Chemistry panel with LDH, Bili (9)	X	X	X	X	X	X	X	X	
Hemolysis assay/LDH, haptoglobin (6)		X		X	X		X	X	X
NT-proBNP (4)		X			X			X	X
EPO, Ferritin (3)		X			X			X	
Hemoglobin F (HbF) (3)		X			X			X	
Gardos channel studies / Central lab (3)		X			X			X	
6 Minute walk distance (2)		X						X	
Dyspnea scale, NY AHA score (5)		X			X			X	
QOL (FACIT fatigue scale, ASCQME) (2)		X						X	
Adverse events, Pain Scale (8)		X	X	X	X	X	X	X	X
Concomitant meds (7)		X	X	X	X	X	X	X	
ECHO (2)		X						X	
Telephone call (weekly) (14)		X	X	X	X	X	X	X	
Pharmacy Compliance & Med reconciliation (7)		X	X	X	X	X	X	X	
Red cell survival (subset) (3-5)		X						X	
Interval History (crisis, hospitalization, or transfusions?) (7)			X	X	X	X	X	X	X

Timelines and Tasks

- Develop Protocol and Analytic Plan
- Create and pilot of forms/assessments
- Design/construct data systems
  - Data Collection/entry
  - Participant/Data Tracking
- Subject recruitment
- Data collection (baseline and follow-up)
- Data cleaning, auditing, and QA
- Preliminary analysis
- Manuscript preparation & submission

BOSTON UNIVERSITY

Create a Visual *Timeline*

- It doesn't have to be fancy
- More detail is better but something simple is better than nothing
- Plan to review and revise it often

BOSTON UNIVERSITY

Simple overview Timeline

Study Timeline

Selected Activities	1-6	6-12	12-18	18-24	24-30	30-36	36-42	42-48	48-54	54-60
Hiring & training										
Finalize instruments & IRB										
Enrollment		X	X	X	X					
Intervention		X	X	X	X	X	X	X		
Follow-up			X	X	X	X	X	X		
Data QA/clean			X	X	X	X	X	X	X	
Primary & secondary data analyses									X	X
Presentations & Publication									X	X
Study meetings	X	X	X	X	X	X	X	X	X	X

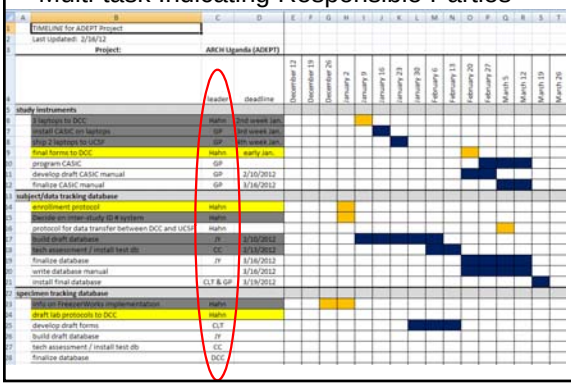
Design/construct data collection systems

Sample Task-based Gantt

Tasks	Year 1				Year 2			
	Months 1-3	Months 4-6	Months 7-9	Months 10-12	Months 13-15	Months 15-18	Months 29-21	Months 22-24
Finalize CRFs								
IRB Approval								
Finalize data platforms								
Finalize protocol								
Build eCRF								
Build database								
Pilot CRFs/protocol								
Build website								
Enrollment/data collection								
Query data /monitor								
Automate data reports								
Update website, reports								
DSMB data freeze, reports, meeting								
Follow up visits								

BOSTON UNIVERSITY

Multi-task Indicating Responsible Parties



TIMELINE for ADEPT Project

Last Updated: 2/16/12

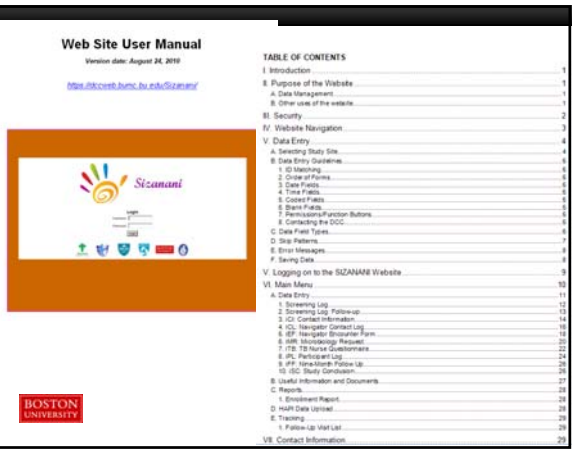
Project: ARCH Uganda (ADEPT)

	leader	deadline
study instruments		
3 laptops to DCC	Hahn	2nd week jan.
install CASIC on laptops	GP	3rd week jan.
ship 2 laptops to UCSF	GP	4th week jan.
final forms to DCC	Hahn	early jan.
program CASIC	GP	
develop draft CASIC manual	GP	2/10/2012
finalize CASIC manual	GP	3/16/2012
subject/data tracking database		
enrollment protocol	Hahn	
Decide on inter-study ID # system	Hahn	
protocol for data transfer between DCC and UCSF	Hahn	
build draft database	JY	3/16/2012
tech assessment / install test db	CC	3/16/2012
finalize database	JY	3/16/2012
write database manual		3/16/2012
install final database	CLT & GP	3/16/2012
specimen tracking database		
info on FreezerWorks implementation	Hahn	
draft lab protocols to DCC	Hahn	
develop draft forms	CLT	
build draft database	JY	
tech assessment / install test db	CC	
finalize database	DCC	

