Download Files for Class

http://bit.ly/repro_files

(https://github.com/ncontaxis/reproworkshop)

Reproducibility Workshop

July 9, 2018

Alisa Surkis, Assistant Director, Research Data and Metrics
Fred LaPolla, Research and Data Librarian
Health Sciences Library

Mark Butler, Clinical Data Analyst
Department of Population Health

Learning Objectives

Apply research data management best practices to your research

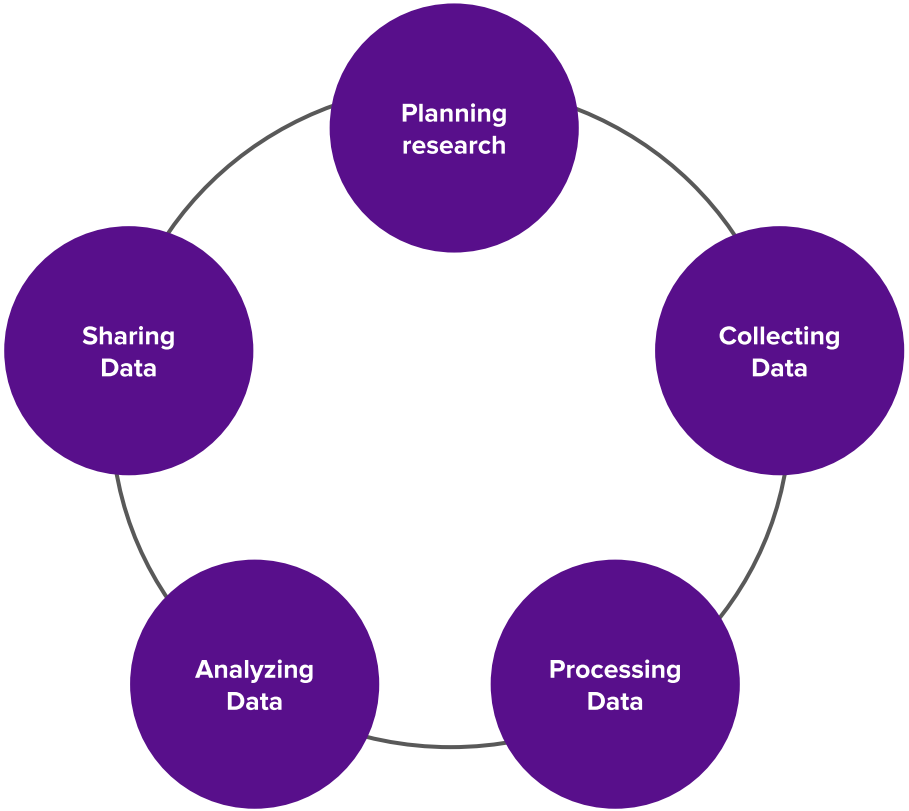
Construct REDCap projects that minimize data collection errors

Import/Export data into REDCap projects to improve research efficiency

Develop code for data processing and analysis in R Markdown

Apply good practices to coding process to improve reproducibility

Data Lifecycle



Reproducibility: Why

nature REVIEWS DRUG DISCOVERY

Nature Reviews Drug Discovery **10**, 712 (September 2011) | doi:10.1038/nrd3439-c1

Believe it or not: how much can we rely on published data on potential drug targets?

- Analysis of target identification and validation projects at Bayer in oncology, women's health, cardiovascular diseases
- 21% where data in literature was consistent with in-house data

Reproducibility: Why

nature

Drug development: Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis

- Scientists in haematology and oncology departments at Amgen tried to confirm findings from 53 “landmark” studies
- Findings confirmed in only 6 (11%) cases.

Reproducibility: Why

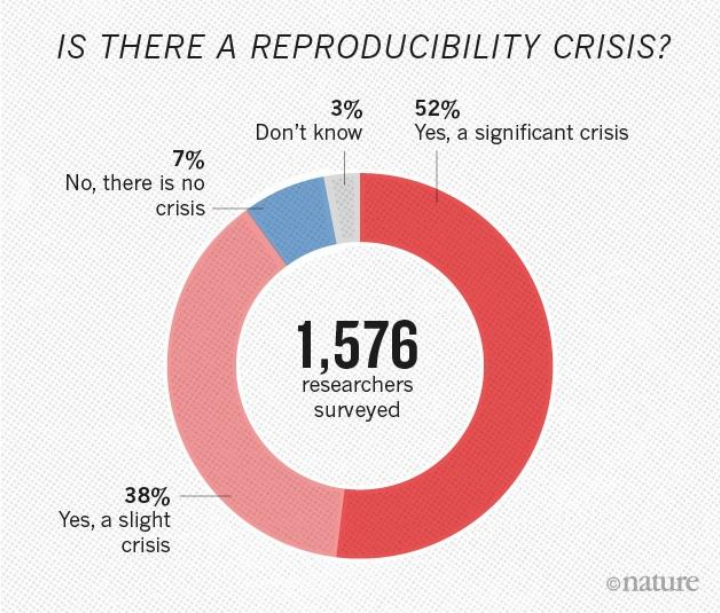
NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

