Optimizing Research Data Planning/Capture/Reuse: How to Leverage Extant Tooling For Data Across the Research Lifecycle

Sungyoung Auh, James Welch, & Ken Wilkins

2026-08-20

Disclaimer: KW works across Intramural AND Extramural spaces, so may cite capabilities limited to one *or the* other*; please reach out to any named resource teams for the latest capabilities for* your *use*

Strategic research project planning, data collection, and data management: Prior Talk

Optimize Research Data Collection and Sharing

Research Project Management 101

Prior Learning Objectives: [Emily Leary & team]

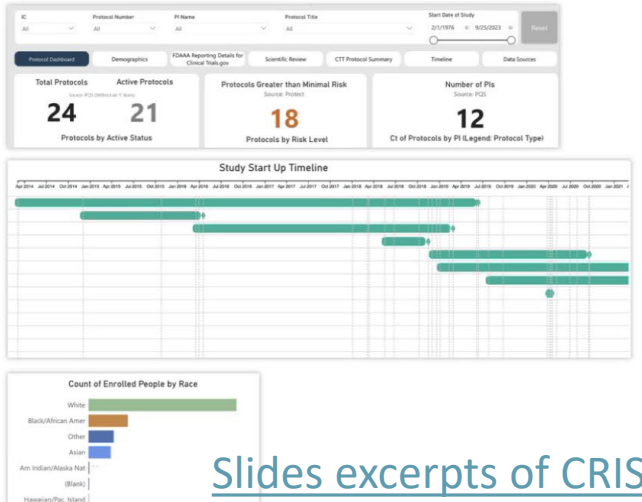
1. Understand research project management principles and how effective project management is critical for NIH-funded research
2. Accurately and completely outline the scope of work with respect to each research project component and research personnel
3. Create, implement, and adhere to a project communication plan for all relevant stakeholders
4. Create, implement, and adhere to a project organization plan and include needed documentation for all outlined scope of work

Need Project Management tools? **Key take-home *as we transition to next set of objectives: CRISPI can support you with it!***

NIH CRISPI Clinical Research Informatics Supporting Principal Investigators

Clinical Research Dashboard

- Key features
  - View & manage protocols by
    - Accrual status
    - Overdue quadrennial reviews
    - Risk level status
    - And more
  - View aggregate and PI-level spending (IC Directors only)
  - View protocol patient demographics
  - View study timelines and upcoming dates



CRISPI links for you!

CRISPI Clinical Research Informatics Supporting Principal Investigators

Clinical Research Dashboard

Goal: Create a "one stop shop" for PIs and IC Directors to manage and view active and historical study information.

Links to and pulls data from
- PROTECT
- ClinicalTrials.gov
- REDCap
- CRIMSON
- PQS
- And more...

Slides excerpts of CRISPI presentation Apr2026 to NIH Data Science Town Hall (+REDCap!)

Strategic research project planning, data collection, and data management

Optimize Research Data Collection & Sharing

The standard project-management life cycle, applied to research, from project start ("Development") to closure. Image from JHU All Children's Hospital.

https://rwdcollaborative.github.io/guide-to-n3c-v1/chapters/practices.html

Leveraging Extant Tools For Data Collection

Initial Learning Objectives:

[applicable across research lifecycle]

1. What can medical terminologies (& *other semantic codings*) do for you in data collection (*from capture to curation*)?

2. How to relate to integrate data across studies/sources (added value) .

3. How to choose a data repository

What are some of your dreams for data capture?

Would you* NOT *want:

- *to collect variables used by others in your field of study?*
- *to "start from scratch" when designing your data collection forms?*
- *to get patient/participant-generated data* known *to be* resilient *to peer critique?*
- *to "head off" questions from peer reviewers about specialized procedures/assays/methods?*
- *to minimize time 'transcribing' info from a source electronic system into study database(s)?*
- *to glean research-question info from free-text clinical notes in NIH's Clinical Research Informatics System or CRIS?*

Before we fully get into things…

some shout-outs to start your pursuit of such dreams

Leverage *Extant* Tools – & see *OTHER* SPACES TO WATCH FOR SOON-TO-BE-AVAILABLE TOOLING* -- *team up with informaticists!*

See the *RE-branded* **CRISPI**

CRISPI — Clinical Research Informatics Supporting Principal Investigators

***Impressive** array of capabilities, from extracting 'signal' from clinical notes, to automated integration of multi-source data*

CRIS/BTRIS use of Medical Terminologies: *in-house* workflow (UFOs)

User Facing Ontologies (UFOs)

AND capacity for HealthLevel7 Fast Healthcare Interoperability Resource (FHIR) exports those under CDIS of REDCap

Clinical Data Interoperability Services

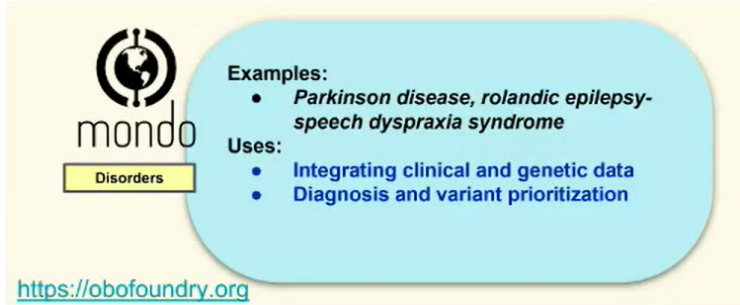
allows your NIH Clinical Center protocols to obtain (with participant consent) external health records!

NIH Cloud Platform Interoperability program – wealth of re-usable tooling publicly accessible via GitHub

NCPI

National Library of Medicine's **Unified Medical Language System** will include (in Nov'26 release) condition/disease ontology

Linked-data Markup Language, **LinkML** will make using interoperability standards for your data *easier*



More to look out for, ahead of fully getting into things…

Leveraging Extant Tools – see *OTHER* SPACES TO WATCH FOR SOON-TO-BE-AVAILABLE TOOLING* -- team up *with informaticists!*

Would your research participants prefer to fill out structured surveys on their own devices?

- Make this readily possible by leveraging **MyREDCap / REDCap mobile app** for ***direct*** data capture!

Could your data be augmented by participants' full health context via their *own* external data records?

- ***Inquire with NIH informatics resource teams on consent-predicated means to leverage extant NIH-funded tools listed under link below***

Other Office of Data Science Strategy (ODSS) funded* efforts

National Institutes of Health
Office of Data Science Strategy

Now, let's get into things…

* *ODSS leads implementation of the NIH Strategic Plan for Data Science 2025-2030*

CADRs=*controlled-access data repositories*

Why should you care about medical terminologies/standards?

Would you* NOT *want:

peer reviewers to* readily appreciate *variables used to yield findings in your study?

to shorten the timeframe when designing/revising research project data collection?

to ease how independent safety-monitoring experts aggregate 'safety signals'?

to smooth the process when submitting your Data Mgmt & Sharing plan(s) for review?

to explore additional research-questions by re-investing in prior-collected data?

be ready to meet any unanticipated request by others, per your DMS plan on file?

Use items mentioned in a talk, & search data artifacts in ways reminiscent of 'eminence-based' medicine?

OR?

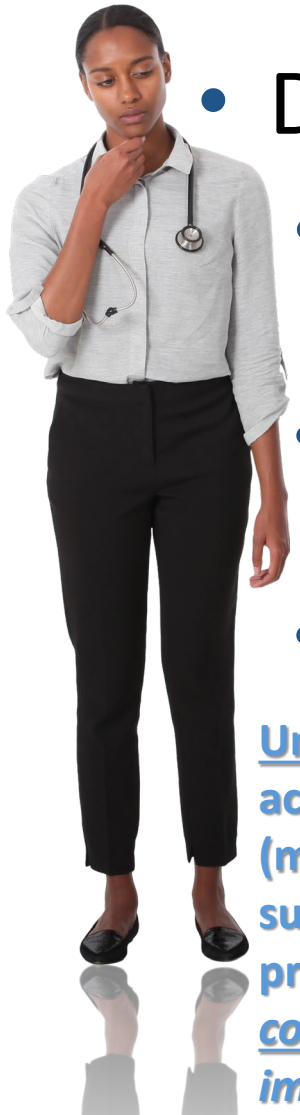
Operate accelerated timelines within a modern, AI-catalyzed set of digital research workflows, lending to *evidence- based* medicine?

https://www.dreamstime.com/old-big-library-lots-books-generative-ai-atmospheric-image279699332

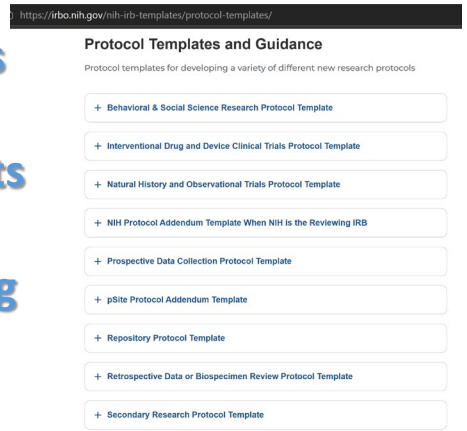
https://www.freepik.com/premium-ai-image/futuristic-library-with-woman-holographic-books_134018639.htm