Tools Of The Trade

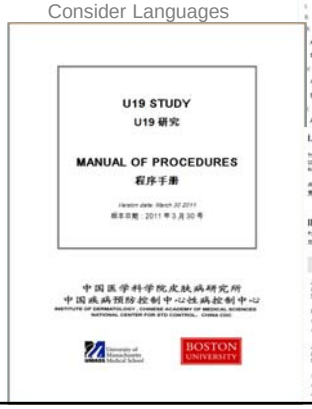
- Analytic plan
- Detailed protocol
- Well designed data collection forms
- Tracking system
- Data capture/entry system
- Plan for data query (checking/cleaning)
- Manuals
- Data dictionaries

BOSTON UNIVERSITY



BOSTON UNIVERSITY

Consider Languages



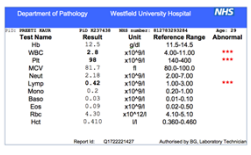
Sample Data Dictionary: SAS formatted dataset

Number	Variable	Label	Type	Format
1	adept_study_id	ADEPT Study ID	Num	8.
2	MissedAssessment	MissedAssessment	Num	11.
3	Interviewer	Interviewer	Char	$50.
4	PatientLanguage	PatientLanguage	Char	$20.
5	Completed	Completed	Num	11.
6	DtAssessStart	DtAssessStart	Num	DATETIME22.3
7	DtAssessEnd	DtAssessEnd	Num	DATETIME22.3
8	datenow	What is today's date?	Num	DATETIME22.3
9	serv_initials	Please enter your (the research assistant) initials	Char	$1024.
10	lang	Language of interview	Num	LANGF
11	sex	Gender	Num	SEXF.
12	weight	Participant's weight in Kg	Num	20.2
13	height	Participant's height in cm	Num	20.1
14	age	Age	Num	8.
15	DOBDay	DoB: Day	Num	8.
16	DOBMonth	DoB: Month	Num	8.
17	DOBYear	DoB: Year	Num	8.
18	tribenkole	Tribe: Nkole	Num	NOYES
19	tribekiga	Tribe: Kiga	Num	NOYES
20	tribehima	Tribe: Hima	Num	NOYES
21	tribeganda	Tribe: Ganda	Num	NOYES
22	tribetoro	Tribe: Toro	Num	NOYES
23	tribekenyo	Tribe: Kenyo	Num	NOYES
24	tribeothe	Tribe: Other	Num	NOYES
25	tribedk	Tribe: Don't Know	Num	NOYES
26	tribena	Tribe: Not applicable	Num	NOYES

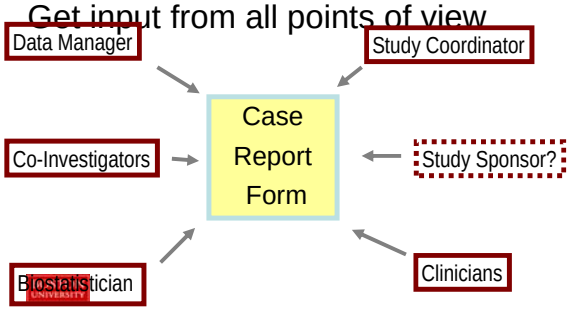
BOSTON UNIVERSITY

Data Collection Forms

- Data will usually end up on some type of "form" whether it is interview, chart review, or imaging results
- Make sure you plan carefully and leave time as this can be a lengthy process



Creating a Data Collection Form: Get input from all points of view



Real-world example why you should get input from multiple perspectives

What is your relationship to [person with illness] (i.e., [person with illness] is your _____)?

- ❑ Parent
- ❑ Child
- ❑ Sibling
- ❑ Spouse/Partner
- ❑ Other _____

BOSTON UNIVERSITY

Why is this topic important?

- ❑ Sloppy forms indicate sloppy research
- ❑ CRF may not answer study questions
- ❑ Danger of collecting:
  - ❑ too much data
  - ❑ too little data
  - ❑ the wrong data
- ❑ Annoyed:
  - ❑ Participants
  - ❑ Study Coordinator
  - ❑ Data Analyst…

BOSTON UNIVERSITY

When designing forms…

Think Google

Not Yahoo

BOSTON UNIVERSITY

What makes a good case report form?