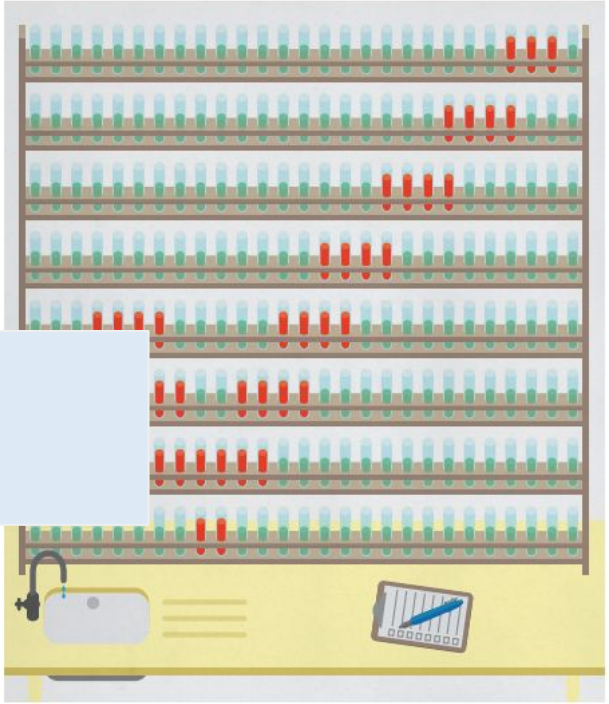
Monya Baker

25 May 2016 | Corrected: 28 July 2016



Reproducibility: Why

Policy announced October 2015
In effect as of January 2016



CHRIS RYAN/NATURE

NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring

shorter term, however, the checks and balances that once ensured scientific fidelity have been hobbled. This has compromised

outnumbered by the hundreds of thousands published each year in good faith.

Instead, a complex array of other factors seems to have contributed to the lack of reproducibility. Factors include poor training of researchers in experimental design; increased emphasis on making provocative statements rather than presenting technical details; and publications that do not report basic elements of experimental design[4]. Crucial experimental design elements that are all too frequently ignored include blinding, randomization, replication, sample-size calculation and the effect of sex differences. And some scientists reputedly use a 'secret sauce' to make their experiments work — and withhold details from publication or describe them only vaguely to retain a competitive edge[5]. What hope is there that other scientists will be able to build on such work to further biomedical progress?

Exacerbating this situation are the policies and attitudes of funding agencies, academic centres and scientific publishers. Funding agencies often uncritically encourage the overvaluation of research published in high-profile journals. Some academic centres also provide incentives for publications in such journals, including promotion and tenure, and in extreme circumstances, cash rewards[6].

Then there is the problem of what is not published. There are few venues for researchers to publish negative data or papers that point out scientific flaws in previously published work. Further compounding the problem is the difficulty of accessing unpublished data — and the failure of funding agencies to establish or enforce policies that insist on data access.

PRECLINICAL PROBLEMS

Reproducibility is potentially a problem in all scientific disciplines. However, human clinical trials seem to be less at risk because they are already governed by various regulations that stipulate rigorous design and independent oversight — including randomization, blinding, power estimates, pre-registration of outcome measures in standardized, public databases such as ClinicalTrials.gov and oversight by institutional review boards and data safety monitoring boards. Furthermore, the clinical trials community has taken important steps towards adopting standard reporting elements[7].

Reproducibility: Why

New guidelines for grants started January 25, 2016

- Scientific premise must describe strengths/weaknesses of prior research
- Scientific rigor to ensure robust/unbiased experimental design, methodology, analysis, interpretation, reporting of results
- Consideration of relevant biological variables
- Authentication of key biological/chemical resources

Reproducibility: Why

New guidelines for grants started January 25, 2016

- Scientific premise must describe strengths/weaknesses of prior research
- **Scientific rigor to ensure robust/unbiased** experimental design, methodology, analysis, interpretation, **reporting of results**
- Consideration of relevant biological variables
- Authentication of key biological/chemical resources

Reproducibility: What

Reproducibility

Reproducibility: What

Reproducibility

Replicability

Reproducibility: What

Reproducibility

Replicability

Repeatability

Reproducibility: What

Reproducibility

Replicability

Repeatability

?