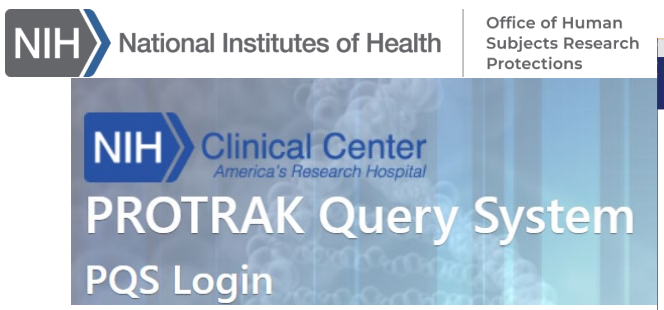
What can medical terminologies (MTs) do for you in data collection/capture? *Give you a uniform vocabulary*

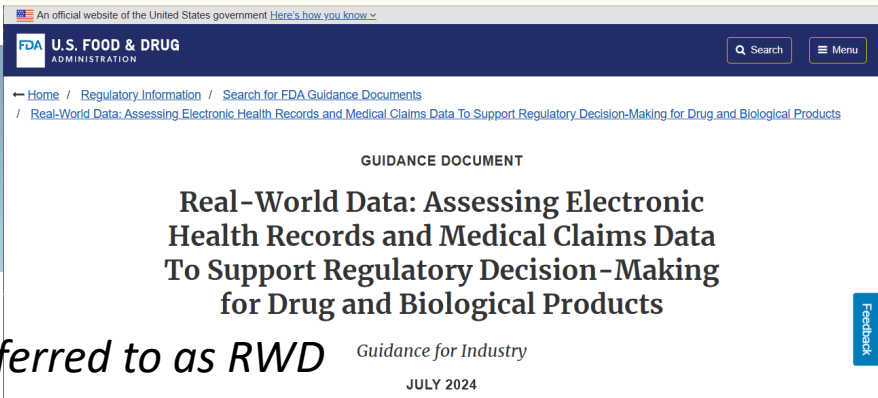
- *Increases the potential to* **re***use data for more research!*
- Data Curation spans the research lifecycle so MTs help in:
  - More rapid conception, informs planning/execution
    - *Quicker ways to* iterate *on your research* portfolio*, report* **safety** *data*
  - Disseminating study findings, Preparing data for **re**-use
    - *Increased potential for* recognized *findings, and measurable* impact
  - Integrating data across distinct studies, *with* Real World Data* sources



Uniform terms across your (meta)data sets support your projects having *consistent impact*







* *hereafter, referred to as RWD*

What can medical terminologies do for you in data collection (from *capture* through to *curation*)?

- Data Capture/Curation spans entire research lifecycle
  - Study conception informing planning/execution
  - Disseminating study findings, Preparing data for re-use
  - Integrating data across distinct studies, RWD sources

- Medical Terminologies' semantics underly data in *each* context:
  - Clinical settings (from generic to condition/disease specific)
  - Biomedical Research Esoterica (e.g., cutting-edge assays/methods)
  - Human contexts (socio-behavioral, SDoH) are ***not*** overlooked

OBO Foundry
Open Biological & Biomedical Ontology Foundry

NIH-wide Social Determinants of Health Research Coordinating Committee

gravity PROJECT

What can medical terminologies do for you in data collection (from *capture* through to *curation*)?

- Medical Terminologies' semantics underly data in *each* context:
  - Clinical settings (from generic to condition/disease specific)
  - Biomedical Research Esoterica (e.g., cutting-edge assays/methods)
  - Human contexts (socio-behavioral) are ***not*** overlooked

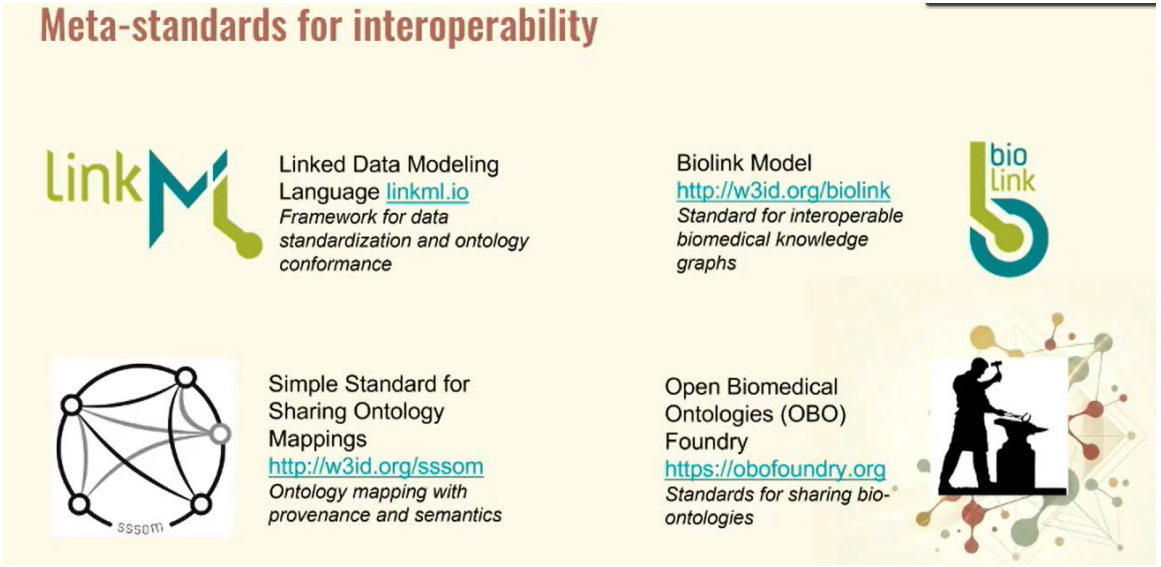
Thus, can *leverage* MT semantics to have data 'born interoperable'

- Existing *mappability* for (re-)use of data over time, for varied purposes
- ***BONUS***: ties to Data Mgmt &Sharing (DMS) policy practices! (*next* talk)

NIH

Data Management and Sharing Plan

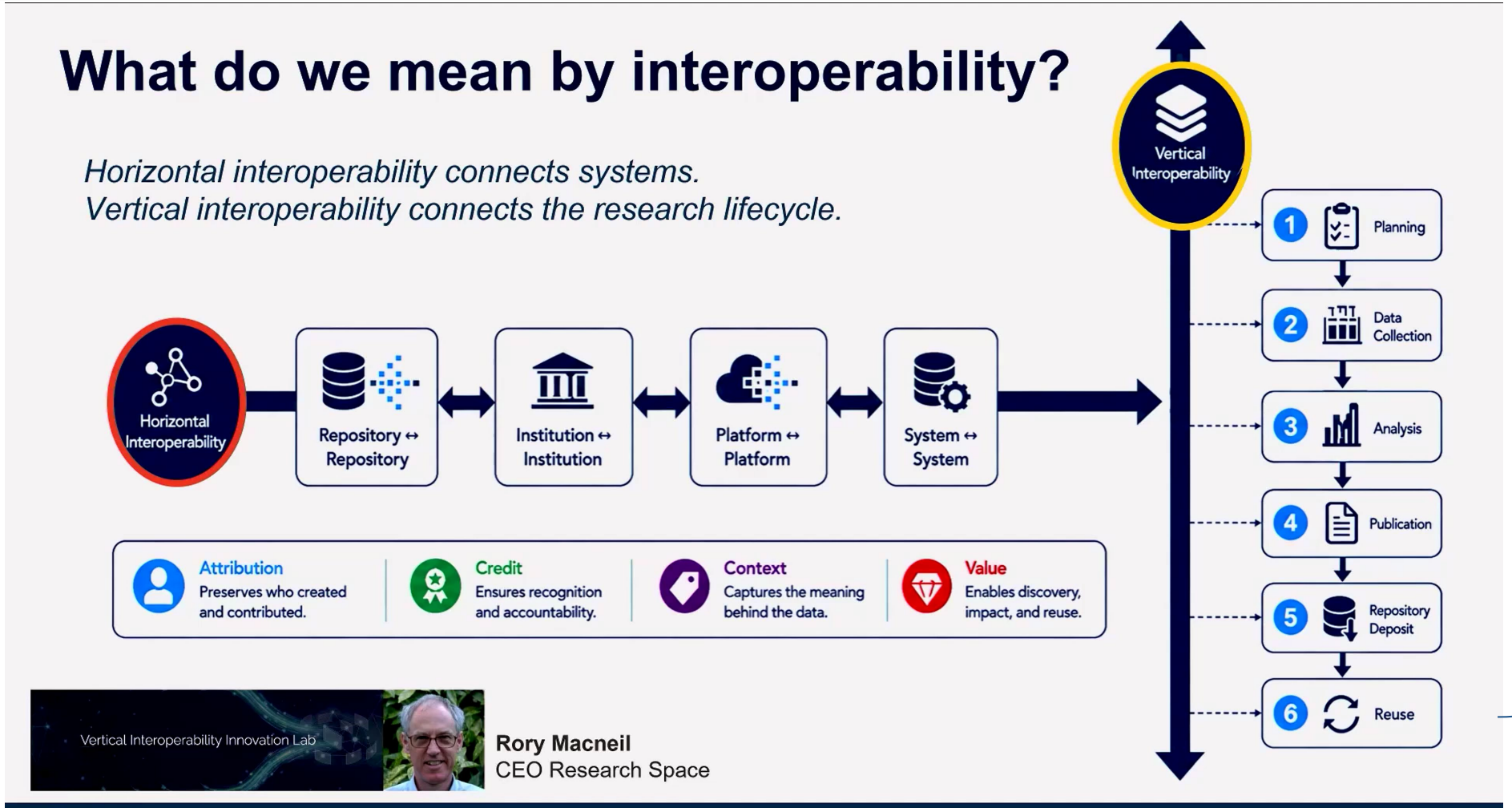
= **"Data Sharing Plan"** in next talk
(avoids confusion of 'DMS' plan,
let alone Data & Safety Monitoring
Plans or 'DSMPs'😳)



What can medical terminologies do for you in data collection (*AND ALSO IN RESEARCH PROJECT MANAGEMENT*)?