- ❑ User-friendly, uncluttered, well organized
- ❑ Provides clear instructions for completion
- ❑ Terminology familiar to person filling out
- ❑ Reading level matches study participants/evaluators
- ❑ Unambiguous questions
- ❑ Questions only asked/data collected in one place and *only* one place
- ❑ Easy to refer back and clean data

BOSTON UNIVERSITY

Account For Missing Data

CBC	Unit	Value	
1. Hemoglobin	g/dl	__ __.__	☐ Not Done
2. Hematocrit	%	__ __.__	☐ Not Done
3. RBC	M/mm^3	__ __.__	☐ Not Done

BOSTON UNIVERSITY

Specify the Units

Categorize Anticipated Responses

- ☐ 1 USA
- ☐ 2 Guatemala
- ☐ 3 Mexico
- ☐ 4 Dominican Republic
- ☐ 5 Other

- ☐ 1 USA
- ☐ 2 Guatemala
- ☐ 3 Mexico
- ☐ 4 Dominican Republic
- ☐ 6 El Salvador
- ☐ 5 Other

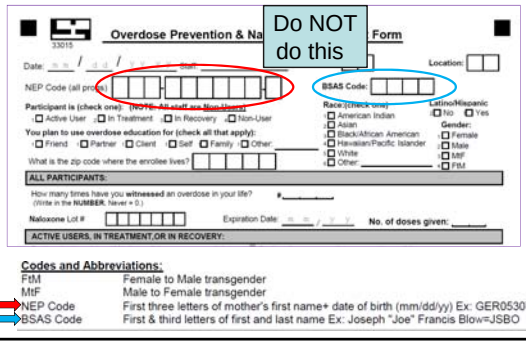
BOSTON UNIVERSITY

ID Assignment

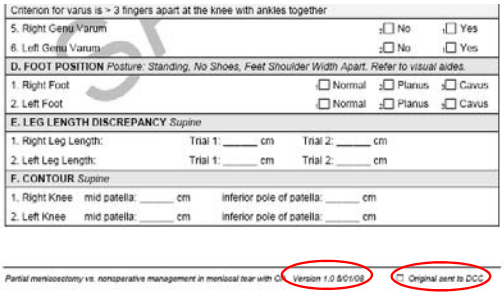
- Must be UNIQUE for each subject
- Should appear on every form (preferably page)
  - Primary key for records database
  - "Merge key" to join various data elements
  - Serves as identifier (no names!)
- May be a simple number 1001
- May be multi-part: 102101
  - 1 = Site
  - 02 = Language
  - 101= ID

BOSTON UNIVERSITY

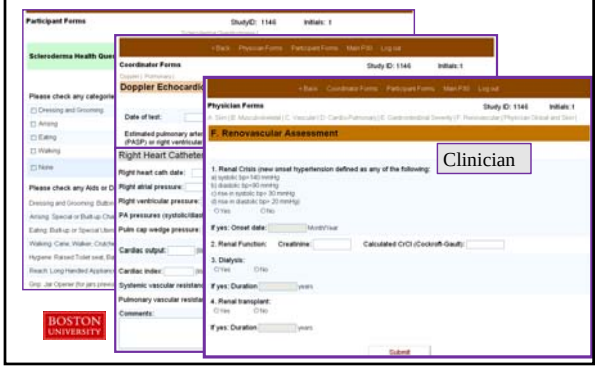
Example of an ID that is not unique



Don't Underestimate Need for Version# /Date



Consider who is Completing a Form



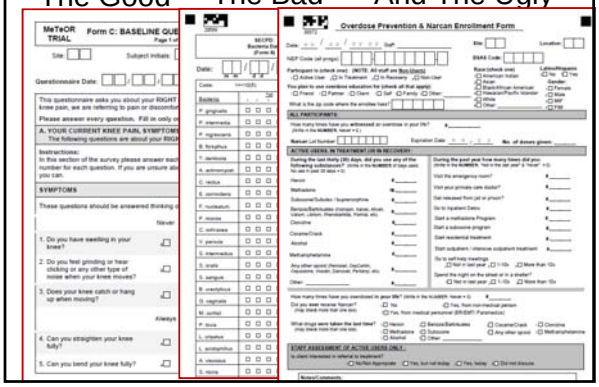
BOSTON UNIVERSITY

In Summary, when designing questions:

- Avoid ambiguous questions and open ended responses
- Include clear instructions
- Be sure form complexity matches collector (self, study coordinator, clinician)
- Collect data elements in correct format ("continuous" or "categorical")
- Make categories mutually exclusive
- Pilot your forms in the target population

BOSTON UNIVERSITY

The Good The Bad And The Ugly



Paper or Electronic?

BOSTON UNIVERSITY

Paper Forms / Manual Entry

Advantages
- The old "standard"
- Shorter start-up time (Word/PDF)
- Relatively easy to train staff
- Hardcopy document to refer back to
- Can be done anywhere

Disadvantages
- Costs: data entry, storage and shipping
- Longer time from collection to database
- Errors in data collection (missing, out of range, skips)

BOSTON UNIVERSITY

Electronic Data Capture

Advantages
- Cleaner data at entry (required fields, skips, ranges)
- Can use data in real time (or close to it)
- No extra data entry costs
- Data can inform next visit even for short follow up

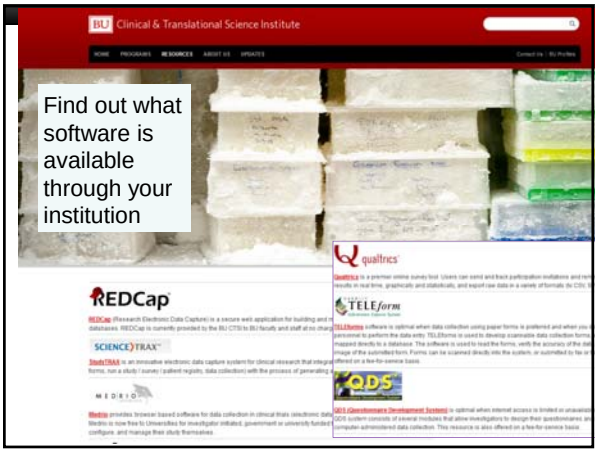
Disadvantages
- Programming time and costs
- Increased hardware and software costs
- Infrastructure concerns (software versions, internet connection, back-up equipment)
- Data security

