2-22-2022

Data and how to manage it

Lori Sloane
Health Sciences Library and Informatics Center, University of New Mexico, lsloane@salud.unm.edu

HEALTH SCIENCES LIBRARY
& INFORMATICS CENTER

February 15, 2022

MANAGING YOUR DATA

LORI SLOANE, HSLIC DATA MANAGER

NIH Data Management and Sharing Policy & Related Announcements

NOT-OD-21-013 - Final NIH Policy for Data Management and Sharing

NOT-OD-21-014 – Common Elements of an NIH Data Management and Sharing Plan

NOT-OD-21-015 – Allowable Costs for Data Management and Sharing

NOT-OD-21-016 – Selecting a Repository for Data Resulting from NIH-Supported Research

NOT-OD-21-131 – Developing Consent Language for Future Use of Data and Biospecimens

Logistics

1. Effective Date: January 25, 2023
2. Required submission of DMSP **at the time of application**
3. Compliance will be monitored
   - Extramural Awards: The Plan will become a Term and Condition of the Notice of Award.
   - After the end of the funding period, non-compliance with the NIH ICO-approved Plan may affect future funding decisions for the recipient institution
4. Scope: “The DMS Policy applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data.”

New definition of Scientific Data

"The recorded factual material commonly accepted in the scientific community as **of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. "**

Scientific data *do not include*

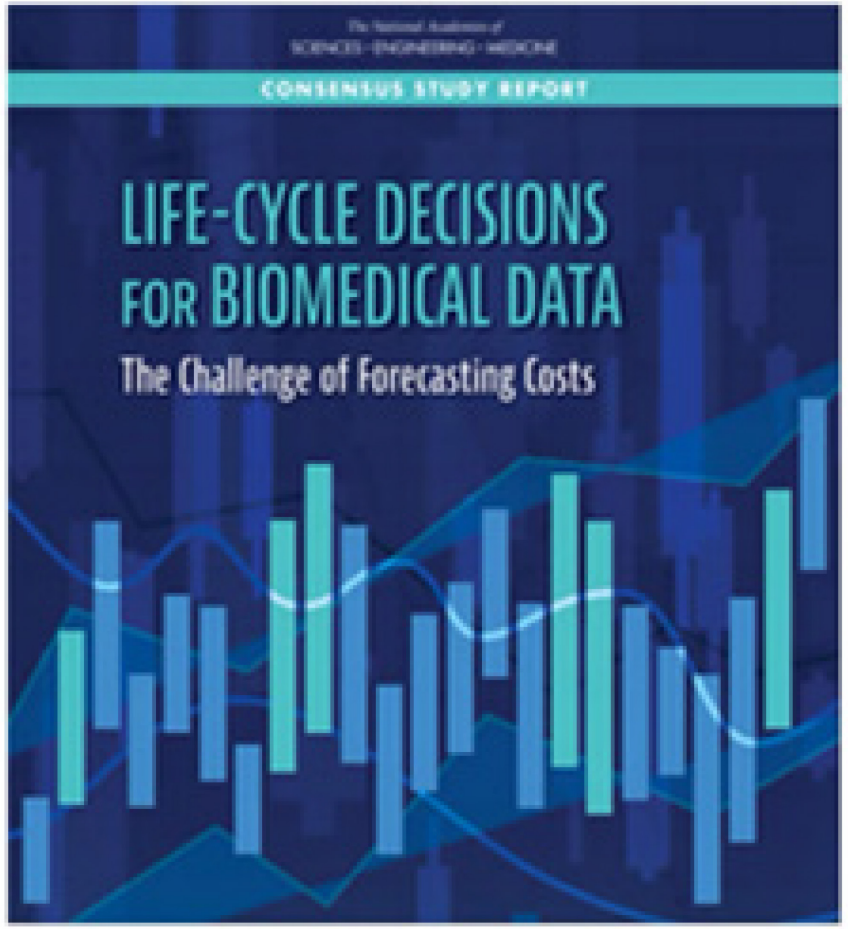
- *laboratory notebooks*
- *preliminary analyses*
- *completed case report forms*
- *drafts of scientific papers*
- *plans for future research*
- *peer reviews*
- *communications with colleagues*
- *physical objects (such as laboratory specimens)*

Costs & Budget

"Personnel costs required to perform the types of data management and sharing activities described in the final Supplemental Information are allowable."

"…funds for these activities must be spent during the performance period, even for scientific data and metadata preserved and shared beyond the award period. NIH funds cannot legally be spent after the award period."

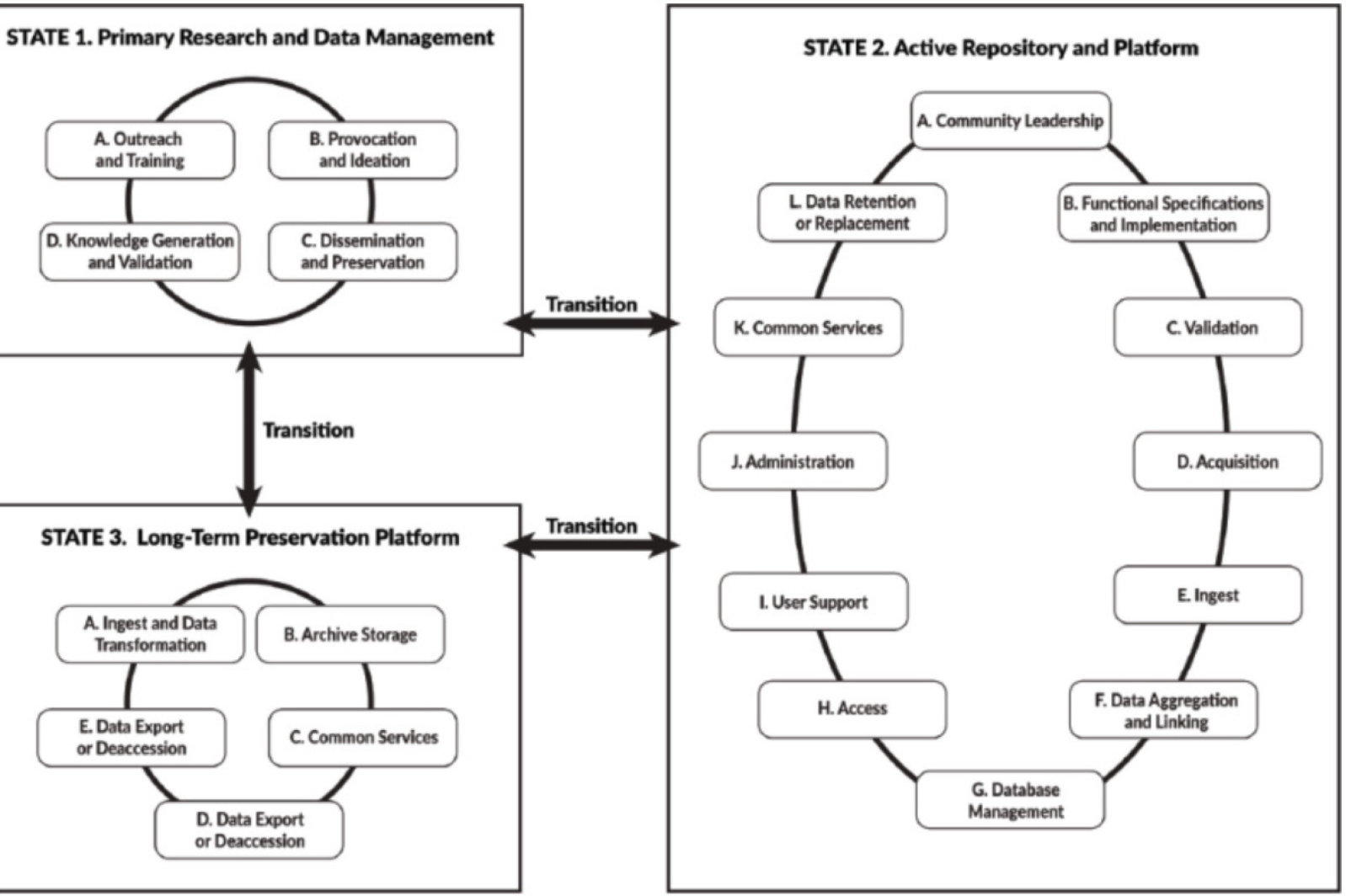
Forecasting Cost for Preserving, Archiving, and Promoting Access to Biomedical Data