Medical Terminologies' semantics CAN ALSO underly **METAdata** in *all* project contexts: **VERTICAL INTEROPERABILITY**

[TIES RIGHT INTO OUR LAST WEBINAR ON RESEARCH PROJECT MANAGEMENT 101]



NIH National Library of Medicine

From National Library of Medicine's "Foundations of Interoperability: Data, Infrastructure, and Culture" - see recordings Day 1 & Day 2 under the NIH Videocast's 'Past Events'

How to relate a given study's data to broader data integrated across studies/sources (added value)

- Data Integration "A-B-C's"
  - A: *access* (meta)data across source types, leverage med.termin.(**MTs**)
    - MTs anchor semantics of Common Data Elements, reference points of data ***re***use
    - MTs anchor 'concept sets' behind Common Data Models for EHRs, claims, etc.
    - MTs are *actively* curated to *expand* when *new* frontiers in know*quire* it
      - *Who* better than leading-edge science at NIH to inform MTs' growth & reuse?

Common Data Elements & other approaches to data interoperability & ***re***use, is an *ongoing* theme today

How to relate a given study's data to broader data integrated across studies/sources (added value)

- Data Integration
  - A: *access* (meta)data across source types, leverage med.termin.(MT's)
    - MTs anchor the semantics of CDEs, which serve as reference point for data
    - MTs anchor 'concept sets' behind Common Data Models for EHRs, claims, etc.
    - MTs are actively curated to *expand* when new frontiers in knowledge require it
  - B: *broaden* the interdisciplinary team bench
    - Leverage **informaticists** most familiar with a source (e.g., BTRIS team, CRISPI)
      - IRP branches with ongoing BTRIS collaborations have netted bespoke re-usable workflows
      - Specific specialties in CC have helped inform CRISPI-led innovations—add *your* team to that
    - Complementary experts on generating *original* data *AND* analyzing *final* data
      - Any specialized assays, instruments, patient-generated data (PD)? *Involve experts in your team*
      - Do you collect repeated measures or multiple/longitudinal outcomes? *Consult methods expert!*
        - *Such expertise helps you* set up *data in ways that minimize error (later talk)*

How to relate a given study's data to broader data integrated across studies/sources (added value)

- Data Integration
  - A: *access* (meta)data across source types, leverage med.termin.(MT's)
    - MTs anchor the semantics of CDEs, which serve as reference point for data
    - MTs anchor 'concept sets' behind Common Data Models for EHRs, claims, etc.
    - MTs are actively curated to *expand* when new frontiers in knowledge require it
  - B: *broaden* the interdisciplinary team bench
    - Leverage informaticists most familiar with a source (e.g., BTRIS team, CRISPI)
    - Complementary experts on generating *original* data *AND* analyzing *final* data

- C: *curate* the data to best suit analytic approaches ↑ to integrated data
  - Frequent use case for integrating data in an ongoing study: *safety monitoring… aggregate 'signal'*
    - *Motivating context for 'Common Data Element' use:* one pairs a given **prompt** ("Any safety-related events?") with unambiguous **response** to address research question at hand (e.g., accruing risk:benefit tradeoffs)
    - Independent safety monitoring experts/panels may consider *which demographic* exhibit certain events
    - Certain safety events get *graded* on degree of impact to participants (severity); example **MT** system is **CTCAE***
    - USE CASE: MT-based-curation of data labs' normal range by source, adverse events, ancestry
      - Adverse kidney functioning associated with APOL1 genotypes (West African ancestry)

** Common Terminology Criteria for Adverse Events, e.g., version 5 – crucial for AE reporting*

Common critique of *Common* Data Elements (CDEs): *SOME ARE WAY* TOO *COMMON*

ALWAYS > *1 way to pair a question (AGE) with set(s) of answers: #s*

Safety considerations vary by age

Units can ***vary*** by specific ***context of application:*** months for infants

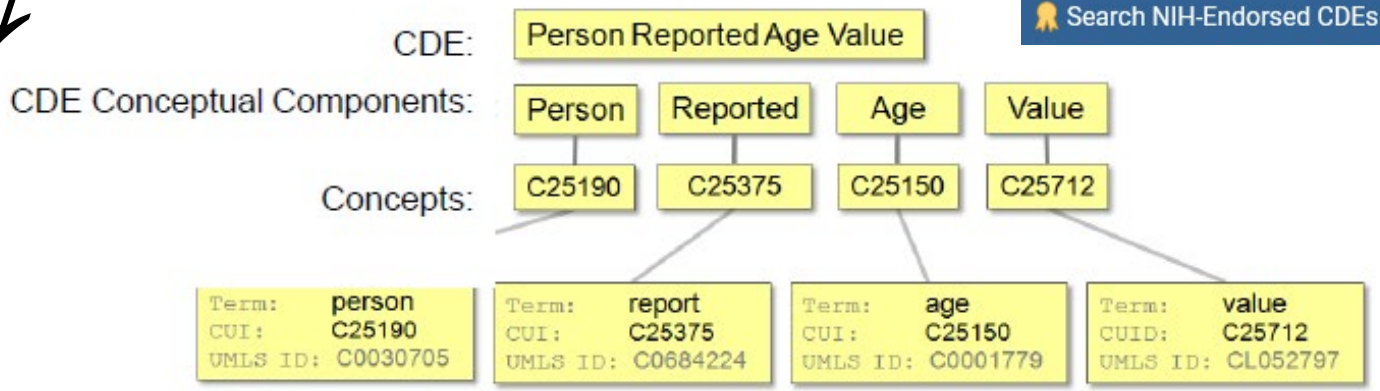
<u>Best way to select among such 'common' options?</u> Look for *in*

NIH-endorsed *meet* key criteria ↓

so have semantic underpinnings

National Library of Medicine
NIH CDE Repository

Search NIH-Endorsed CDEs



Term: the word in the terminology (human readable and understanding what is meant)

CUI: in this case →NCI Thesaurus Concept Codes

UMLS CUI: the Unified Medical Language System (UMLS) Concept Unique Identifier

NIH National Institutes of Health
Endorsed CDEs Accelerate Research

- Clear definition of variable, measure with prompt and response
- Documented evidence of reliability and validity
- Human- and machine-readable format preferred
- Recommended/designated by a recognized NIH body (Institute/Center, Committee, etc)
- Licensing and intellectual property status clear (prefer Creative Commons-by or open source)

Case study on the need for *Common* Data Elements 'CDEs'

There is *>1 way to pair a question with set(s) of answers:* 'race'⇔APOL1

From Nat'l Health & Nutrition Examination Survey (NHANES) **v3 c.1990s**

Hispanic Origin (Y or N):
Race Code: Race Codes: 0=White 1=Black 2-Asian/Pacific 3=Am Indian/Eskimo 4=Other

Need to decode the combinations
"Yx0" "Yx1" "Yx2" "Yx3" "Yx4"
"Nx0" "Nx1" "Nx2" "Nx3" "Nx4"

…for "_x4" need to know what if 'Other' is 'Multiple races' and whether that includes Black

to US Race/Ethnicity by 2029Q2

Per public comment & expert panels:

"respondents should select all categories that apply to them"

In this case, for compliance by 2029, Fed'l Gov't to move toward the way at right->

Figure 3. Race and Ethnicity Question with Minimum Categories Only

https://spd15revision.gov/

What is your race and/or ethnicity?
Select all that apply.

- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Hispanic or Latino
- ☐ Middle Eastern or North African
- ☐ Native Hawaiian or Pacific Islander
- ☐ White

ClinicalTrials.gov
A service of the U.S. National Institutes of Health

Find Studies | About Clinical Studies

Terminology Code Mapping

	NCI Thesaurus	→ UMLS CUI	→ LOINC
American Indian or Alaska Native	C41259	C0282204	LA10608-0
Asian or Asian American	C41260	C0003988	LA6156-9
Black of African American	C16352	C0085756	LA10610-6
Hispanic, Latino, or Spanish	C17459	C0086409	LA6214-6
Native Hawaiian or Other Pacific Islander	C41219	C1513907	LA10611-4
Middle Eastern or North African	C43866	C1553353	No Match
White	C41261	C0043157	LA4457-3

U.S. Office of Management and Budget's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity

Note how answers are *mappable* across MTs!

Case study on the need for *Common* Data Elements 'CDEs'

There is > *1 way to pair a question with set(s) of answers:*

Adverse Events (AEs) re: kidney disease-can ***vary*** by specific ***context***

In non-kidney-disease diagnosed, urine protein excursions may lead to AEs, or *if reporting* race as Black or African-American may look to genotype for APOL1 for more thorough risk assessment; forward-looking studies may use **MT**-anchored CDE with **African heritage**…

Western African origin being a starting *proxy for risk assessment*

Race/Ethnicity Self-Identification

A textual description of a person's **race**. C17049 |The ethnicity of a person.