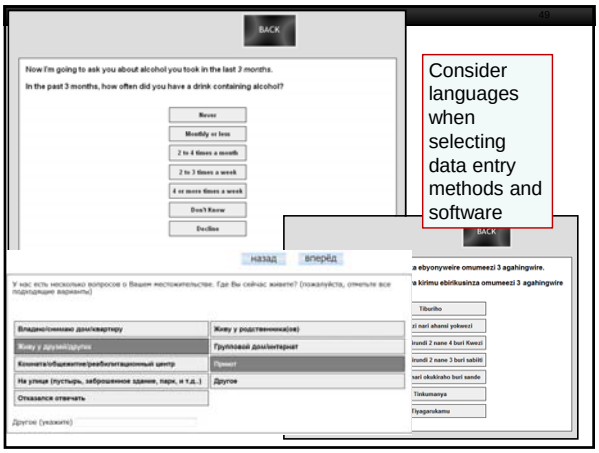
BOSTON UNIVERSITY

A Word About "Canned" Software

- Many "canned" sofware packages available
- No single best choice
- Cost can vary widely
- Database structures vary
- Do your homework to make sure wha for your project

BOSTON UNIVERSITY





Consider languages when selecting data entry methods and software

What to use…?

- To determine what software is best suited for your project see:
  - What is available to you?
  - What is the cost (can you afford it)?
  - What has the features you need (e.g., language)

BOSTON UNIVERSITY

50

Once data are collected…

- Get your data into a useful format
- No "right" format – use what works for you
  - SQL database
  - SAS datasets
  - SPSS
  - Excel (be careful!)

BOSTON UNIVERSITY

Keep Personal Identifiers Separate

- Do not store any identifier unless you have a good reason for it
- Do not store identifiers in same files with study data. Identifiers should be kept separate!
- Create "crosswalk" files of identifiers and store them someplace secure.

BOSTON UNIVERSITY

Personal Identifiers

First Name	Last Name	Study ID	Screen ID	Phone #
Joseph	Blow	1234	50001	555-131-1111

Identifiable Data

Crosswalk file

Subject ID	MRN	SSN
1234	64322	******

Study Data

ID	Visit	Var 1	Var 2	Var 3
1234	1	5	2	3

HIPAA Identifiers

1. Names
2. Addresses other than state, and first three digits of the zip code
3. All elements of date other than year, and all specific ages over 89 years
4. Telephone numbers
5. fax numbers
6. Email addresses
7. Social Security numbers
8. Medical Record numbers
9. Health plan beneficiary numbers

BOSTON UNIVERSITY

HIPAA Identifiers (cont)

10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers
13. Device identifiers and serial numbers
14. Web universal resource locators (URLs; web site addresses)
15. Internet protocol (IP) addresses
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

BOSTON UNIVERSITY

Participant Tracking

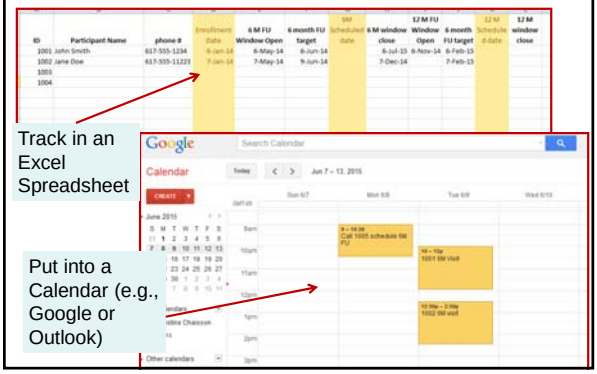
Tracking the Participants

You need a system to track participants

- Tracking for Study Management:
  - Screened, Eligible, Enrolled
  - Monitor and report progress
- Tracking tools for study staff:
  - Schedule/reminders follow up visits
  - Collection of all data points at each visit
- Small study may use Outlook or Excel; large study may need a tracking system

BOSTON UNIVERSITY

Simple Participant Tracking Tools



More Complex Tracking Tool

Tracking the Data Elements

- Identify what data have been collected
  - For each Subject at each Visit:
    - Questionnaires
    - Imaging, labs results
    - Other external data
- Missing data: can you still get it?
- Data cleaning / QA / auditing
- Create "clean" frozen datasets that you append new data to over time

BOSTON UNIVERSITY

Reports

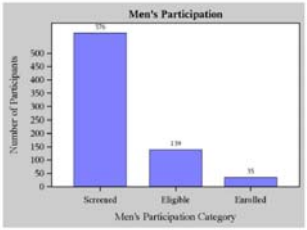
- For the study team to view at regular meetings or online as needed