Reproducibility: What

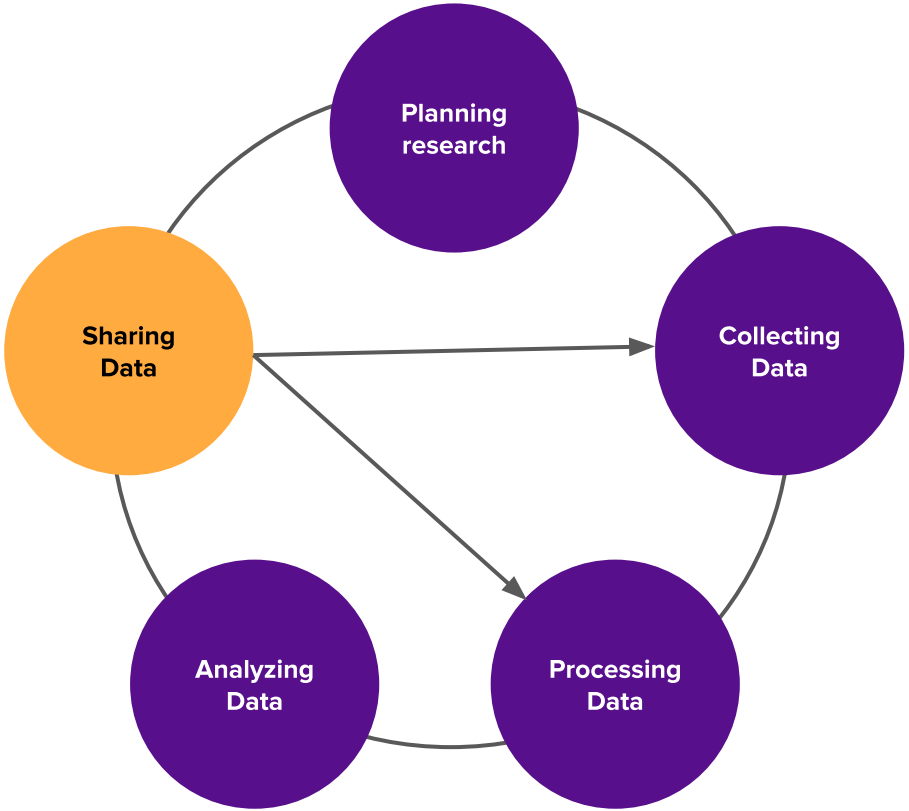
Reproducibility

Replicability

Repeatability

Different (often conflicting) meanings in different disciplines!

Reproducibility: What



Reproducibility: What

Reproducibility

Replicability

Repeatability

Different (often conflicting) meanings in different disciplines!

Key concept:

independently reimplementing an experiment	vs.	running the same code using the same data to get the same result

Reproducibility: How

- Documentation

- Workflows

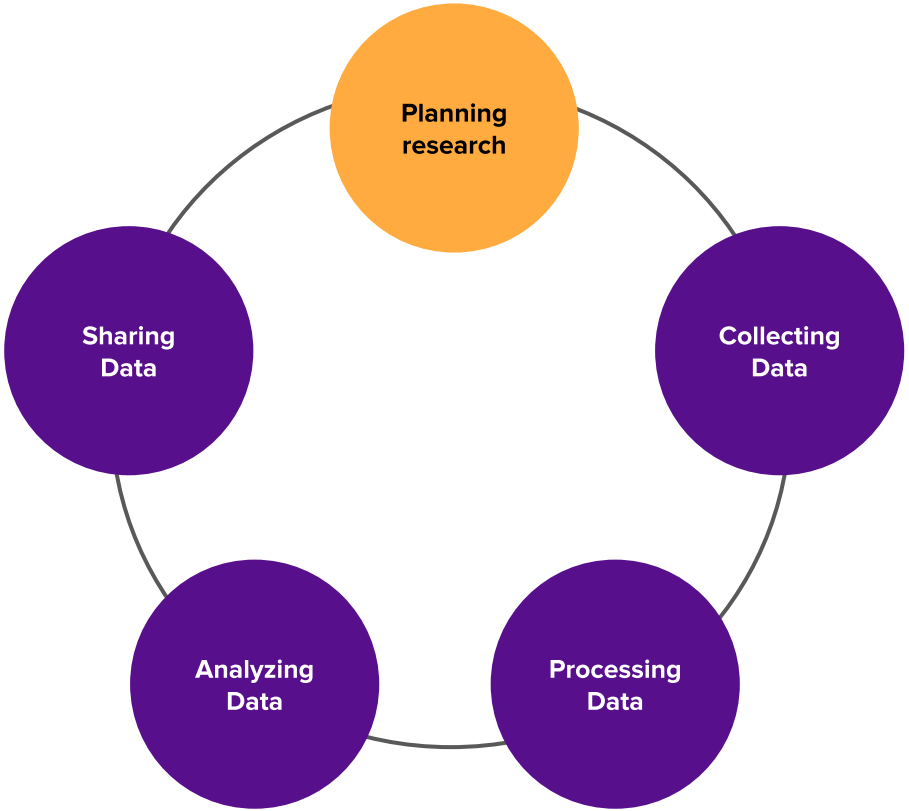
Reproducibility: How

- Documentation
  - Data dictionary
  - Experimental protocol
  - Well-documented code
  - Code inputs and versions
  - Workflows
- Workflows