Funded by the National Library of Medicine, this study by the National Academies of Sciences, Engineering, and Medicine developed and demonstrated a **framework for forecasting long-term costs** for preserving, archiving, and accessing various types of biomedical data and estimating potential future benefits to research.

https://vimeo.com/447601572

https://pubmed.ncbi.nlm.nih.gov/33026757/



Three data states

- State 1: Primary research/data management environment; data are captured and analyzed
- State 2: Active repository and platform; data may be acquired, curated, aggregated, accessed, and analyzed
- State 3: Long-term preservation platform

Elements of a Data Management and Sharing Plan

Data type
- Identifying data to be preserved and shared

Related tools, software, code
- Tools and software needed to access and manipulate data

Standards
- Standards to be applied to scientific data and metadata

Data preservation, access, timelines
- Repository to be used, persistent unique identifier, and when/how long data will be available

Access, distribution, reuse considerations
- Description of factors for data access, distribution, or reuse

Oversight of data management
- How plan compliance will be monitored/ managed and by whom

Choosing a data repository

"Primary consideration should be given to data repositories that are discipline or data-type specific to support effective data discovery and reuse."

Desirable characteristics (from NOT-OD-21-016)	
Unique persistent identifiers (DOI)	Clear use guidance
Long-term sustainability	Security and integrity
Metadata	Confidentiality
Curation & quality assurance	Common format
Free and easy access	Provenance
Broad and measure reuse	Retention policy

Additional Considerations for Human Data
(even if de-identified)

Fidelity to Consent: documented procedures to restrict dataset access and use to those that are consistent with participant consent

Restricted Use Compliant: documented procedures to communicate and enforce data use restrictions

Privacy: documentation to protect human subjects' data from inappropriate access

Plan for Breach: measures that include a response plan for detected data breaches

Download Control: Controls and audits access to and download of datasets (if download is permitted).

Violations: addressing violations of terms-of-use by users and data mismanagement by the repository

Request Review: makes use of an established and transparent process for reviewing data access requests

Selecting a repository: the NIH data sharing landscape

NIH strongly encourages
subject-specific, open access Data Sharing Repositories
as a first choice.
https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Datasets up to **2 gigabytes**

PubMed Central

Stores publication-related supplemental materials and datasets directly associated publications.

PMC

Datasets up to **20 gigabytes**

Generalist Repositories

Datasets associated with publications or otherwise and links to PubMed.

DRYAD, Dataverse, zenodo, OSF, figshare, MENDELEY

High priority datasets, **petabyte-scale**

Cloud Partners
(STRIDES Program)

Store and manage large scale, high priority NIH datasets.

aws

BMIC Repository list

Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

To help researchers locate an appropriate resource for sharing their data, as well as to promote awareness of resources where datasets can be located for reuse, BMIC maintains lists of several types of data sharing resources:

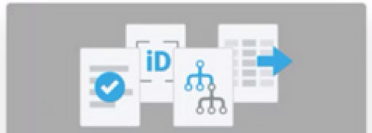
- Open NIH-supported domain-specific repositories that house data of a specific type or related to a specific discipline;
- Other NIH-supported domain-specific resources, including repositories and knowledgebases, that have limitations on submitting and/or accessing data; and
- Generalist repositories that house data regardless of type, format, content, or subject matter.

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

re3data is a global registry of research data repositories. The registry covers research data repositories from different academic disciplines. re3data presents repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions. re3data aims to promote a culture of sharing, increased access and better visibility of research data.

FAIRsharing.org
standards, databases, policies



1578 Standards

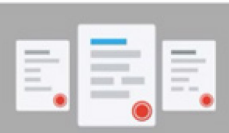
Terminology Artifact	829
Model/Format	504
Reporting Guideline	226
Identifier Schema	19

VIEW ALL

1858 Databases

Repositories	953
Knowledgebases	787
Knowledgebase/Repositories	118

VIEW ALL



158 Policies

Journal	94
Funder	23
Project	14
Society	13

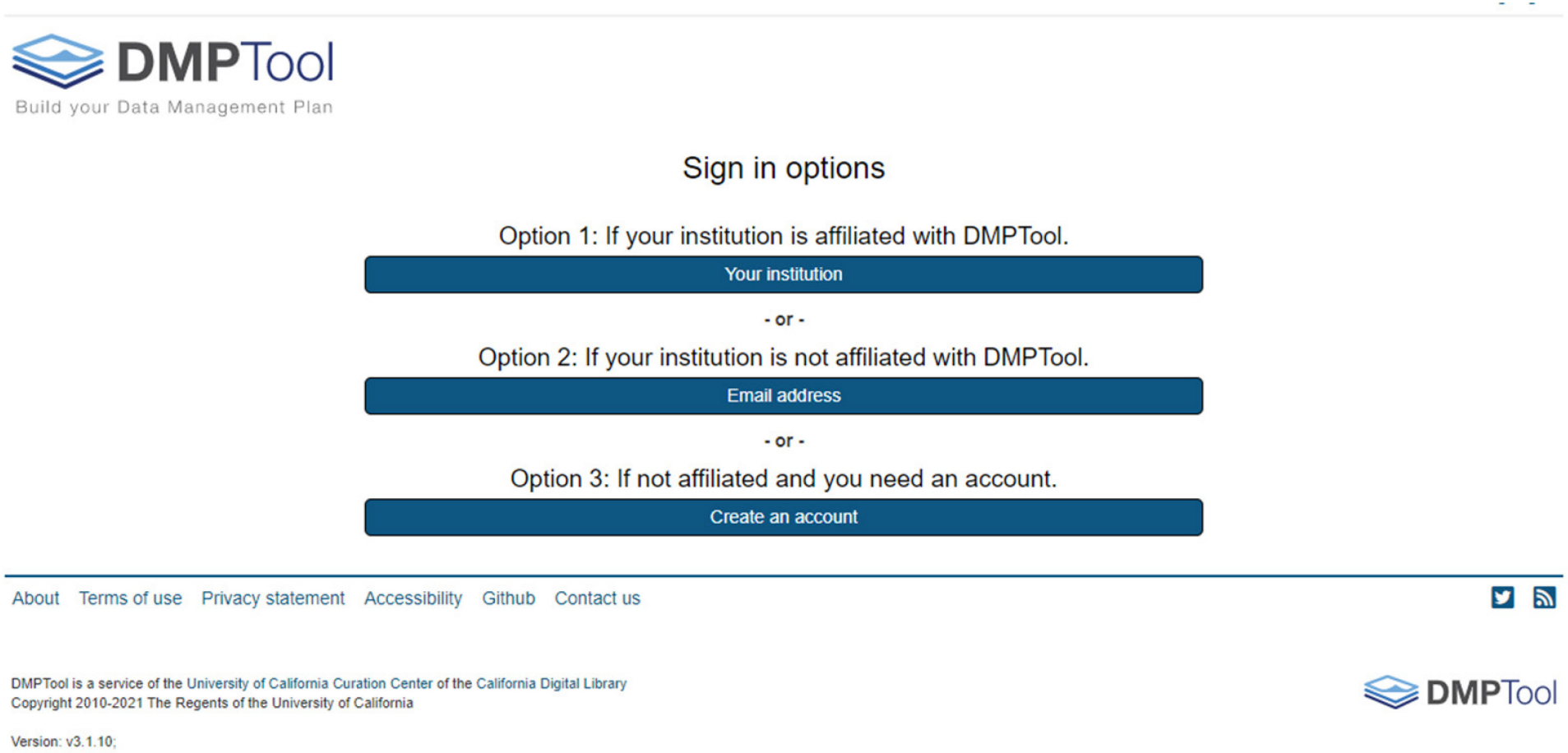
VIEW ALL

Acceptable reasons to not share

- Informed consent limitations on sharing
- Existing consent (e.g. for previously collected biospecimens) prohibits sharing
- Privacy or safety of research participants would be compromised
- Explicit federal, state, local, or Tribal law, regulation, or policy prohibits disclosure