BOSTON UNIVERSITY

Reports: Visual as well as Tabular

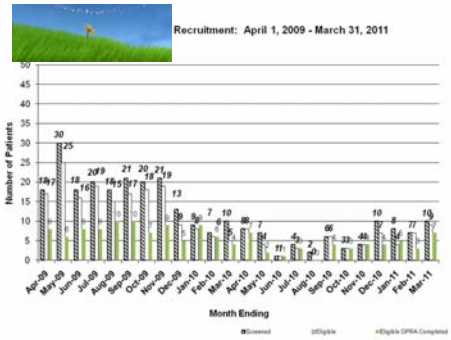
Men's Participation

gender	Status	N	%
MALE:	Screened	576	100
	Eligible	139	24
	Enrolled	35	6



BOSTON UNIVERSITY

Screened, Eligible and Enrolled



Enrollment

December 2, 2013

Screening and Enrollment Summary

	This Week	Total	HERMITAGE	Other Source
Completed Screening A	_	215	145	70
Eligible for Screening B	_	134 (62.3%)	66 (45.5%)	68 (97.1%)
Completed Screening B	_	98 (73.1%)	37 (56.1%)	61 (89.7%)
Missed Screening B		30 (23.4%)	28 (43.1%)	2 (3.2%)
Eligible for study	_	98 (100.0%)	37 (100.0%)	61 (100.0%)
Enrolled	_	98 (100.0%)	37 (100.0%)	61 (100.0%)

Reasons for Ineligibility on Screening A

	n
Ever ART use, Current ART use	70, 46
Out of catchment area	7
Can't verify HIV status	4
Can't provide two contacts	4
Can't verify ART naive status	4
Doesn't have HIV	1
No phone	1

Baseline Alcohol Use

	Total	Males	Females
Baseline TLFB Complete	98	67	31
Heavy Drinking - Past 7 Days	37 (38%)	22 (33%)	15 (48%)
Heavy Drinking - Past 30 Days*	45 (46%)	28 (42%)	17 (55%)
Moderate Drinking - Past 30 Days	20 (20%)	13 (19%)	7 (23%)
Abstinence - Past 30 Days	33 (34%)	26 (39%)	7 (23%)

Phlebotomy

December 2, 2013

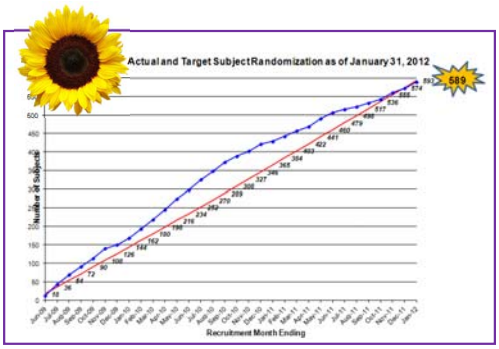
Phlebotomy and Lab Tracking Summary

	n (%)
Blood Drawn: Complete	95 (96%)
Blood Drawn: Incomplete	1 (1%)
Blood Not Drawn: 2 attempts	2 (2%)
Blood Not Drawn: 3 attempts	1 (1%)
Quality of Venipuncture: Clean	64 (66.7%)
Quality of Venipuncture: Traumatic	32 (33.3%)

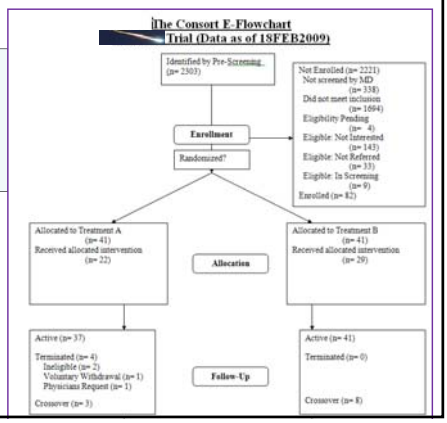
	Mean (Min, Med, Max)	Within 60 Minutes	Within 90 Minutes
Draw Time (minutes)	3.2 (1, 2, 18)		
Tourniquet Time (seconds)	72 (20, 68, 120)		
Time to Receipt (minutes)	42 (20, 39, 80)		
Time to Processing, Tube 1 (minutes)	44 (28, 40, 116)	85 (90.4%)	93 (98.9%)
Time to Processing, Tube 2 (minutes)	44 (28, 41, 116)	83 (90.2%)	91 (98.9%)
Time to Processing, Tube 3 (minutes)	58 (28, 58, 146)	49 (54.5%)	88 (97.7%)

Tube	Drawn	Full	Storage Consent	Aliquots Mean (Min, Med, Max)	Flow Cytometry Done	DBS Done
Tube 1 (Purple Top, EDTA)	94 (97.9%)	91 (94.8%)	83 (96.9%)	8.3 (5, 8, 10)	n/a	n/a
Tube 2 (Red Top, SST)	92 (95.8%)	89 (92.7%)	81 (94.8%)	7.6 (3, 8, 10)	n/a	n/a
Tube 3 (Green Top, Heparin)	92 (95.8%)	89 (92.7%)	81 (94.8%)	n/a	83 (96.9%)	84 (100%)