Reproducibility: How

- Documentation
  - Data dictionary
  - Experimental protocol
  - Well-documented code
  - Code inputs and versions
  - Workflows
- Workflows
  - Clearly outlined procedures
  - Well-defined roles
  - Always maintain untouched version of raw data

Data Lifecycle



Creating instruments/forms/CRFs

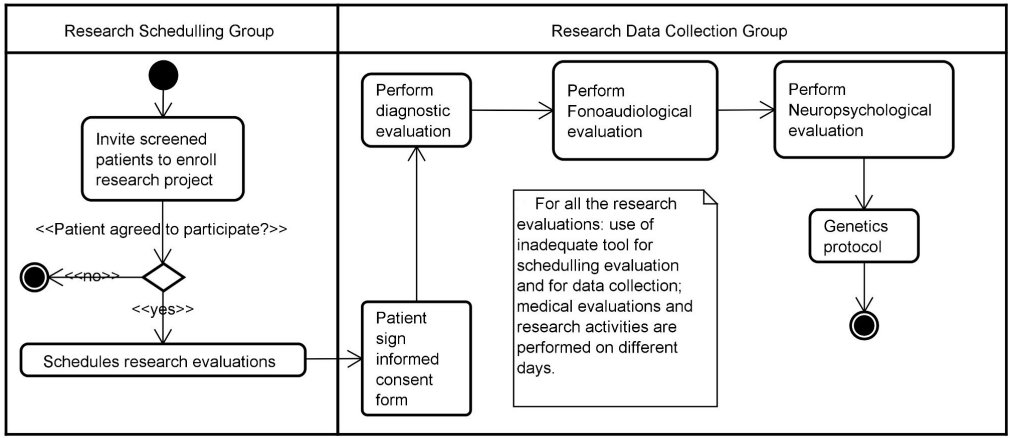
A data collection plan is not an afterthought

Creating instruments/forms/CRFs

Questions to ask before you begin a study:

- **What** are you collecting?
- **When** are you collecting?
- **Where** are you collecting?
- **Who** is doing the data collection?
- **How** are you collecting data?

Documenting workflows



Cofiel L, Bassi DU, Ray RK, Pietrobon R, Brentani H (2013) Detecting Dissonance in Clinical and Research Workflow for Translational Psychiatric Registries. PLOS ONE 8(9): e75167. https://doi.org/10.1371/journal.pone.0075167

Population: Teens with depression

Target: 100 subjects

Data collection information needed:
- Demographics, clinical measures, depression scale, MoCA

Resources:
- Data collection:
  - iPads (4)
  - REDCap software (https://openredcap.nyumc.org)
- Statistical Analysis Plan

Sites: After school programs in NYC
- 25 NYCHA Housing Authorities

Procedure: Meet with students during after school programs, collect data on iPads using REDCap

DATA COLLECTION AND DATA MANAGEMENT DOCUMENTATION

Creating instruments/forms/CRFs

What is the best method for collecting your data?

- Single fields
- Multiple choice/select all
- Matrix of fields

Best practices:

- Keep values consistent
- Use validation whenever possible

When do errors commonly occur?

When do errors commonly occur?

Data entry

- Human error

When do errors commonly occur?

Data entry

Flawed data entry process

- Unclear questions/variables
- Lack of instructions
- Missing validation

When do errors commonly occur?

Data entry

Flawed data entry process

Null problem

- Not known vs. not applicable

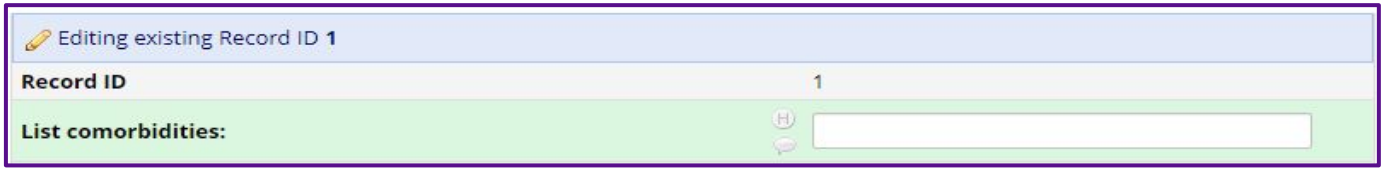
Why do errors commonly occur?