Creating your DMP

DMPTool - https://dmptool.org/



University of New Mexico

DMPTool
Build your Data Management Plan

Data curation at University of New Mexico

My Dashboard Create plan Admin Features

Descriptive research on COViD-19 materials provided to Medical Advisory Team

Project Details | Collaborators | Write Plan | Research outputs | Download | Finalize / Publish

* Project title

Descriptive research on COViD-19 materials provided to Medical Advisory Team

☐ mock project for testing, practice, or educational purposes

Project abstract

A team was assembled at the University of New Mexico Health Sciences Center to help advise the Medical Advisory
Team to the Governor of New Mexico and Health Sciences Leadership in April of 2020.
An overwhelming amount of information was coming out in the news and literature in regards to the COVID-19

Research domain

Health sciences

Project Start

04/06/2020

Project End

06/30/2020

☐ Research outputs may have ethical concerns

Select Guidance

To help you write your plan, DMPTool can show you guidance from a variety of organizations.

Select up to 6 organizations to see their guidance.

☑ DMPTool
☑ University of New Mexico (UNM)

Find guidance from additional organizations below

See the full list

Save

Data collection best practices

- Have a plan developed beforehand and document it in detail
- Consider who, what, when, where, how you are going to collect your data from the outset
- Speak to a statistician and consider analysis throughout the process

Data Collection and Data Management Documentation

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://hsc.unm.edu/ctsc/services/informatics/redcap-login.html)
  - Statistical software
  - Statistical Analysis Plan

Sites: After school programs in NM

- 25 YMCA, NM Teen Project, Camp Fire NM, etc.

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

Creating instruments/forms/CRFs

What is the best method for collecting your data?

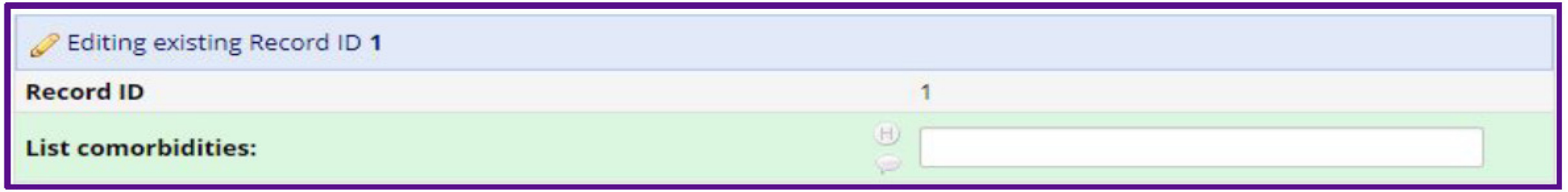
- Single fields
- Multiple choice/select all
- Matrix of fields

Best practices:

- Keep values consistent
- Use validation whenever possible

Collecting data: Example

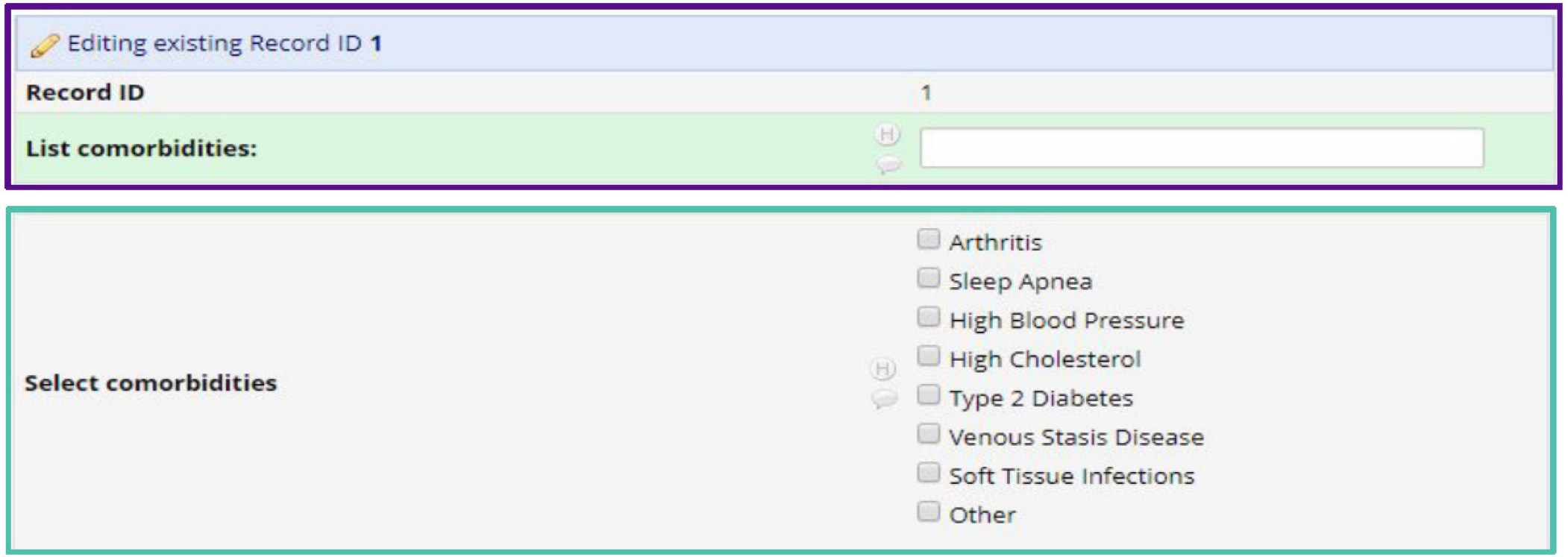
Scenario: Collecting data on comorbidities