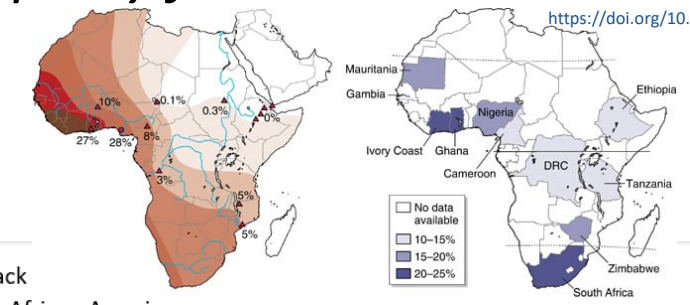
Qualified

Steward: NIMHD
Used By: NIMHD
Sources: ScHARe working group preference based on potential classifications in 2030 census (see References)

Permissible Values

Label	Code	ConceptID
American Indian…	Login to see the…	C41259
Asian or Asian A…	Login to see the…	C41260
Black or African …	Login to see the…	C16352
Hispanic, Latino…	Login to see the…	C17459

(7 total) See full table in Detail View



☐ Black
- ○ African American
- ○ Jamaican
- ○ Haitian
- ○ Nigerian
- ○ Ethiopian
- ○ Somali
- ○ [write in] Enter, for example, Trinidadian and Tobagonian, Ghanaian, Congolese,

Case study on the need for *Common* Data Elements 'CDEs'

There is *> 1 way to pair a question with set(s) of answers:*

Adverse Events (AEs) re: kidney disease-can ***vary*** by specific ***context***

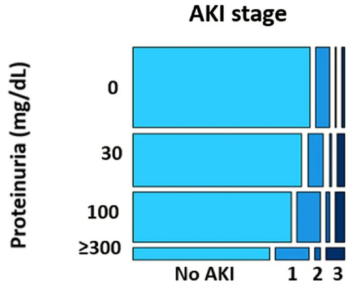
Safety monitoring may screen for unusual lab values, say, *albumin in urine ⇔ acute kidney injury*

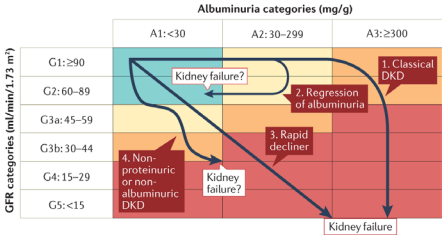
May be *graded for severity per consensus criteria* as well as anchored in **MTs**:

According to Kidney Disease Improving Global Outcomes (KDIGO) guidelines, albuminuria is categorized into 3 grades with increasing severity of associated kidney injury: A1, A2, and A3. This grading scale, along with estimated glomerular filtration rate, is used to assess risk of progression and severity of kidney disease, particularly in chronic kidney disease.[2]

KDIGO Albuminuria Categories Based on Urine ACR

Albuminuria Category	ACR (mg/mmol)	ACR (mg/g)
A1 (normal to mildly increased)	<3	<30
A2 (moderately increased)	3–30	30–300
A3 (severely increased)	>30	>300





https://link.springer.com/article/10.1186/s12882-022-02677-y/figures/1 | https://www.nature.com/articles/s41581-021-00462-y/figures/1

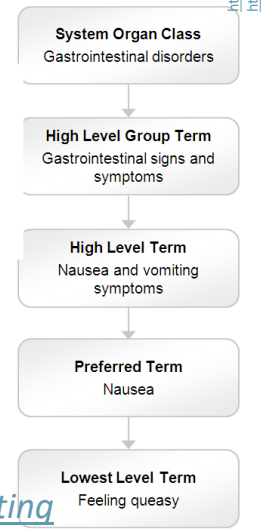
In *one* kidney test article's trial, certain AEs were 'expected' (could solicit comorbidities in a structured 'participant diary')

Expected adverse events are abdominal bloating and diarrhea;
if any subject experiences a severe level of these adverse events that is not explained by a co-existing condition (e.g. acute gastroenteritis), we may reduce the second dose level

Another case for MTs: *AEs' safety 'signal' best aggregated by MedDRA hierarchy,*
regardless of 'which way' they get recorded/captured, for *monitoring*
As well as mapping to other MTs for risk-based monitoring!

MedDRA
Medical Dictionary for Regulatory Activities

CTCAEv4 – already LLTs for AE reporting

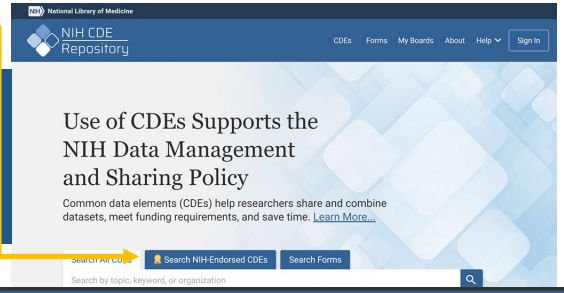


CDEs *ease* Data Capture Plans: elicit *clear* data SOPs!

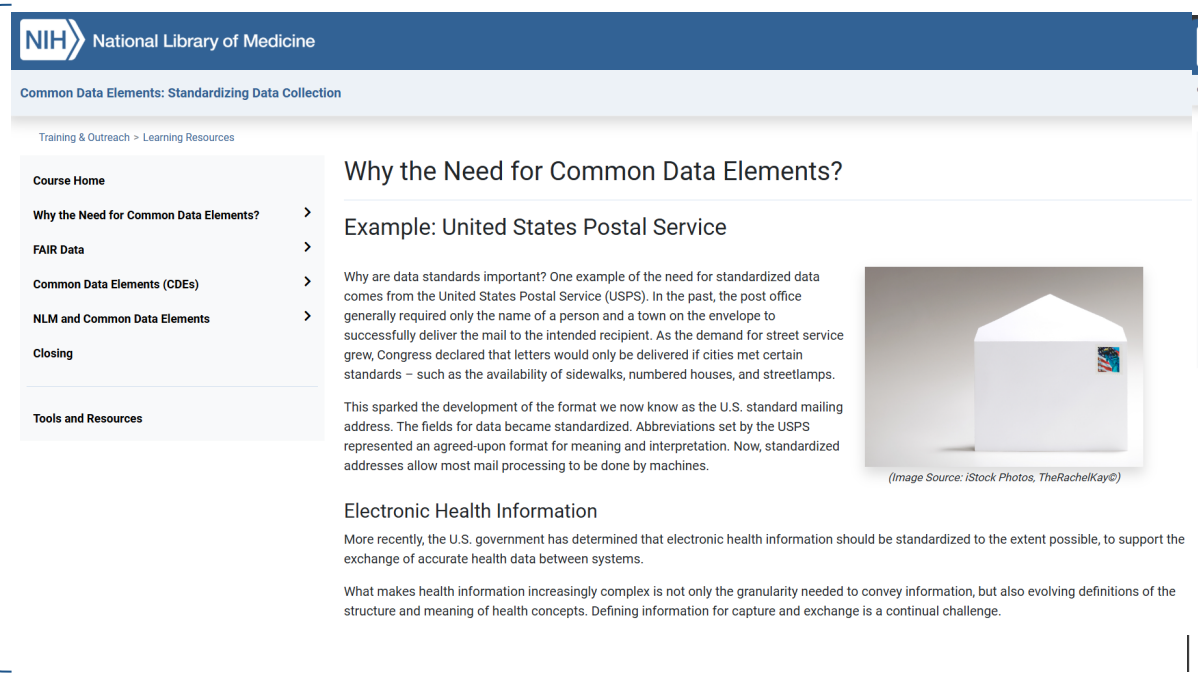
- Leveraging NIH CDE Repository *public* & *log-in* features in order to *future-proof data* – *Natural way to leverage **MTs** to set of Case Report Forms (CRFs) useful for yr study*
  - *CRF generation can MOST READILY be done by 'template' forms made of CDEs!*
  - *IDEALLY, use NIH-endorsed **Common Data Elements***

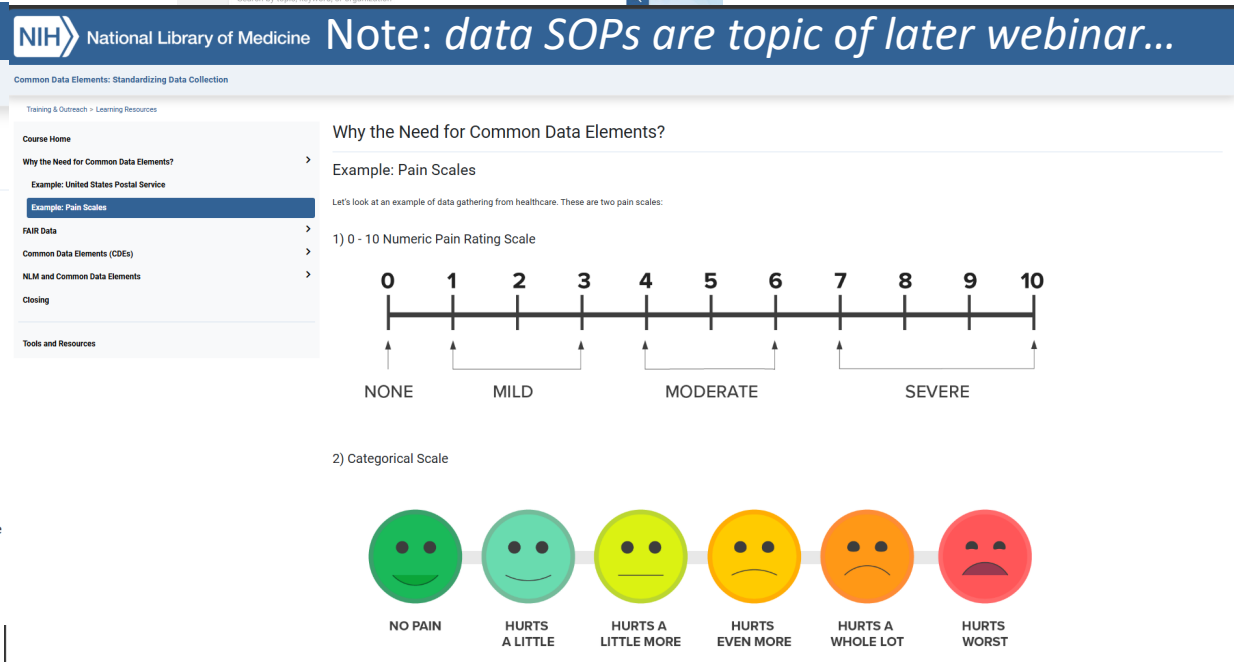
plan → code → data

NIH CDE Repository uses the Unified Medical Language System (UMLS) Terminology Service (UTS) Sign on Service which lets you set up an account and sign in using your NIH credentials, your account with a research organization, or a personal account such as Google, Microsoft, or Login.gov. User account holders can create Boards and save CDEs and Forms to them, remember your preferences on all your devices