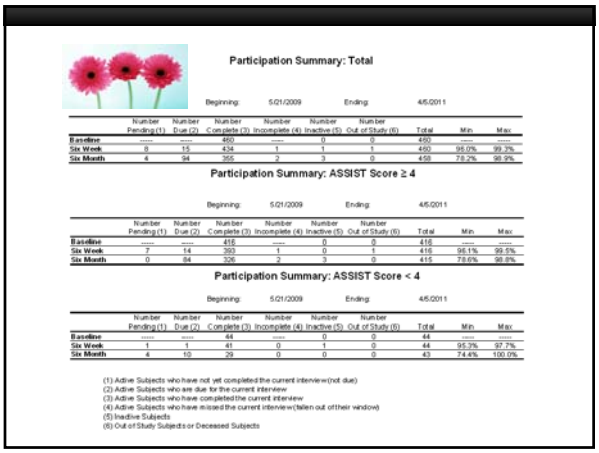
Actual vs. Targeted Enrollment



Consort Diagram For DSMB Meeting



BOSTON UNIVERSITY



Participation Summary: Total

Beginning: 5/21/2009 Ending: 4/5/2011

	Number Pending (1)	Number Due (2)	Number Complete (3)	Number Incomplete (4)	Number Inactive (5)	Number Out of Study (6)	Total	Min	Max
Baseline	-----	-----	460	-----	0	0	460	-----	-----
Six Week	8	15	434	1	1	1	460	96.0%	99.3%
Six Month	4	94	355	2	3	0	458	78.2%	98.9%

Participation Summary: ASSIST Score ≥ 4

Beginning: 5/21/2009 Ending: 4/5/2011

	Number Pending (1)	Number Due (2)	Number Complete (3)	Number Incomplete (4)	Number Inactive (5)	Number Out of Study (6)	Total	Min	Max
Baseline	-----	-----	416	-----	0	0	416	-----	-----
Six Week	7	14	393	1	0	1	416	96.1%	99.5%
Six Month	0	84	326	2	3	0	415	78.6%	98.8%

Participation Summary: ASSIST Score < 4

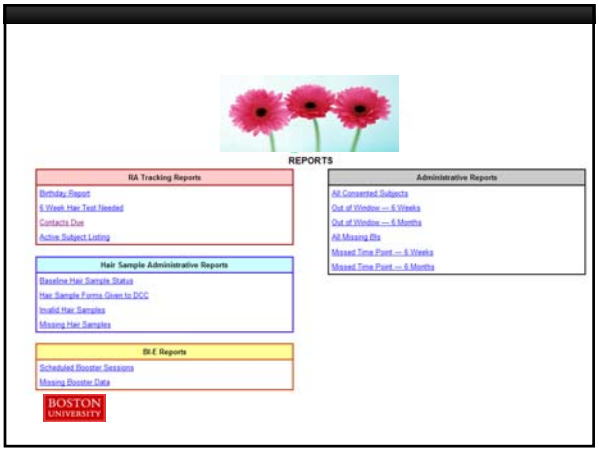
Beginning: 5/21/2009 Ending: 4/5/2011

	Number Pending (1)	Number Due (2)	Number Complete (3)	Number Incomplete (4)	Number Inactive (5)	Number Out of Study (6)	Total	Min	Max
Baseline	-----	-----	44	-----	0	0	44	-----	-----
Six Week	1	1	41	0	1	0	44	95.3%	97.7%
Six Month	4	10	29	0	0	0	43	74.4%	100.0%

(1) Active Subjects who have not yet completed the current interview (not due)
(2) Active Subjects who are due for the current interview
(3) Active Subjects who have completed the current interview
(4) Active Subjects who have missed the current interview (fallen out of their window)
(5) Inactive Subjects
(6) Out of Study Subjects or Deceased Subjects

Reports

- For the study staff to help manage tasks and know what needs to be done, when



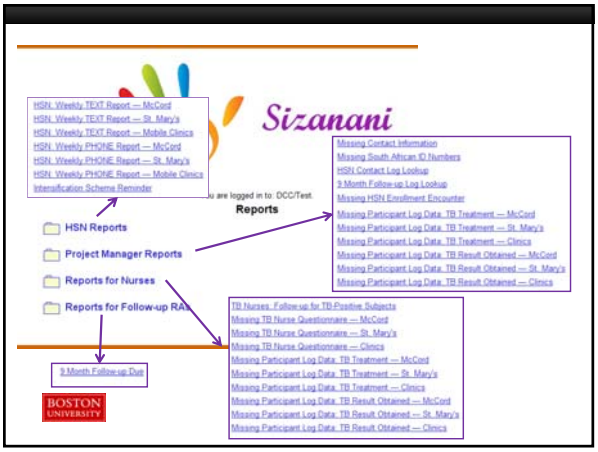
REPORTS

RA Tracking Reports
- Birthday Report
- 6 Week Hair Test Needed
- Contacts Due
- Active Subject Listing

Administrative Reports
- All Consented Subjects
- Out of Window — 6 Weeks
- Out of Window — 6 Months
- All Missing Bls
- Missed Time Point — 6 Weeks
- Missed Time Point — 6 Months

Hair Sample Administrative Reports
- Baseline Hair Sample Status
- Hair Sample Forms Given to DCC
- Invalid Hair Samples
- Missing Hair Samples