https://hsc-redcap-idp.health.unm.edu/my.policy

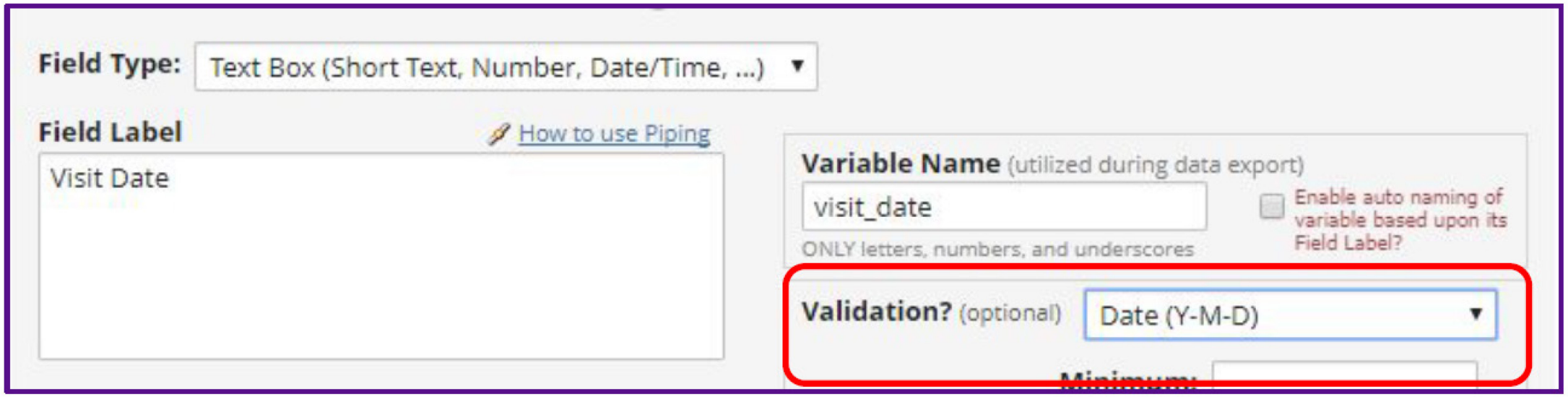
Collecting data: Example

Scenario: Collecting data on comorbidities



Collecting data: Example

Scenario: Data validation



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

AUTHOR_DATASET_ReadmeTemplate.txt

```
This DATSETNAMEreadme.txt file was generated on YYYY-MM-DD by NAME
<help text is included in angle brackets, and can be deleted before saving>


GENERAL INFORMATION

1. Title of Dataset:

2. Author Information
    A. Principal Investigator Contact Information
        Name:
        Institution:
        Address:
        Email:

    B. Associate or Co-investigator Contact Information
        Name:
        Institution:
        Address:
        Email:

    C. Alternate Contact Information
```

Medical coding dictionaries

Helps categorize medical terms (e.g. diagnoses, symptoms, adverse events) so that they can be analyzed/reviewed

Standardizes data across studies

Dictionary/Coding Example:

- ICD-10 codes

Record what medical coding dictionaries, validated instruments or standards will be used

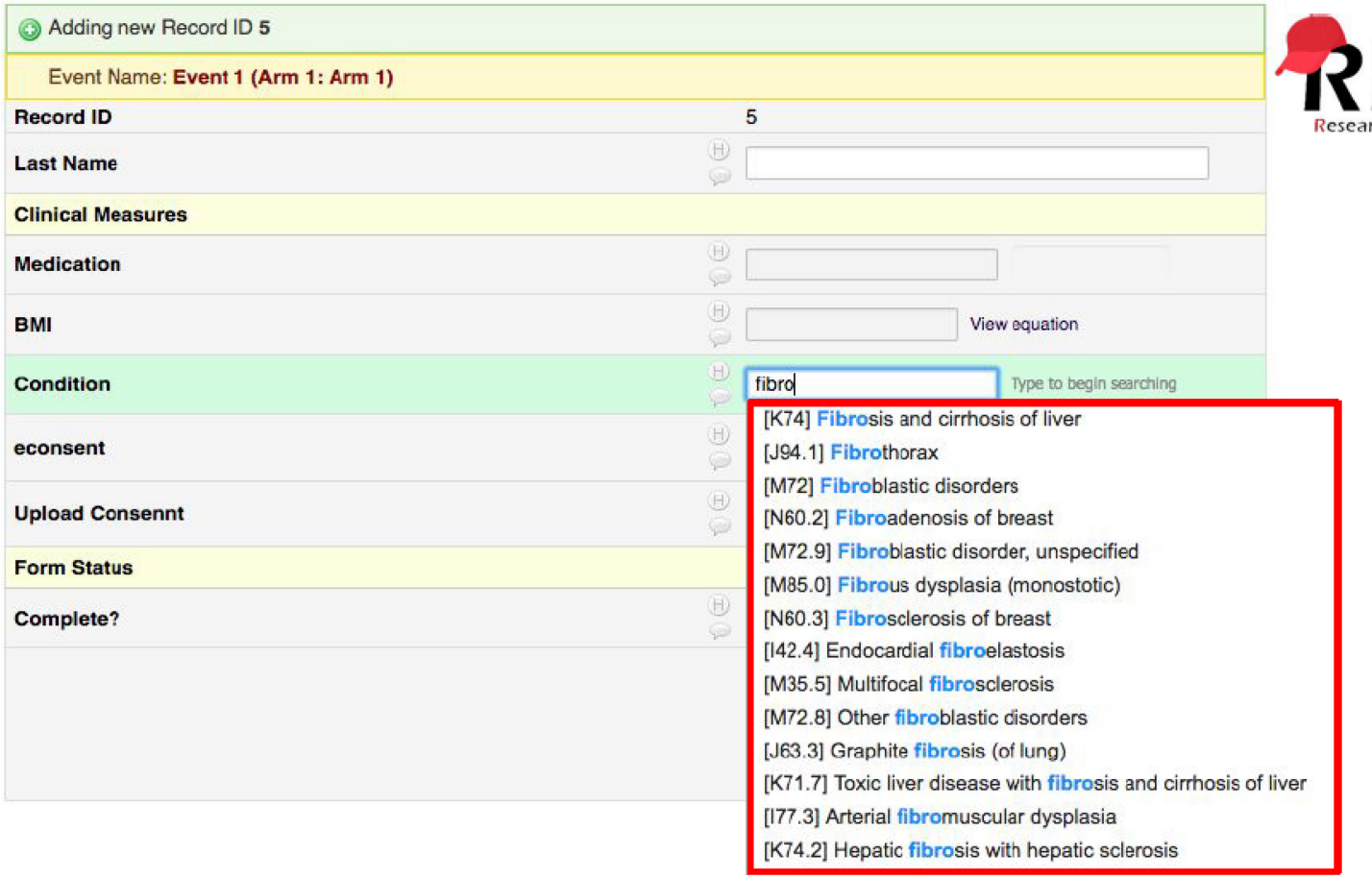
Indicate version of dictionary/instrument/standard used

Medical coding dictionaries

ICD-10 codes:

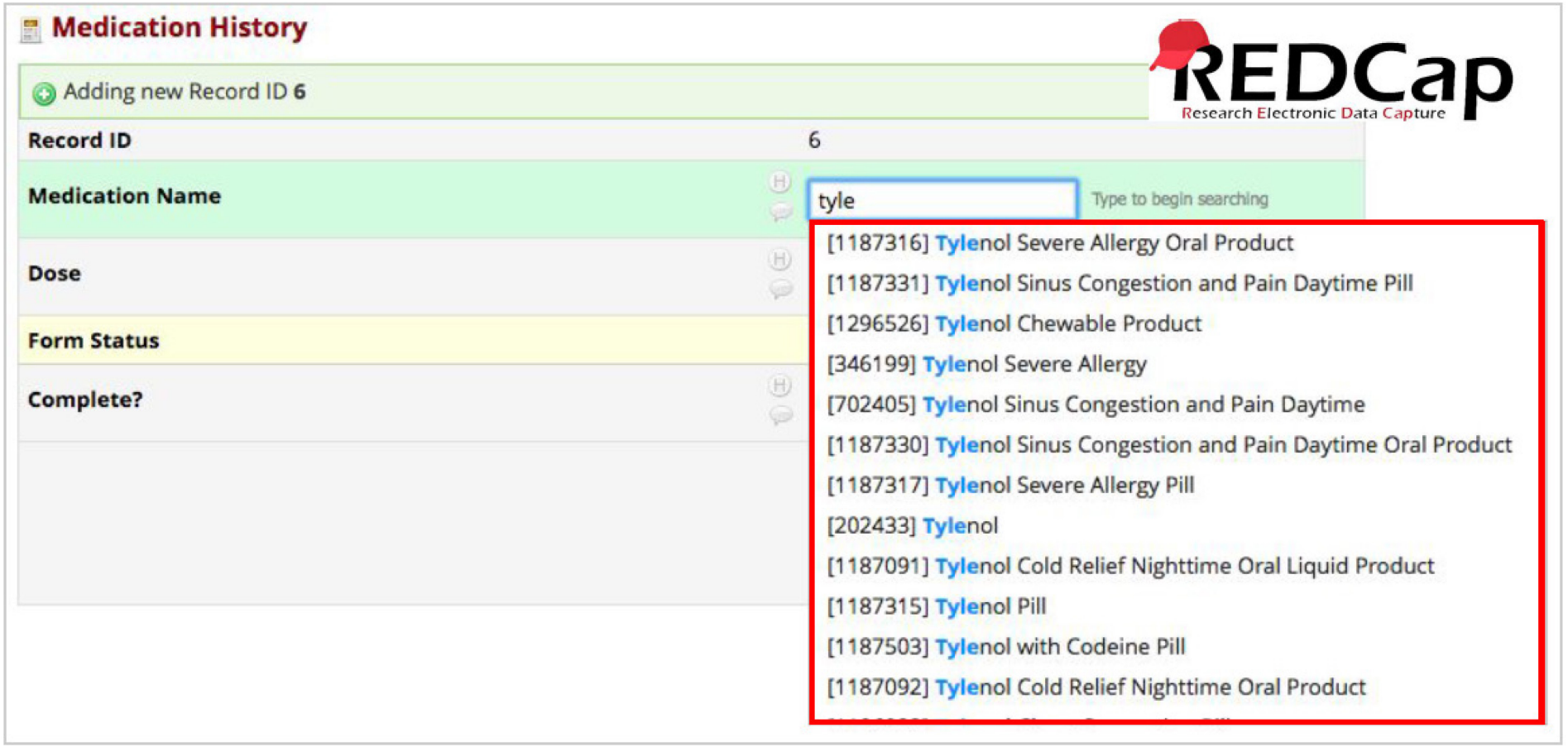
Medical coding dictionaries

ICD-10 codes:

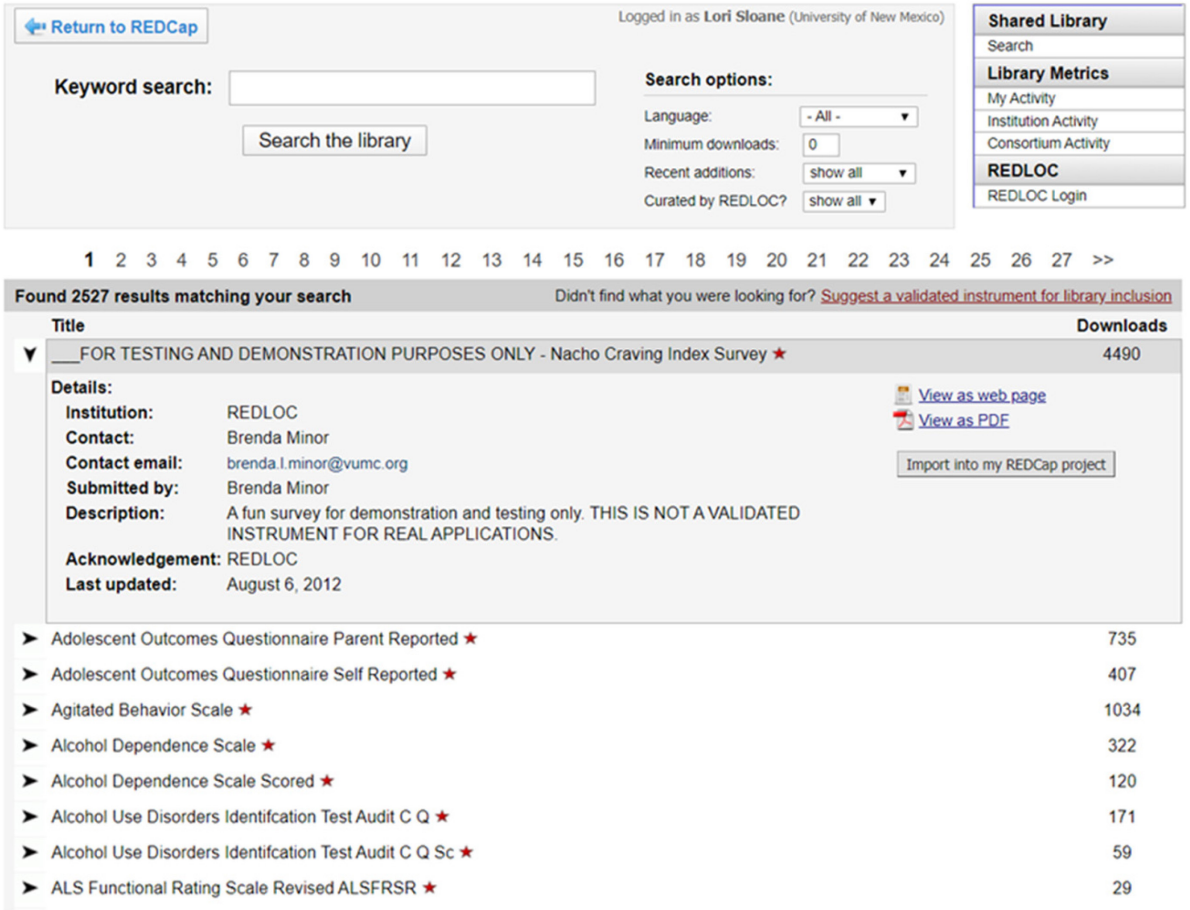


Medical coding dictionaries

Drug codes:



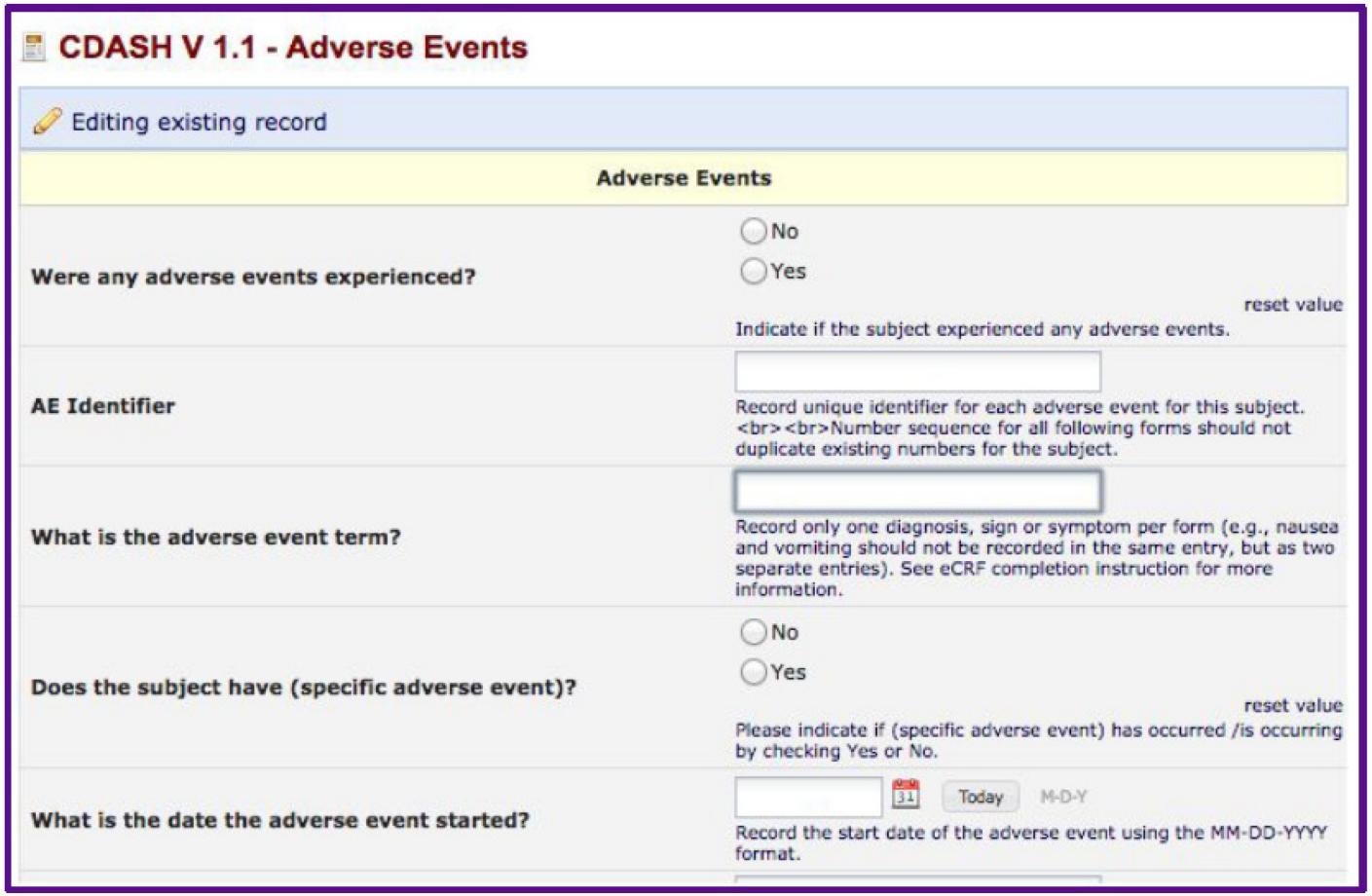
REDCap Shared Library



Data standards: CDASH

REDCap
Research Electronic Data Capture

Adverse Events:



CDASH V 1.1 - Adverse Events

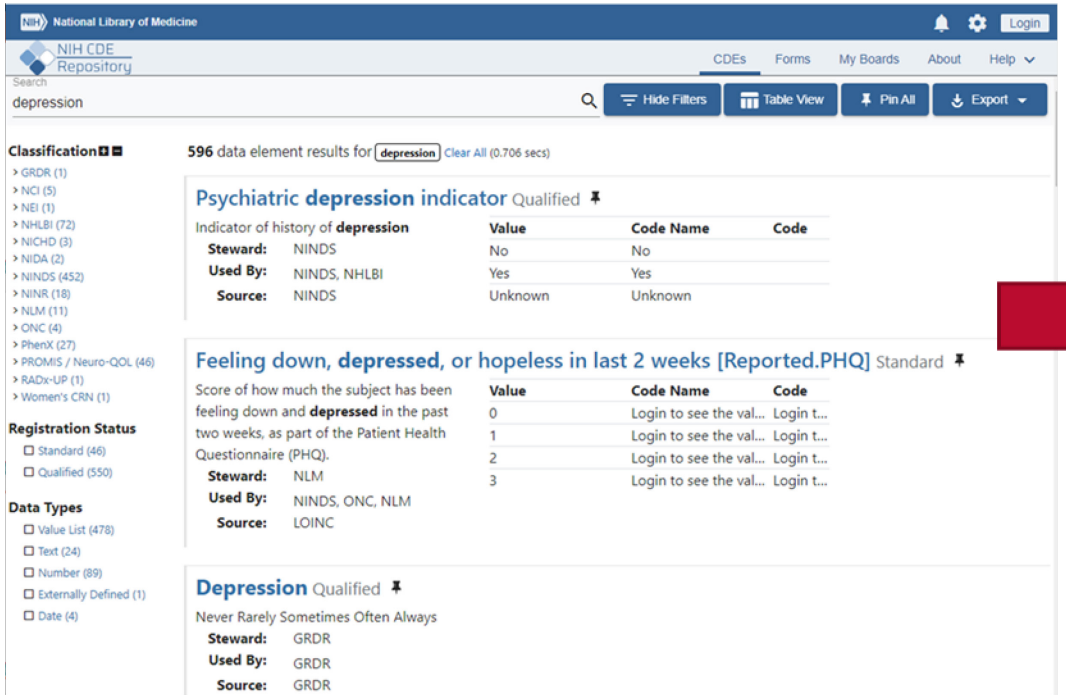
Editing existing record

Adverse Events

Were any adverse events experienced?	○ No ○ Yes reset value Indicate if the subject experienced any adverse events.
AE Identifier	Record unique identifier for each adverse event for this subject. Number sequence for all following forms should not duplicate existing numbers for the subject.
What is the adverse event term?	Record only one diagnosis, sign or symptom per form (e.g., nausea and vomiting should not be recorded in the same entry, but as two separate entries). See eCRF completion instruction for more information.
Does the subject have (specific adverse event)?	○ No ○ Yes reset value Please indicate if (specific adverse event) has occurred /is occurring by checking Yes or No.
What is the date the adverse event started?	Today M-D-Y Record the start date of the adverse event using the MM-DD-YYYY format.