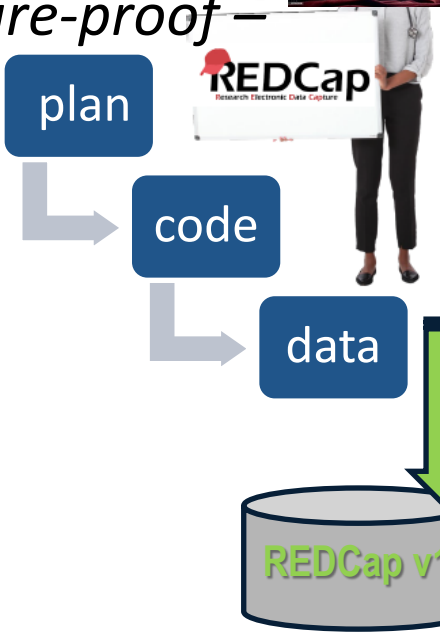
Plenty of NLM training materials on Common Data Elements





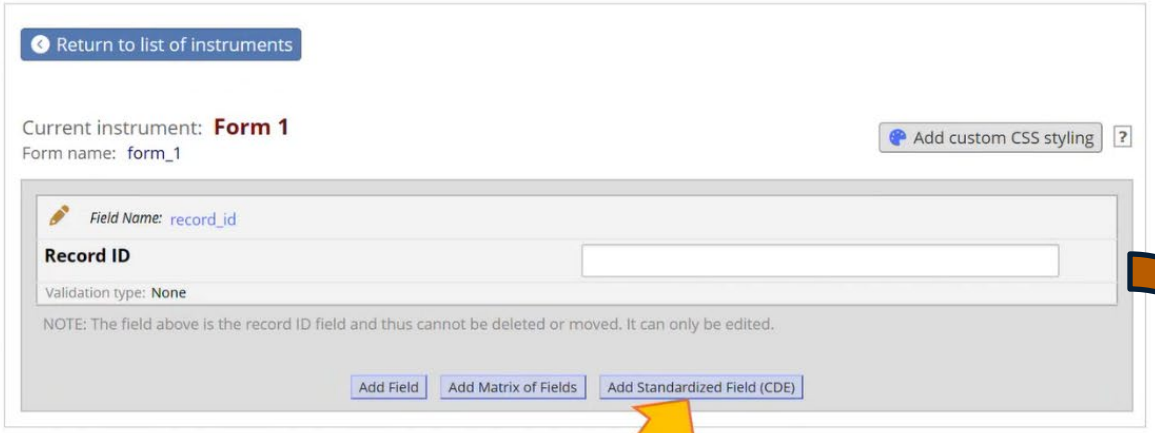
REDCap: tooled* for CDE RE-use

- Leveraging NIH CDE Repo Application Programming Interface (API): *future-proof – Natural way to leverage **MTs** via 'semantic search' e.g., HIV-based CDEs*



CDEs in REDCap (quick walkthrough)

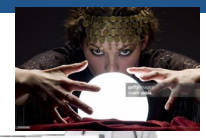
- When building a data collection form in REDCap, you can import CDEs.
- Click "Add Standardized Field (CDE)" in the Online Designer (formerly "Import from Field Bank")



* NOT a REDCap User? Please note ALL HHS-/NIH-accessible genAI tools can help too; see Digging Deeper into Common Data Elements at NIH by Using Generative AI | ODSS

REDCap: tooled for CDE RE-use

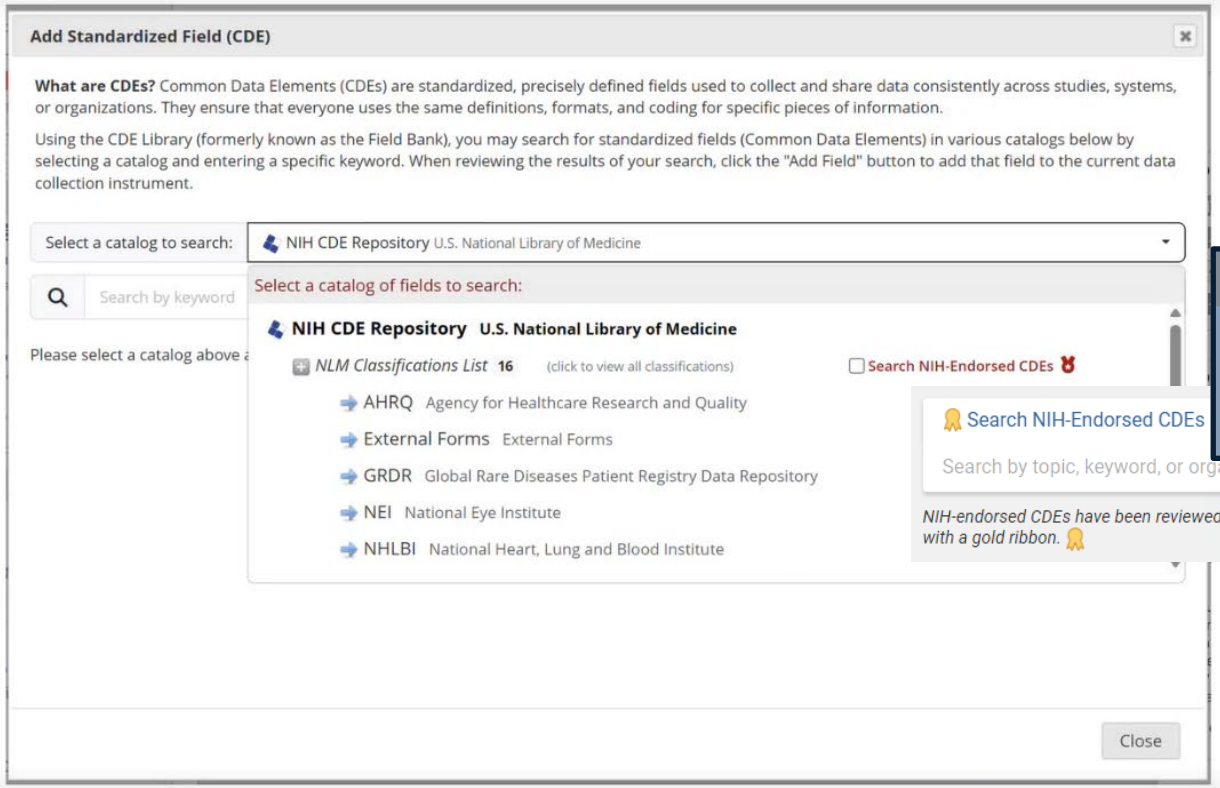
- Leveraging NIH CDE Repo Application Programming Interface (API): *future-proof – Natural way to leverage **MTs** via 'semantic search' e.g., HIV status*

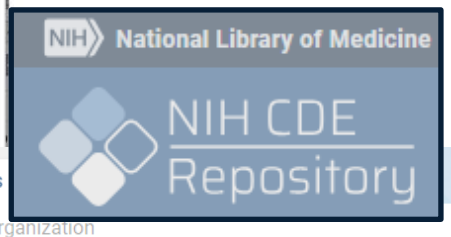


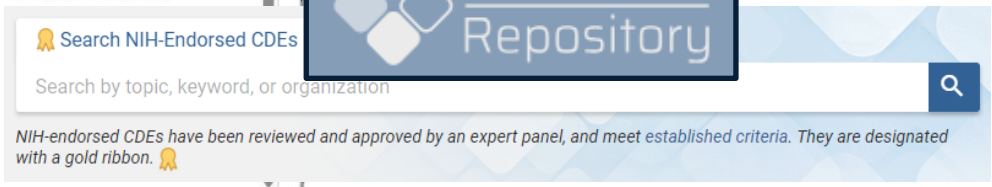
plan → code → data

REDCap Screenshot (2)

In the box that pops up, you can search against the NIH CDE Repository for CDEs and import them into your forms.