	DETECTION METHODS				
SOURCES OF ERROR	Programmatic Data Checks	Source Data Verification	Data Validation	Aggregate Statistics	CRF-to-Database Inspection
Subject completes questionnaire incorrectly or provides incorrect or incomplete answers to questions (lack of tool validation or bad form design)			X		
Subject does not follow trial conduct instructions		X			
Inadequate instructions given to the subject				X	
Site personnel trial conduct error (protocol violation)		X		X	
Data captured incorrectly on the source	X	X			
Site personnel transcription error	X	X	X		
Site equipment error				X	
Human error in reading equipment or print out or inter-rater-reliability		X			
Data entry error	X	X	X		X
Electronic data acquisition error (power glitch, back up that didn't run, lead not attached securely)			X		X
Data linked to the wrong subject		X	X		X
Database updated incorrectly from data clarification form or query					X
Missing data	X	X			
Outliers	X				
Data inconsistencies	X	X			
Programming error in user interface or database or data manipulations					X
Lost data		X	X		
Fraud		X		X	

- GCDMP, 2013

Collecting data: Example

Scenario: Collecting data on comorbidities



Collecting data: Example

Scenario: Collecting data on comorbidities

Editing existing Record ID **1**

Record ID	1
List comorbidities:	

VS

Select comorbidities	☐ Arthritis ☐ Sleep Apnea ☐ High Blood Pressure ☐ High Cholesterol ☐ Type 2 Diabetes ☐ Venous Stasis Disease ☐ Soft Tissue Infections ☐ Other

Collecting data: Example

Scenario: Collecting age

Age:

VS

Age range:

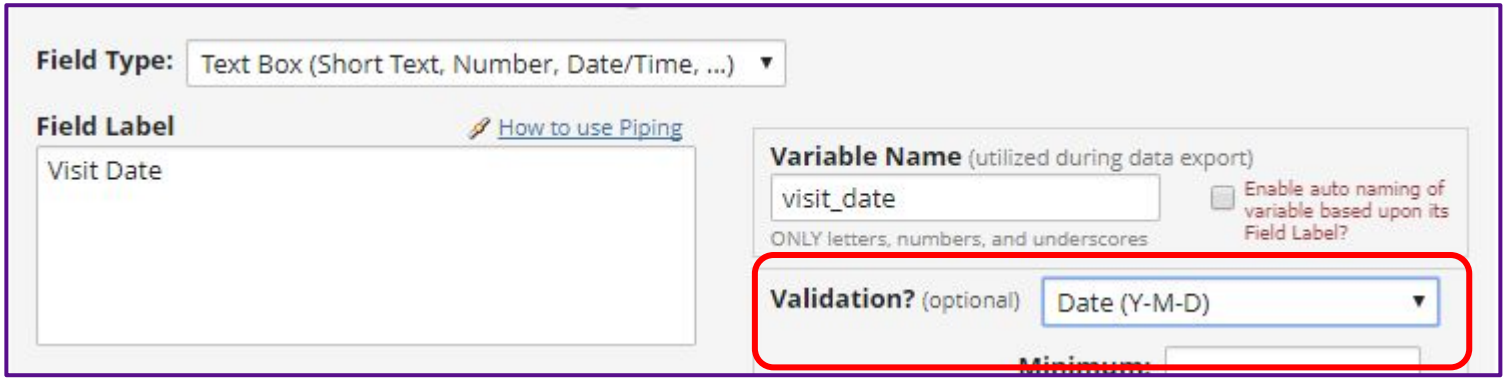
- 18-30
- 31-45
- 46-60
- 60-80
- +80

Collecting data best practice

RAW ✓ vs CATEGORICAL

Collecting data: Example

Scenario: Data validation



Use consistent codes and questions

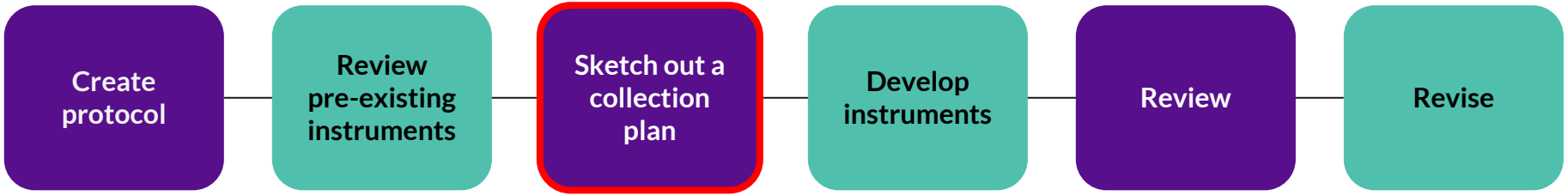
Alternating questions

- abnormal = yes | normal = yes

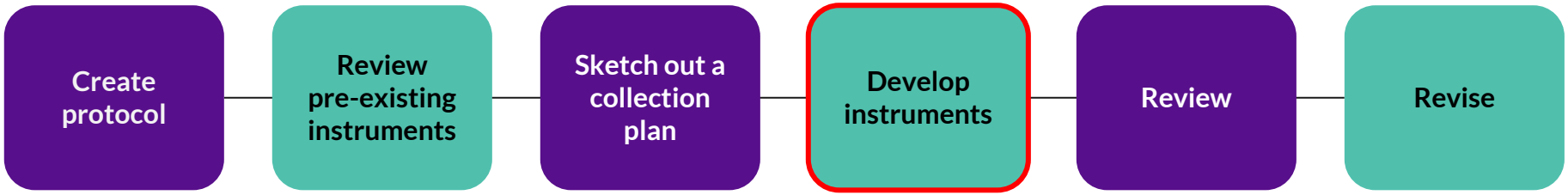
Ensure likert scales are the same across your forms

- 1=Not at all | 2=Sometimes | 3=Mostly | 4=Always
- 1=Always | 2=Mostly | 3=Sometimes | 4=Not at all

Data collection plan workflow



Data collection plan workflow



Instrument clarity and documentation

Provide detailed descriptions and instructions for all variables and data elements:

- What is the purpose of each variable?
- Do the instructions clearly explain how to enter data for that variable?
- Are your variable names understandable in a raw data file (e.g., spreadsheet, stats program)?
- What are the values associated with each variable?

Variable Name	Instructions/format	Values
brthdat	Section Header: *Demographics Information* What is the subject's date of birth? *Record the date of birth using the DD-MM-YYYY format.*	text (date_dmy)

Common data collection issues

Enter subject's weight:

Enter subject's height:

Weight	Height

Common data collection issues

Enter subject's weight:

185 ?

Enter subject's height:

180 ?

? Weight	? Height
185	180
70	6'2
10	165
145	5'9
48	120

Common data collection issues

Enter subject's weight in lbs:

185

Enter subject's height in cm:

180

Weight_lbs	Height_cm
185	180
155	185
210	165
145	155.5
130	120

Exercise: Data Collection Form and Spreadsheet

Definitions and acronyms

Crucial to the comprehensibility of your data

All potentially ambiguous terminology or acronyms should be defined in:

- Protocol
- Data management plan
- Data collection instruments
- Instructional information

Examples:

- ICF = informed consent form
- SVT = site visit template

Keep all source documentation!

- hospital records
- clinical and office charts
- laboratory notes
- memoranda
- subjects' diaries or evaluation checklists
- pharmacy dispensing records
- recorded data from automated instruments
- copies or transcriptions certified after verification as being accurate copies
- Microfiches
- photographic negatives
- microfilm or magnetic media X-rays
- subject files

Bargaje, C. (2011). Good documentation practice in clinical research. *Perspectives in Clinical Research*, *2*(2), 59–63. http://doi.org/10.4103/2229-3485.80368

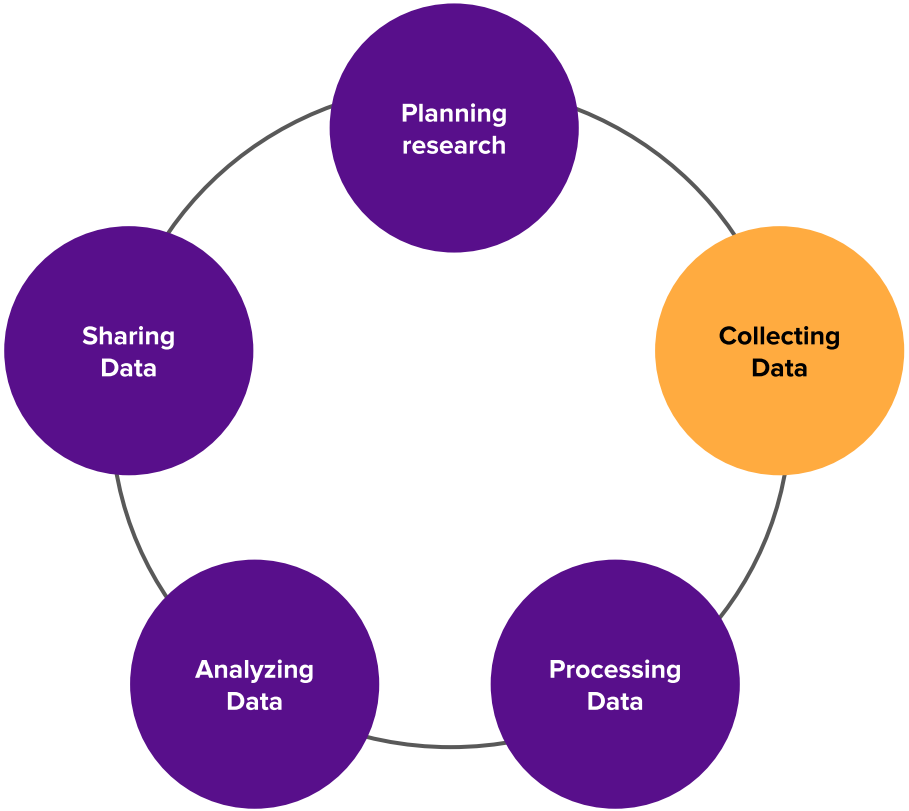
Documentation: Data Dictionary/Codebook

Instrument: **CDASH V 1.1 - Demographics** (cdash_v_11_demographics)