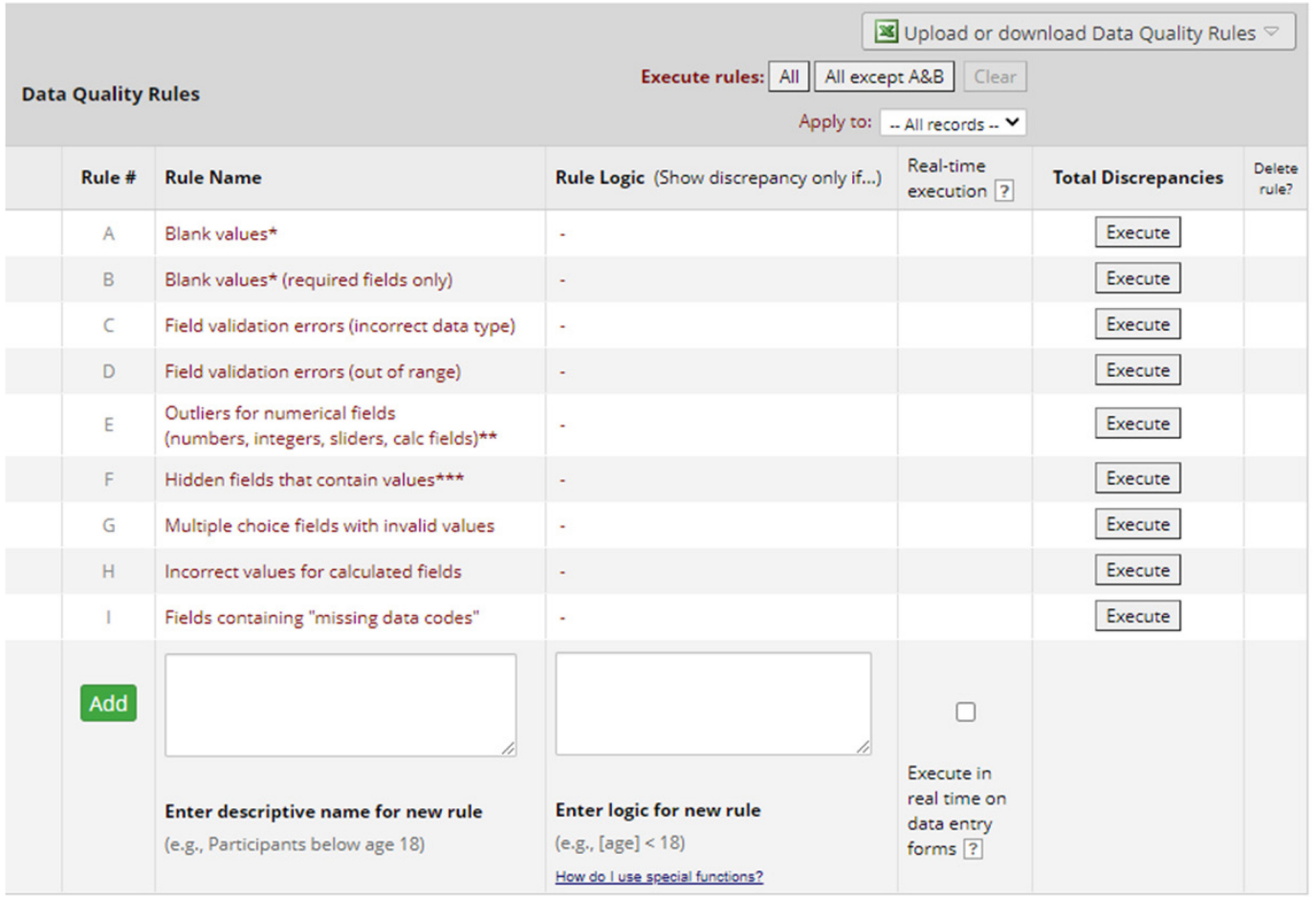
NIH CDE Repository

https://cde.nlm.nih.gov/home



	A	B	C	D
1	Name	Question Texts	Value Type	Permissible Values
2	Psychiatric depression indicator		Value List	No; Yes; Unknown
3	Feeling down, depressed, or hopeless in last 2 weeks [Reported.PHQ]	Feeling down, depressed, or hopeless?; Over the last 2 weeks how often have you been bothered by feeling down depressed or hopeless?	Value List	0; 1; 2; 3
4	Depression		Externally Defined	
5	Depression	Depression	Value List	1; 0
6	PROMIS - Emotional Distress depression depressed scale		Value List	5; 4; 3; 2; 1
7	Dysphoria - depression		Value List	1; 2
8	Depressed BFI	I see myself as someone who is depressed, blue.	Value List	1; 2; 3; 4; 5
9	Depressed 30D K6	During the past 30 days, about how often did you feel...so depressed that nothing could cheer you up?	Value List	1; 2; 3; 4; 5
10	Patient Health Questionnaire Depression (PHQ) - Depressed score	Over the last 2 weeks how often have you been bothered by feeling down depressed or hopeless?	Value List	0; 1; 2; 3
11	Geriatric Depression Scale (GDS) Long Form - Depression scale		Value List	Mild depressive; Normal; Severe depressive
12	Center for Epidemiologic Studies-Depression Scale (CES-D) - depressed indicator		Value List	0; 1; 2; 3
13	Hamilton Depression Rating Scale (HDRS) - depressed mood indicator		Value List	0; 1; 2; 3; 4
14	Quality of Life - Depression assessment past week scale		Value List	5; 4; 3; 2; 1
15	Psychiatric depression onset date		Date	

Data quality



Upload or download Data Quality Rules

Data Quality Rules

Execute rules: All | All except A&B | Clear

Apply to: -- All records --

	Rule #	Rule Name	Rule Logic (Show discrepancy only if...)	Real-time execution [?]	Total Discrepancies	Delete rule?
	A	Blank values*	-		Execute	
	B	Blank values* (required fields only)	-		Execute	
	C	Field validation errors (incorrect data type)	-		Execute	
	D	Field validation errors (out of range)	-		Execute	
	E	Outliers for numerical fields (numbers, integers, sliders, calc fields)**	-		Execute	
	F	Hidden fields that contain values***	-		Execute	
	G	Multiple choice fields with invalid values	-		Execute	
	H	Incorrect values for calculated fields	-		Execute	
	I	Fields containing "missing data codes"	-		Execute	
	Add	**Enter descriptive name for new rule** (e.g., Participants below age 18)	**Enter logic for new rule** (e.g., [age] < 18) How do I use special functions?	Execute in real time on data entry forms [?]		

Quality Assurance & Data Management Example

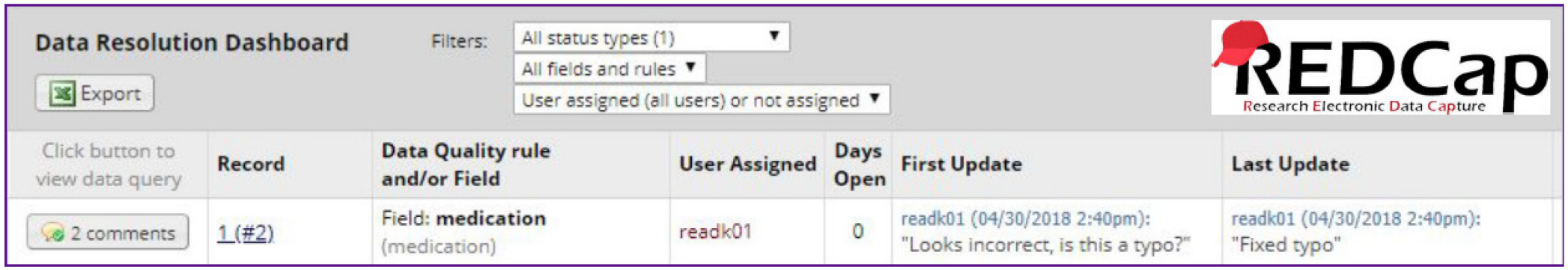
Data reconciliation responsibilities in REDCap:

Staff responsible:

- Research coordinator
- PI

If there is a discrepancy in data from the source documentation, follow this procedure:

1. Error flagged in REDCap
2. Review source documentation
3. Error (if it is not self-evident) to PI
   a) [a] correct error, [b] irresolvable
4. Discrepancy is approved. Decision is documented below by the PI:



Where is the safest place to store my data?

UNM HSC Network Drive (H, O or N drives, Sharepoint)

REDCap (production mode)

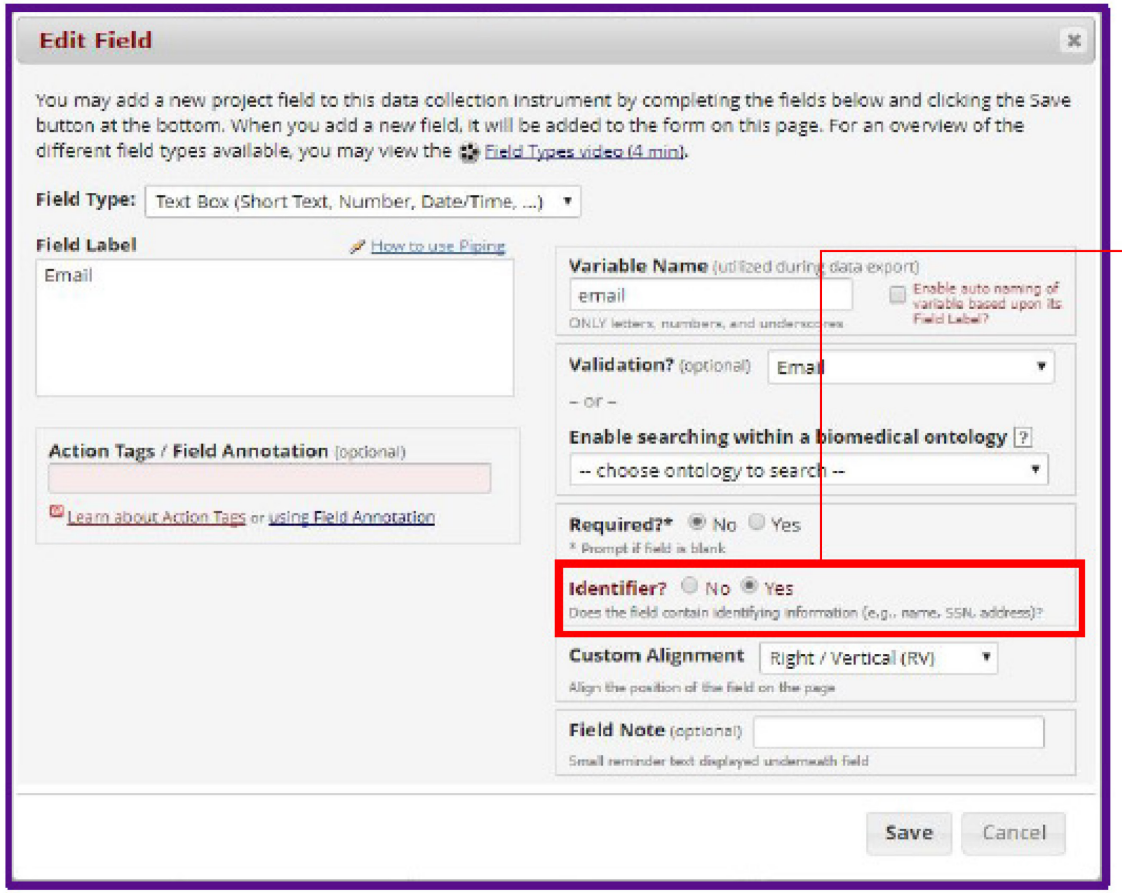
- https://hsc.unm.edu/ctsc/services/informatics/redcap-login.html