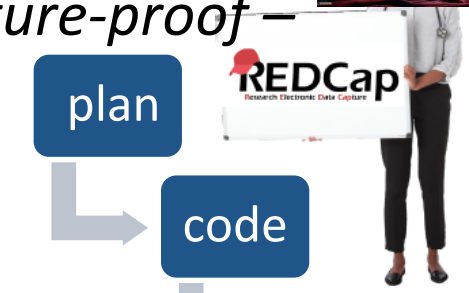






REDCap: tooled for CDE RE-use

- Leveraging NIH CDE Repo Application Programming Interface (API): *future-proof – Natural way to leverage **MTs** via 'semantic search' e.g., HIV status*

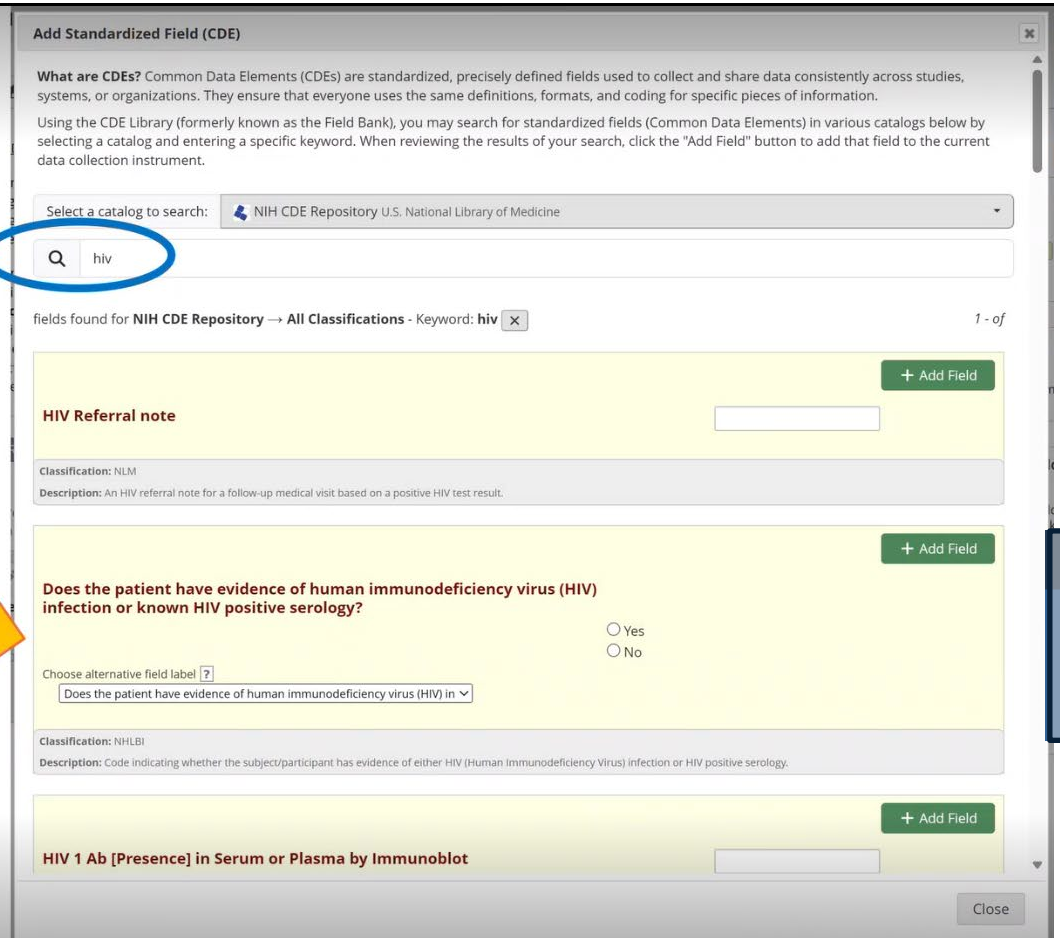


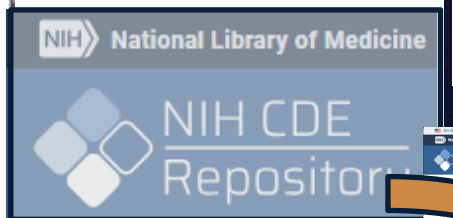
REDCap Screenshot (3)

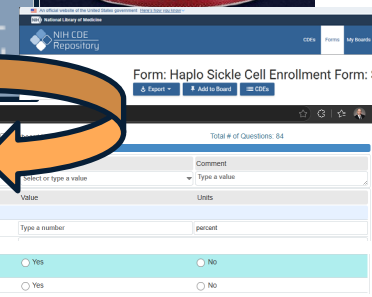
In this example, I typed "HIV" in the search box.

This is the same yes/no CDE that we looked at earlier.

The +Add Field button will insert this question into your REDCap form.

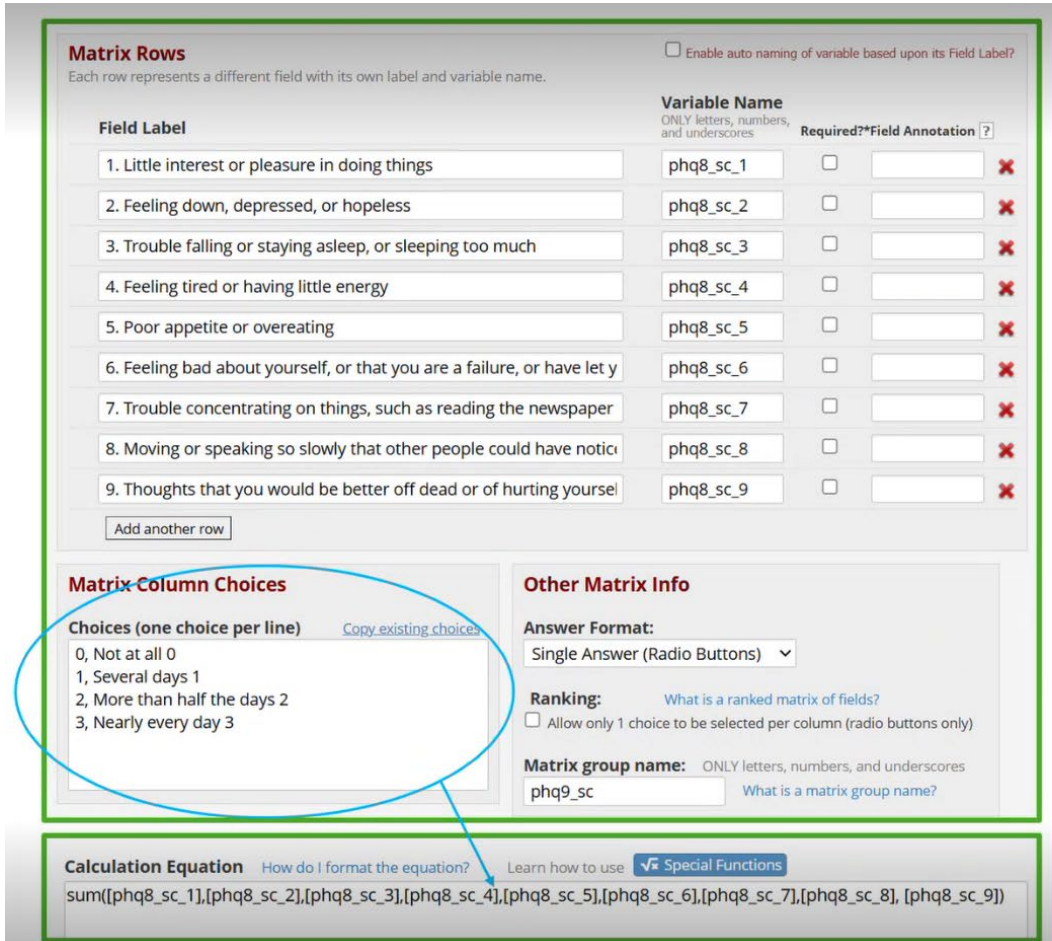






REDCap: tooled for CDE RE-use

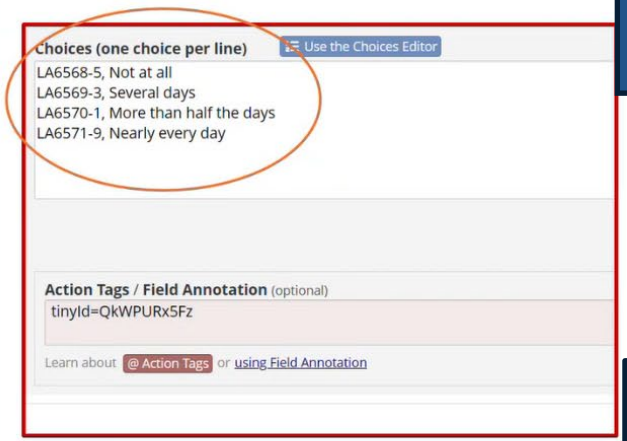
- Leveraging NIH CDE Repo Application Programming Interface (API): *future-proof – Natural way to leverage **MTs** via 'semantic search' e.g., **frequency***



The option you choose determines whether you can easily write a calculation for the PHQ-9 total score.

It's easy to sum up integers, more difficult to sum LOINC codes.

The 4 options also use different naming conventions.