BLE Reports
- Scheduled Booster Sessions
- Missing Booster Data



Sample Tracking Report: Follow UP

Sizanani

You are logged in to: DCC/Test

9-Month Follow-Up Due

Date: 04/04/2013

ID	Enrollment Date	Name	Cell #	Best #	Family/Friend Best #	Family/Friend Name	Verification	Target Date	Window Open	Window Close
20737	14/05/2012							14/02/2013	31/01/2013	[illegible]
10733	14/05/2012							14/02/2013	31/01/2013	[illegible]
20749	28/05/2012							28/02/2013	14/02/2013	11/04/2013
20756	05/06/2012							05/03/2013	19/02/2013	16/04/2013
20767	15/06/2012							15/03/2013	01/03/2013	26/04/2013
10768	18/06/2012							18/03/2013	04/03/2013	29/04/2013

Look at the Data Early and Often

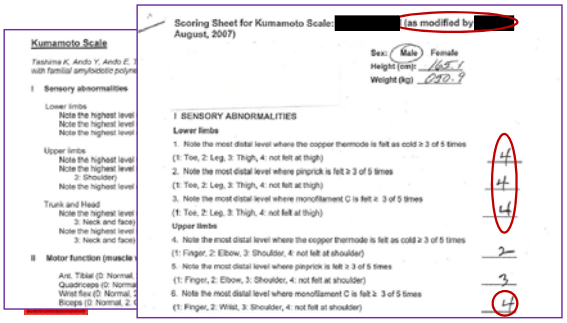
- You cannot fix a problem if you don't know it exists
- Get data into electronic format that can be manipulated ASAP so it can be more easily reviewed
- Monitor every data point for the first few participants
- Ongoing: audit percentage of forms
- Pay extra attention to key variables

Do simple checks

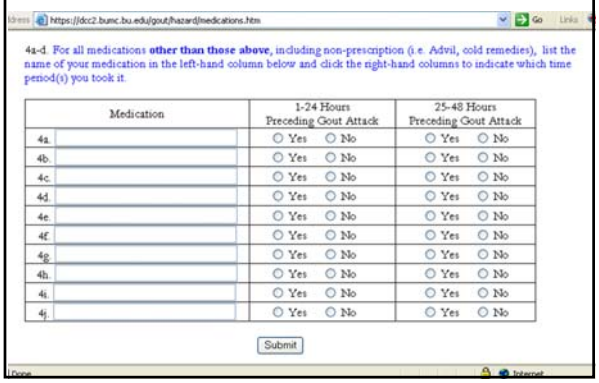
- Frequency (count) and distribution (range) of each and every variable
- Do crosstabs of variables where appropriate
- What is missing?
- What is out of range?
- What contradicts (e.g., pregnant males)
- Are there systemic problems?

BOSTON UNIVERSITY

This is why you check…



Medications are never easy

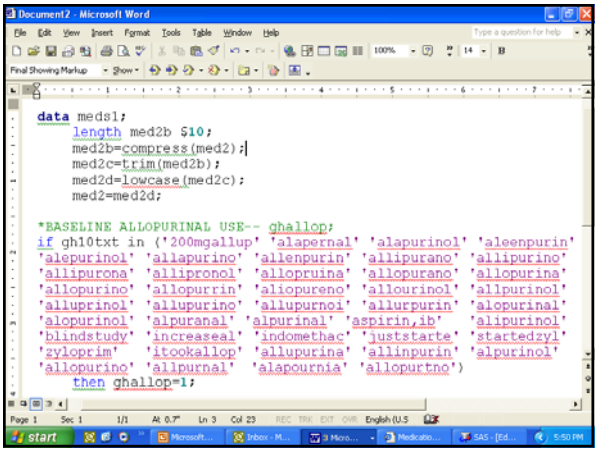


https://dcc2.bumc.bu.edu/gout/hazard/medications.htm

4a-d. For all medications **other than those above**, including non-prescription (i.e. Advil, cold remedies), list the name of your medication in the left-hand column below and click the right-hand columns to indicate which time period(s) you took it.

	Medication	1-24 Hours Preceding Gout Attack	25-48 Hours Preceding Gout Attack
4a.		○ Yes ○ No	○ Yes ○ No
4b.		○ Yes ○ No	○ Yes ○ No
4c.		○ Yes ○ No	○ Yes ○ No
4d.		○ Yes ○ No	○ Yes ○ No
4e.		○ Yes ○ No	○ Yes ○ No
4f.		○ Yes ○ No	○ Yes ○ No
4g.		○ Yes ○ No	○ Yes ○ No
4h.		○ Yes ○ No	○ Yes ○ No
4i.		○ Yes ○ No	○ Yes ○ No
4j.		○ Yes ○ No	○ Yes ○ No

Submit



Document2 - Microsoft Word

```
data meds1;
    length med2b $10;
    med2b=compress(med2);
    med2c=trim(med2b);
    med2d=lowcase(med2c);
    med2=med2d;

*BASELINE ALLOPURINAL USE-- ghallop;
if gh10txt in ('200mgallup' 'alapernal' 'alapurinol' 'aleenpurin'
'alepurinol' 'allapurino' 'allenpurin' 'allipurano' 'allipurino'
'allipurona' 'allipronol' 'allopruina' 'allopurano' 'allopurina'
'allopurino' 'allopurrin' 'aliopureno' 'allourinol' 'allpurinol'
'alluprinol' 'allupurino' 'allupurnoi' 'allurpurin' 'alopurinal'
'alopurinol' 'alpuranal' 'alpurinal' 'aspirin,ib' 'alipurinol'
'blindstudy' 'increaseal' 'indomethac' 'juststarte' 'startedzyl'
'zyloprim' 'itookallop' 'allupurina' 'allinpurin' 'alpurinol'
'allopurino' 'allpurnal' 'alapournia' 'allopurtno')
    then ghallop=1;
```