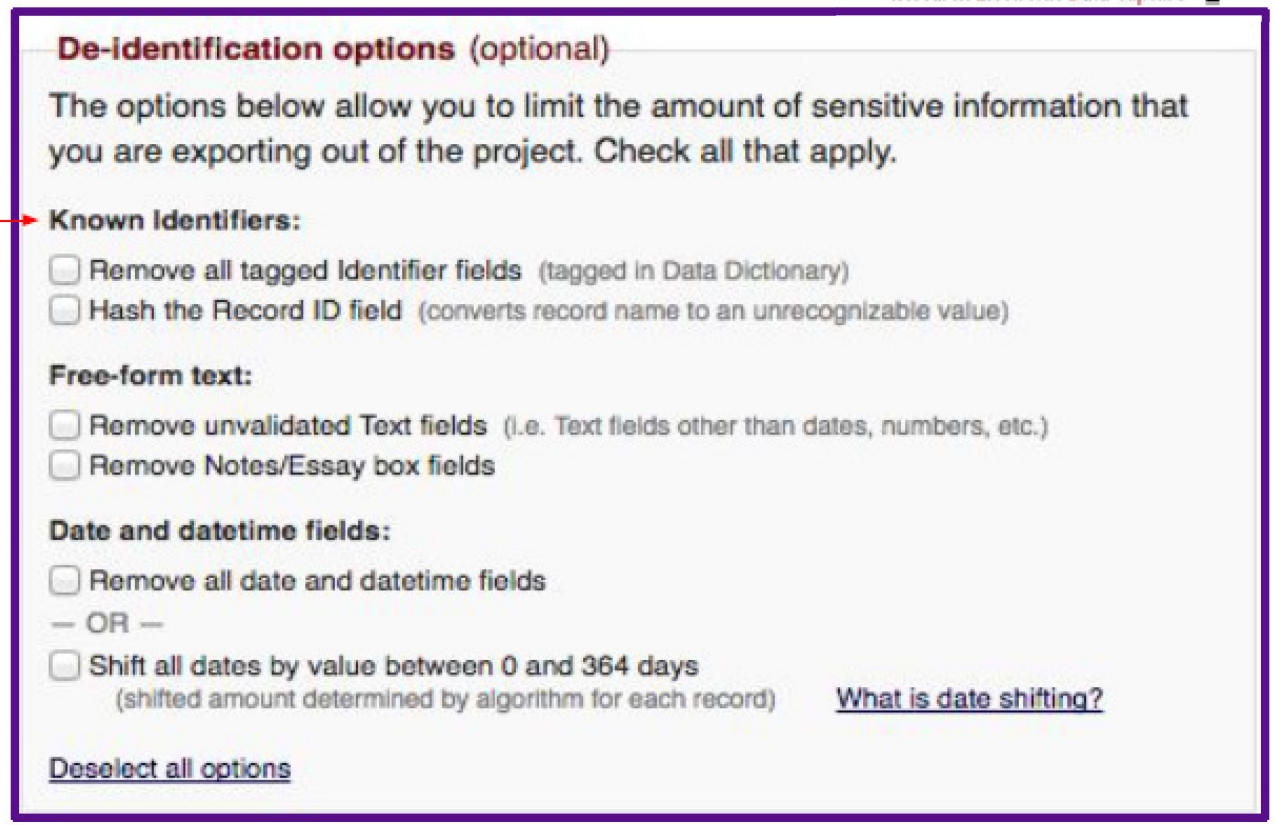
- UNM Main Campus High Performance Computing

- HSC-250 Systems and Network Security

Use open formats

Data privacy: REDCap





Data privacy: HIPAA Compliance

Logging

Download entire logging record to Microsoft Excel (CSV)

This module lists all changes made to this project, including data exports, data changes, and the creation or deletion of users.

Filter by event: All event types (excluding page views)
Filter by user name: All users
Filter by record: All records
Displaying events (by most recent): 1 - 100
Filter by time range from to

Time / Date	Username	Action	List of Data Changes OR Fields Exported
10/30/2017 6:54pm	kdg294	Updated Record 18	item_name = 'None found', data_collection_instrument_worksheet_complete = '2'
10/30/2017 6:54pm	kdg294	Updated Record 18	dataset_information_for_data_collection_instrument_complete = '2'
10/30/2017 6:53pm	kdg294	Updated Record 17	data_collection_instrument_worksheet_complete = '2'
10/30/2017 6:53pm	kdg294	Updated Record 17	item_name = 'None found', data_collection_instrument_worksheet_complete = '0'
10/30/2017 6:53pm	kdg294	Updated Record 17	dataset_information_for_data_collection_instrument_complete = '2'
10/30/2017 6:52pm	kdg294	Updated Record 16	item_name = 'None found', data_collection_instrument_worksheet_complete = '2'
10/30/2017 6:52pm	kdg294	Updated Record 16	dataset_information_for_data_collection_instrument_complete = '2'
10/30/2017 6:52pm	kdg294	Updated Record 15	item_name = 'None found', data_collection_instrument_worksheet_complete = '2'
10/30/2017 6:51pm	kdg294	Updated Record 15	dataset_information_for_data_collection_instrument_complete = '2'

REDCap
Research Electronic Data Capture

Data Storage Documentation Example

Data Storage Procedure and Workflow

Data Storage During Collection:

- Data stored within REDCap
- Lab data uploaded to REDCap File Repository with file name **LabName_DataType_YYYYMMDD**

Data Storage Before Analysis:

- REDCap data exported to H:/ drive **Project Folder** in CSV raw format with file name:
  - **ProjectName_MasterDataset_YYYYMMDD**
- **ProjectName_MasterDataset_YYYYMMDD** is cleaned and prepared for analysis
- Import into **SPSS v25** and saved to **Project Folder → Data Analysis Folder** with file name:
  - **ProjectName_DataSubset_YYYYMMDD**

Data Storage During Analysis:

- All changes to analysis dataset including output logs and monitor files will be stored in **Project Folder → Data Analysis Folder** with the file name:
  - **DataSubset_AnalysisType_FileType_YYYYMMDD**

Data management and storage

Document where your data lives throughout your study

Who is responsible for the data at different stages (e.g., collection vs. analysis)?

How safe is your data in each location? How is that safety guaranteed?

Ensure that your file systems are planned out ahead of time and understandable

Back up data and save in open formats if possible

Data transfer @ HSC

Requesting HSC Central IT Transfer:
The default DUA language requires that transfers be handled by HSC Central IT. Investigators can request secure data transfer services by creating a self-service help ticket and attaching the request number with their documentation confirming that HSC Central IT security transfer has been requested.

1. File transfers
    1. Submit a request through http://help.health.unm.edu and select -- “HSC Central IT secure transfer”
2. Data Entry into an External Web Application or Portal
    1. Submit a request through http://help.health.unm.edu and select – “Request Security Review”

For questions concerning IT Security issues please refer to http://hscsecurity.unm.edu/.

https://hsc.unm.edu/financialservices/preaward/ancillary-agreements/data-agreements/

Summary

Data management should be discussed **BEFORE** beginning a study and **any documentation** is better than no documentation

HEALTH SCIENCES LIBRARY
& INFORMATICS CENTER

February 15, 2022

MANAGING YOUR DATA

LORI SLOANE, HSLIC DATA MANAGER