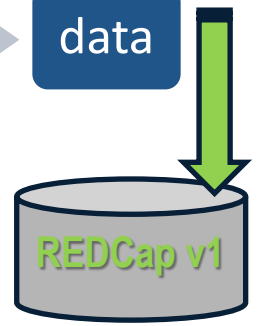


plan → code → data

LOINC® from Regenstrief

CONNECTED BY *NEW* LOINC ONTOLOGY

SNOMED International — Delivering SNOMED CT

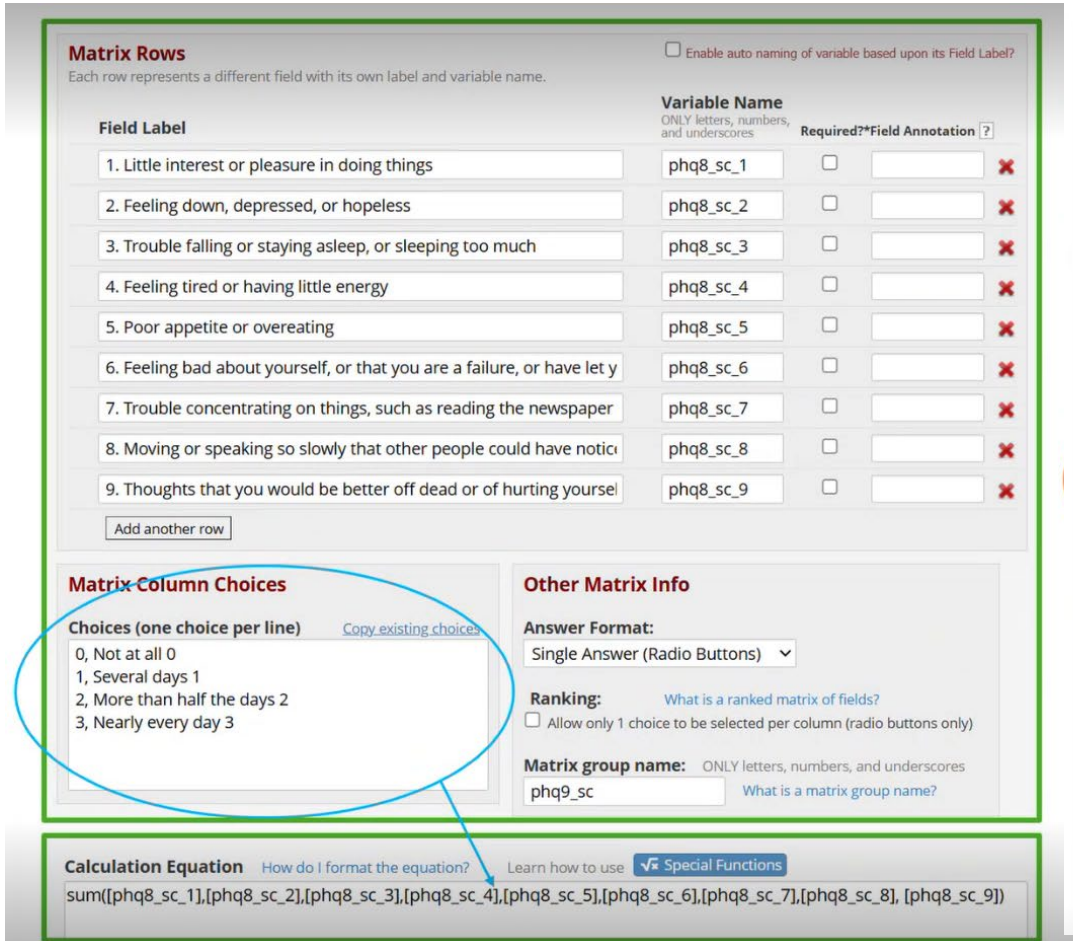


NIH National Library of Medicine

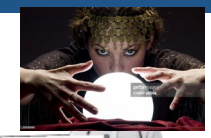
NIH CDE Repository

REDCap: tooled for CDE RE-use

- Leveraging NIH CDE Repo Application Programming Interface (API): *future-proof – Your Case Report Forms arise from 'semantic search'* ***RESULTING ELEMENTS***



To be covered in a talk later in webinar series



plan → code → data → REDCap v1

NIH National Library of Medicine
NIH CDE Repository

Extant tooling for Data: others *key* for planning stage

- *Leverage generative AI to get MTs info from Proj.Mgmt artifacts!*
- Leveraging NIH CDE Repo (& its API to interact with other tooling ↓)

- Leveraging structured-instrument resources
  - E.g., PhenX Toolkit of instruments, Patient-Reported Outcomes
- Leveraging REDCap / Other NIH IRP resources—as you build your '**data SOPs**'
  - CRISPI capacities slide

Clinical Research Informatics Supporting Principal Investigators

[UMLS / NCI-EVS also have APIs for GenAI to leverage]

…LET'S KEEP IN MIND: A ***WHOLE TEAM*** AVAILABLE TO HELP WITH USE OF **BTRIS** DATA: recap in extra slide

Extant tooling for Data: ones for planning stage

Leveraging NIH CDE Repo

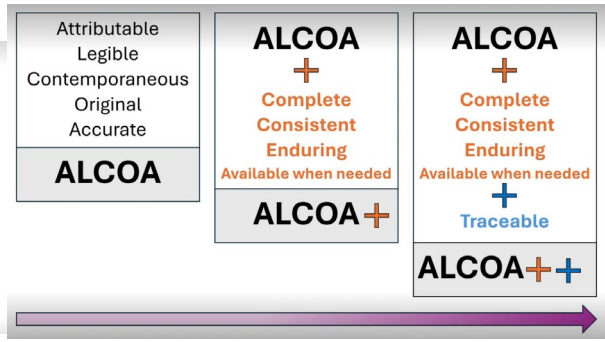
- Leveraging structured-instrument resources
- Leveraging REDCap / Other NIH IRP resources
- Upon study initiation, 'spec out' key metadata (*more on it & ALCOA in later talks*):
  - Data dictionary forms *minimal* extent of metadata to surface *publicly,* more ought to be shared
  - Keep in mind eventual 'Results' to be posted on **ClinicalTrials.gov**: do *metadata* adequately describe how estimates/findings will be obtained from underlying data *as stored?* Code for analysis?
    - *Minority of cases: regulatory-oversight trials & pre-clinical studies that also may need to submit to FDA*

cdisc SDTM — STUDY DATA TABULATION MODEL

cdisc ODM — OPERATIONAL DATA MODEL

cdisc CDASH — CLINICAL DATA ACQUISITION STANDARDS HARMONIZATION

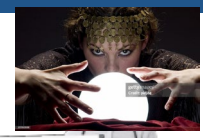
cdisc ADaM — ANALYSIS DATA MODEL



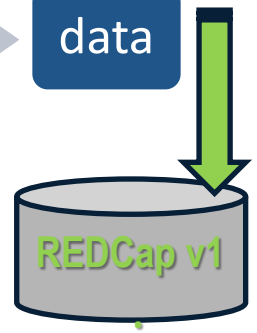
Note how MTs underly *mappability* of data elements curated in REDCap!

REDCap: tooled for CDE RE-use

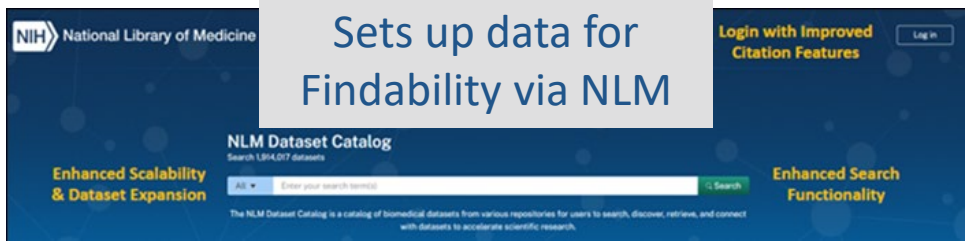
- Leveraging NIH CDE Repo Application Programming Interface (API): *future-proof*
- Leveraging structured instrument availability in other resources
  - E.g., PhenX Toolkit of instruments, CDEMapper

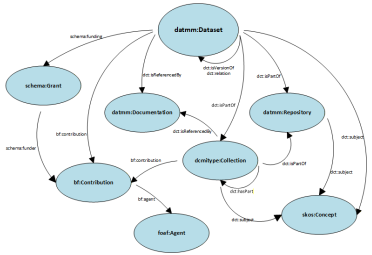


- **Leveraging REDCap as internal repository of project data AND artifacts**
  - **Version control**
  - **Code and underlying data-capture/curation, MTs' annotation → "data SOPs"**

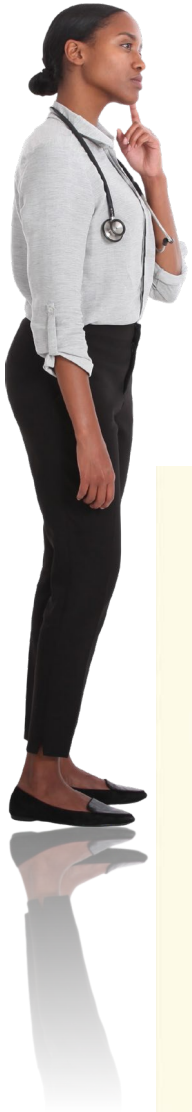


- **Upon study completion, archiving the REDCap projects' collected artifacts**
  - **Use of data lock leaves no question that the data sets *shared* are the same data sets *analyzed***
  - ***Ideally each archived version follows a metadata schema, e.g., DATaset Metadata Model (DATMM)***