1	record_id	Record ID	text
2	brthdat	Section Header: *Demographics Information* What is the subject's date of birth? *Record the date of birth using the DD-MM-YYYY format.*	text (date_dmy)
3	age	What is the subject's age? *Record age of the subject in years.*	text
4	dmdat	What is the date of collection? *Record the date the demographics data were collected using the DD-MM-YYYY format.*	text (date_dmy)
5	sex	What is the sex of the subject? *Record the appropriate sex. Collect the sex or gender, as reported by the subject or caretaker. Select one.*	radio 1 Female 2 Male 3 Undifferentiated 99 Unknown
6	ethnic	What is the ethnicity of the subject? *Study participants should self-report ethnicity, with ethnicity being asked about before race. Select one.*	radio 1 Hispanic or Latino 2 Not Hispanic or Latino
7	race	What is the race of the subject? *Study participants should self-report race, with race being asked about after ethnicity. Check all that apply.*	checkbox 1 race___1 Black or African American 2 race___2 American Indian or Alaska Native 3 race___3 Asian 4 race___4 Native Hawaiian or Other Pacific Islander

REDCap
Research Electronic Data Capture

Data Lifecycle



NYU HEALTH SCIENCES LIBRARY

REDCap

Research Electronic Data Capture

What is REDCap?

What is REDCap?

Electronic data capture for research purposes

User-friendly, web-based program for creating:

- Data collection instruments (forms)
- Surveys

Choosing a REDCap Website

OPEN REDCap

Access from any internet connection

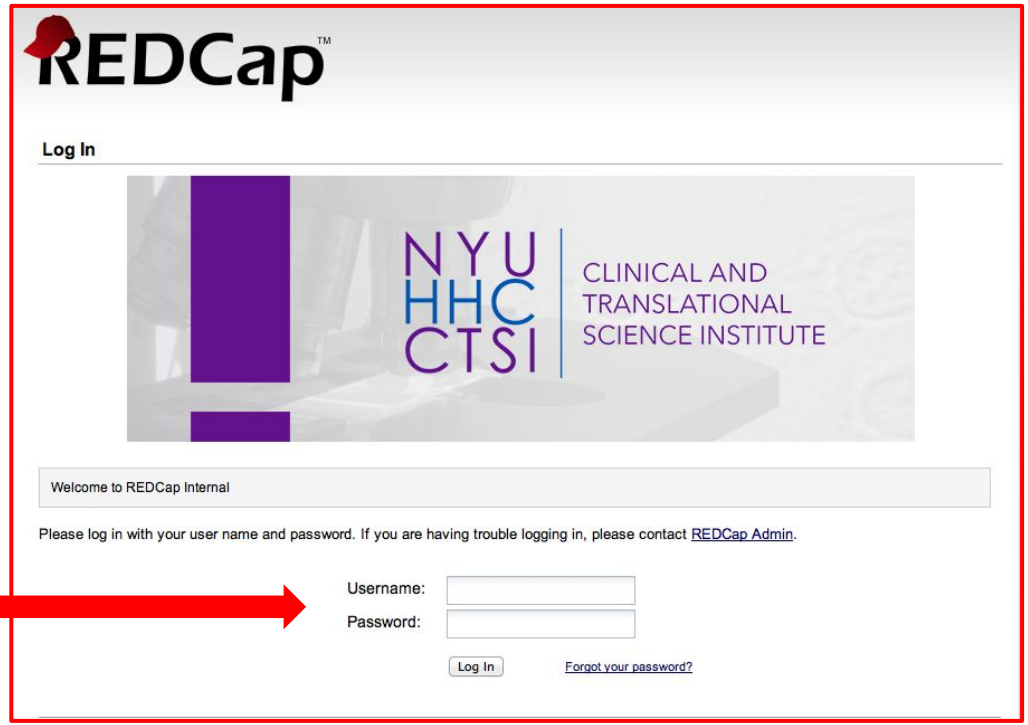
https://openredcap.nyumc.org

INTERNAL REDCap

Functional inside the NYU Langone firewall

https://redcap.nyumc.org

Creating an Account in REDCap



Requires Kerberos ID and password

NYU HEALTH SCIENCES LIBRARY

Creating an Account in REDCap

Sponsored Individuals:

MCIT Service Catalog
Create new account

Authentication Email from REDCap

REDCap **INTERNAL:**

Limited functionality within NYULMC region

https://redcap.nyumc.org

OPEN REDCap

Full function on any internet connection

https://openredcap.nyumc.org

From: redcap_admin@nyumc.org [redcap_admin@nyumc.org]

To:
Subject: [REDCap] Verify your email address

[This message was automatically generated by REDCap]

To complete the process of setting up a new primary email for your REDCap account with username "**Kerberos ID**" you will need to confirm your email address by clicking the link below. You will not be able to fully access your REDCap account until this verification process has been completed. Thank you!

Click here to confirm your email address

If the link above does not work, try copying the link below into your web browser:
http://redcap.nyumc.org/apps/redcap/index.php?user_verify=AgdQQXd2sc67pyAHu5LL

This link is unique to you and should not be forwarded to others.

Authentication Email from REDCap

REDCap **INTERNAL:**

Limited functionality within NYULMC region

https://redcap.nyumc.org

OPEN REDCap

Full function on any internet connection

https://openredcap.nyumc.org

From: redcap_admin@nyumc.org [redcap_admin@nyumc.org]

To:
Subject: [REDCap] Verify your email address

[This message was automatically generated by REDCap]

To complete the process of setting up a new primary email for your REDCap account with username "**Kerberos ID**" you will need to confirm your email address by clicking the link below. You will not be able to fully access your REDCap account until this verification process has been completed. Thank you!

Click here to confirm your email address

If the link above does not work, try copying the link below into your web browser:
http://redcap.nyumc.org/apps/redcap/index.php?user_verify=AgdQQXd2sc67pyAHu5LL

This link is unique to you and should not be forwarded to others.

Why REDCap?

Why REDCap?

Create instruments and surveys

Control users and collaborators

HIPAA compliant

Export data in various formats

Clean, high quality data

Starting a REDCap Project

REDCap Form Design

Online Designer:

- Point and click
- <30 variables

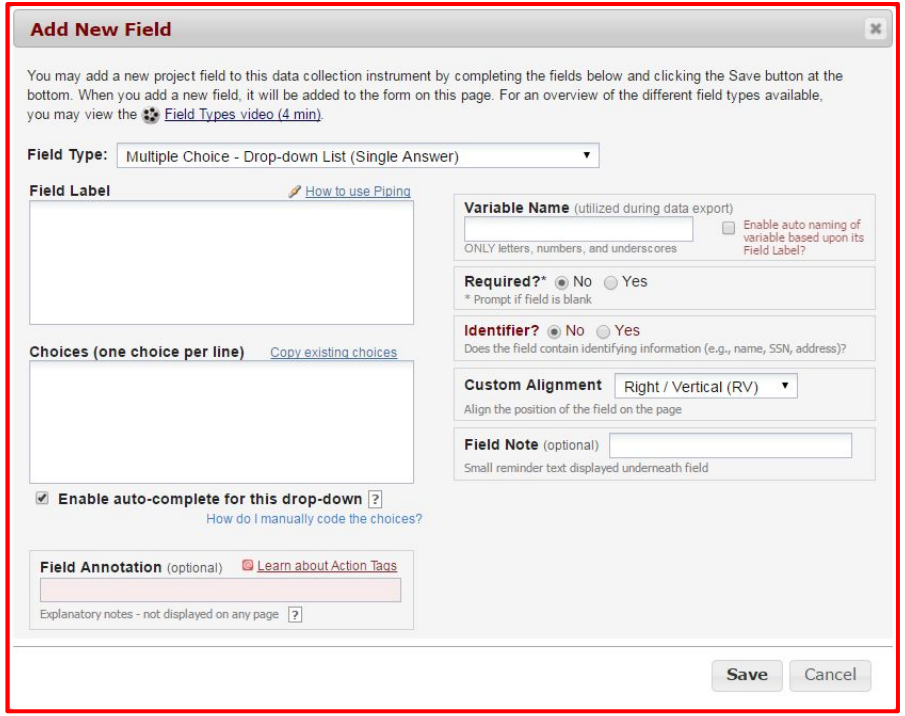
Data Dictionary:

- Spreadsheet to build forms
- >30 variables

REDCap Shared Library

- Existing, validated instruments

Online Designer



REDCap Activity

Data Dictionaries

Data Dictionary: Find the file called “reproworkshop_datadictionary.csv”

Go to Data Dictionary, upload this file and commit changes.

Data Import

Data Import

Import data into REDCap from other locations

Useful for batch data upload from existing spreadsheets/software

Some formatting required

Upload: REDCapData_ForImport to Data Import tool.

Data Export