Perform Systematic Data Audits

- Data forms and source documents are compared with database on X % of forms
- Set an "acceptable" error rate. For example:
  - 0.1% for key variables
  - 0.5% overall
- If audit yields a larger error rate, you must check and correct the database

BOSTON UNIVERSITY

Audit Example (real data)

6-Month Follow-Up Assessment (Interviewer Administered) – Data Discrepancies

Subject ID	Field Name	CRF	Database	Notes
1115	Interdate_6	10/20/08	03/30/2009	*Check entire CRF*
	Site	1	3	
	Site_other	(text)	-888	
	Interstart	12:00	13:30	
	Interfinish	12:30	14:00	
	HIV4A_6	Blank	480	
	HIV4A_OK_6	Checked	blank	
	SP3a_1_6	2	1	
	SP4b_6	3	2	
	SP4e_6	15	10	
	SP4f_1_6	0	1	
	SP4f_2_6	0	1	
	SP4f_3_6	0	1	
	SP4g_6	1	-888	
	SP4h_6	1	-888	
	SP4i_1_6	1	-888	
	SP4g_2_6	0	-888	
	SP4g_3_6	0	-888	
	SP4g_4_6	0	-888	
	SP4g_5_6	0	-888	
	SP13_6	5	0	
	SP14_6	1	0	
	SP15_6	2	0	
	SP18_6	1	0	
	STDIG1_6	3	2	

Entered under incorrect ID?

Pay Extra Attention To Key Data

Be sure to pay particular attention to key data points where applicable.

- Query all entries of critical variables (e.g., primary outcome)
- Extra attention to problematic variables (e.g., time-line-follow-back)
- Query all Serious Adverse Events?

Derived Variables

Many analyses require creation of a derived variable from multiple data points

- Be especially careful in creating derived variables
- Include all relevant data elements
- Don't forget to account for missing data
- Be sure to look at frequencies and cross-tabs of derived variables prior to including in models

Creating a Derived Variable

Q1. Does child smoke

Q2. Do household members smoke?

Q3. Do caretakers smoke?

New Var: Smoke_Exp

Q1. Unprotected sex primary partner?

Q2. Unprotected sex with casual partner?

Q3. Share needles?

New Variable: HIV_Exp

Sample SAS Code for Derived Variable

```
if (q1=1) or (q2=1) or (q3=1) then any_exp=1; else
    any_exp=2;

  proc freq;
    tables any_exp*site;
    run;
```

Any_Exp	Yes	No	Total
Site 1	50	40	90
Site 2	20	70	90
Total	70	110	180

Sample SAS Code for Derived Variable

```
/* Corrected code to account for missing */
    q1=1) or (q2=1) or (q3=1) then any_expM=1; else
        if (q1=0) and (q2=0) and (q3=0) then any_expM=2; else
            any_expM=.;

  proc freq;
    tables any_expM*site;
    run;
```

Any_ExpM	Yes	No	Missing	Total
Site 1	50	40		90
Site 2	20	20	50	90
Total	70	60	50	180

Example 1 missing coded no

The FREQ Procedure

Frequency / Percent / Row Pct / Col Pct	Table of row by col		
row	1	2	Total
1	50 27.78 55.56 71.43	40 22.22 44.44 36.36	90 50.00
2	20 11.11 22.22 28.57	70 38.89 77.78 63.64	90 50.00
Total	70 38.89	110 61.11	180 100.00

Statistics for Table of row by col

Statistic	DF	Value	Prob
Chi-Square		21.0390	<.0001

Example 2 missing coded missing

The FREQ Procedure

Frequency / Percent / Row Pct / Col Pct	Table of row by col		
row	1	2	Total
1	50 38.46 55.56 71.43	40 30.77 44.44 66.67	90 69.23
2	20 15.38 50.00 28.57	20 15.38 50.00 33.33	40 30.77
Total	70 53.85	60 46.15	130 100.00

Statistics for Table of row by col

Statistic	DF	Value	Prob
Chi-Square	1	0.3439	0.5576

What's up with the missing values?

- Go back and look at forms:
  - Is there an explanation?
  - Is the missing data differential?
- What are the implications?
  - Example: There are 2 sites and all the forms with missing values came from a single site
- Did you find this problem early enough to correct it?
- This is why you check "early and often"

Data Security - General

- Keep paper records should be kept in locked cabinets and/or offices
- Store identifiers like names and addresses separate from clinical data
- Keep particularly sensitive data apart from other identifiers (e.g., SSN) – in a separate file, by ID
- Do not collect sensitive data unless you *really* need it

Data Security - Hardware

- Password protect all computers
- Set to automatically timeout if inactive
- Encrypt laptops, flash-drives and other storage devices when possible
- Do not put identifiable data on portable media (e.g., CDs, flash-drives) unless password protected, preferably encrypted

Take Home Message

- Your team should include someone who understands data issues
- Budget for data management
- Planning ahead results in fewer revisions
- Check your data early and often
- If you do things correctly from the beginning:
  - It is less work
  - It is less expensive
  - You are more likely to discover the truth at the end

Questions?