How to choose a data repository: preliminaries

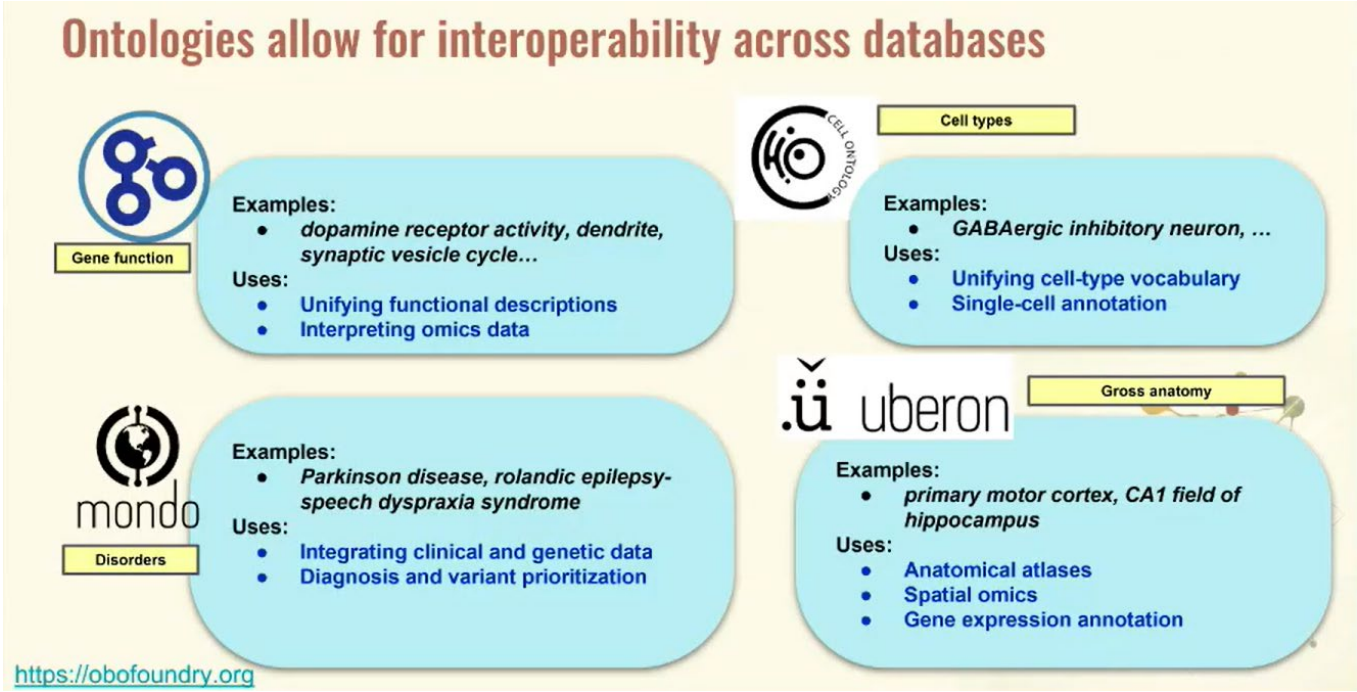


- NIH Generalist Repository Ecosystem Initiative (GREI) repo's
  - *Only as fall back to domain-specific / data-type-specific options, like NIH repo's*

Dataverse · DRYAD · figshare · OSF · MENDELEY DATA · Vivli · zenodo

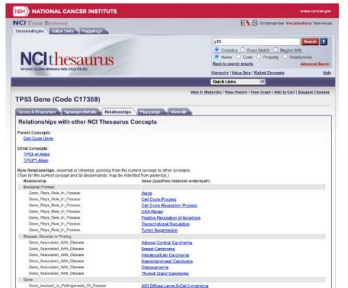
The Generalist Repository Ecosystem Initiative advances FAIR sharing of NIH-funded data — GREI

  - *Latter domain-specific repo's often leverage domain-specific semantics: for examples, ImmPort & Kidney Precision Medicine Projects employ* **ontologies***



**** Ontologies can be viewed as MTs/controlled-vocabularies that* also *specify relationships among terms & concepts; such 'added layer' of meaning can help generative AI/ LLMs perform better when doing any 'human-in-the-loop' (meta)data curation… e.g., see example kidney use case***

- "TP53 Gene" Code C17359
- Concept Relationships
  - Gene_Plays_Role_In_Process
  - Gene_Associated_With_Disease
  - Gene_Involved_In_Pathogensis_Of_Disease
  - Gene_Has_Abnormality
  - Gene_Found_In_Organism



How to choose a data repository: key desiderata

- NIH GREI repositories
  - *Only as fall back to domain-specific / data-type-specific options*
  - *Latter domain-specific repo's often leverage domain-specific semantics: for examples, ImmPort & Kidney Precision Medicine Projects employ* **ontologies**
- Key Question, with more in next talk: *is controlled-access needed*?
  - *CHECK* any human-subjects protocol's consent: scope includes sharing, then YES
    - *May* be some exceptions IF CONSENT LANGUAGE ALLOWS, to 'de-risk' re-identification
  - *Be Sure any Controlled-Access Data Repository (CADR) adheres to new NIH policy*
    - For a conversational-level sense, with Q&A, see this NIH Data Science Town Hall plus its slides
- Not controlled-data-access: ("HIPAA De-ID'd"$^{\Delta}$) Generalist Repo's

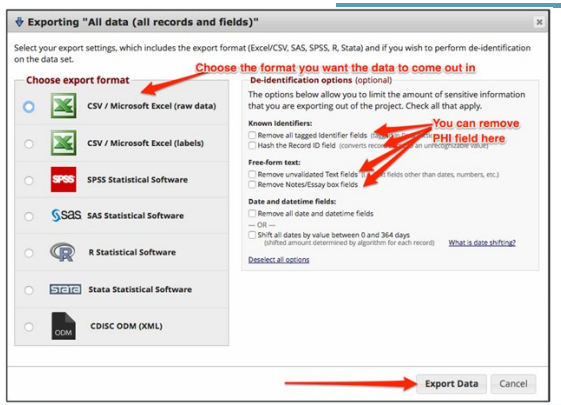
  1996 Health Insurance Portability and Accountability Act (HIPAA)

  Dataverse DRYAD figshare OSF MENDELEY DATA Vivli zenodo
- *Note: your professional profile will improve with increased data sharing, for a better S-index!*
  - In the wake of a NIH-ecosystem-wide challenge, some 'Sharing Indices' will be piloted within parts of NIH IRP

Δ To be distinguished from HIPAA Limited Data Sets; see NIH Grants Policy Statement § 4.1.4 (& links therein) for more detail

How to choose a data repository: nuanced desiderata

- NIH GREI repositories
  - *Only as fall back to domain-specific / data-type-specific options*
  - *Latter domain-specific repo's often leverage domain-specific semantics: for examples, ImmPort & Kidney Precision Medicine Projects employ* **ontologies**
- Key Question: is controlled-access needed?
  - *CHECK* any human-subjects protocol's consent: scope includes sharing, then YES
    - *May* be some exceptions IF CONSENT LANGUAGE ALLOWS, to 'de-risk' re-identification
  - *Be Sure any Controlled-Access Data Repository (CADR) adheres to new NIH policy*
- Not controlled-data-access: "HIPAA De-ID'd"$^{\Delta}$ Generalist Repo's
  - NOTE: REDCap & other systems help!
  - *Can screen for Protected Health Info**
  - *Can export into analysis-ready format:*



- Curator-hosts of GREI can help
  - Can ask NIH-hosted AI tools re:
    - Comparative features across GREI
    - Varied amounts of 1:1 support services

The Generalist Repository Ecosystem Initiative advances FAIR sharing of NIH-funded data

GREI

Dataverse DRYAD figshare OSF MENDELEY DATA Vivli zenodo

Δ To be distinguished from HIPAA Limited Data Sets; see NIH Grants Policy Statement § 4.1.4 (& links therein) for more detail; often means removal of *PHI, PII=personally identifiable information

Strategic research project planning, data collection, and data management: RECAP

https://rwdcollaborative.github.io/guide-to-n3c-v1/chapters/practices.html

The standard project-management life cycle, applied to research, from project start ("Development") to closure. Image from JHU All Children's Hospital.

Optimize Research Data Collection & Sharing

Leveraging Extant Tools For Data Collection

To summarize: *couldn't you now note that you've met 1-2 of below objectives?*

[applicable across research lifecycle]

1. What can medical terminologies (& *other semantic codings*) do for you in data collection (*from capture to curation*)?
2. How to relate to integrate data across studies/sources (added value)
3. How to choose a data repository

UMLS Unified Medical Language System (NIH National Library of Medicine) · MeSH · RxNav Navigating RxNorm Drugs · LOINC from Regenstrief · cpt (AMA) · SNOMED International · Delivering SNOMED CT

Other helpful resources /links on extra slides or in NIH SharePoint:

Any remaining questions? Cover @ end OR via email: WILKINSKJ@NIDDK.NIH.GOV

More to look out for, in more detail from before *talk…*

Leveraging Extant Tools – see *OTHER* SPACES TO WATCH FOR SOON-TO-BE-AVAILABLE TOOLING* -- team up *with informaticists!*

Would your research participants prefer to fill out structured surveys on their own devices?

- Make this readily possible by leveraging **MyREDCap / REDCap mobile app** for ***direct*** data capture!

Could your data be augmented by participants' full health context via their *own* external data records?

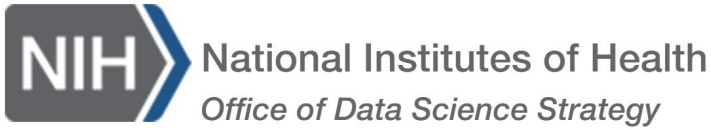
- ***Inquire with NIH informatics resource teams on consent-predicated means to leverage extant NIH-funded tools below***

Other Office of Data Science Strategy (ODSS) funded* efforts (FY26 funding call aspects recapped below)

- Data Repository Service (DRS), for NIH 'CADRs' & within Global Alliance for Genomics and Health (GA4GH)
- Fast Healthcare Interoperability Resources (FHIR®)<->REDCap:
- Observational Medical Outcomes Partnership (OMOP) & other Common Data Models (CDMs)
- Researcher Auth Service (RAS), soon to be subsumed into a broader program
- REDCap Sharing Health-data Access for Research Empowerment (SHARE): study-enrollees consent to ***external*** EHR import!
- US Core Data for Interoperability (USCDI)
- NIH CDE Repository connections to *many* of the above
- *Among others recently innovated*

A USCDI Data Class is an aggregation of Data Elements by a common theme or use case.
A USCDI Data Element is a piece of data defined in USCDI for access, exchange or use of electronic health information.

With that recap, let's get into other additional slides…



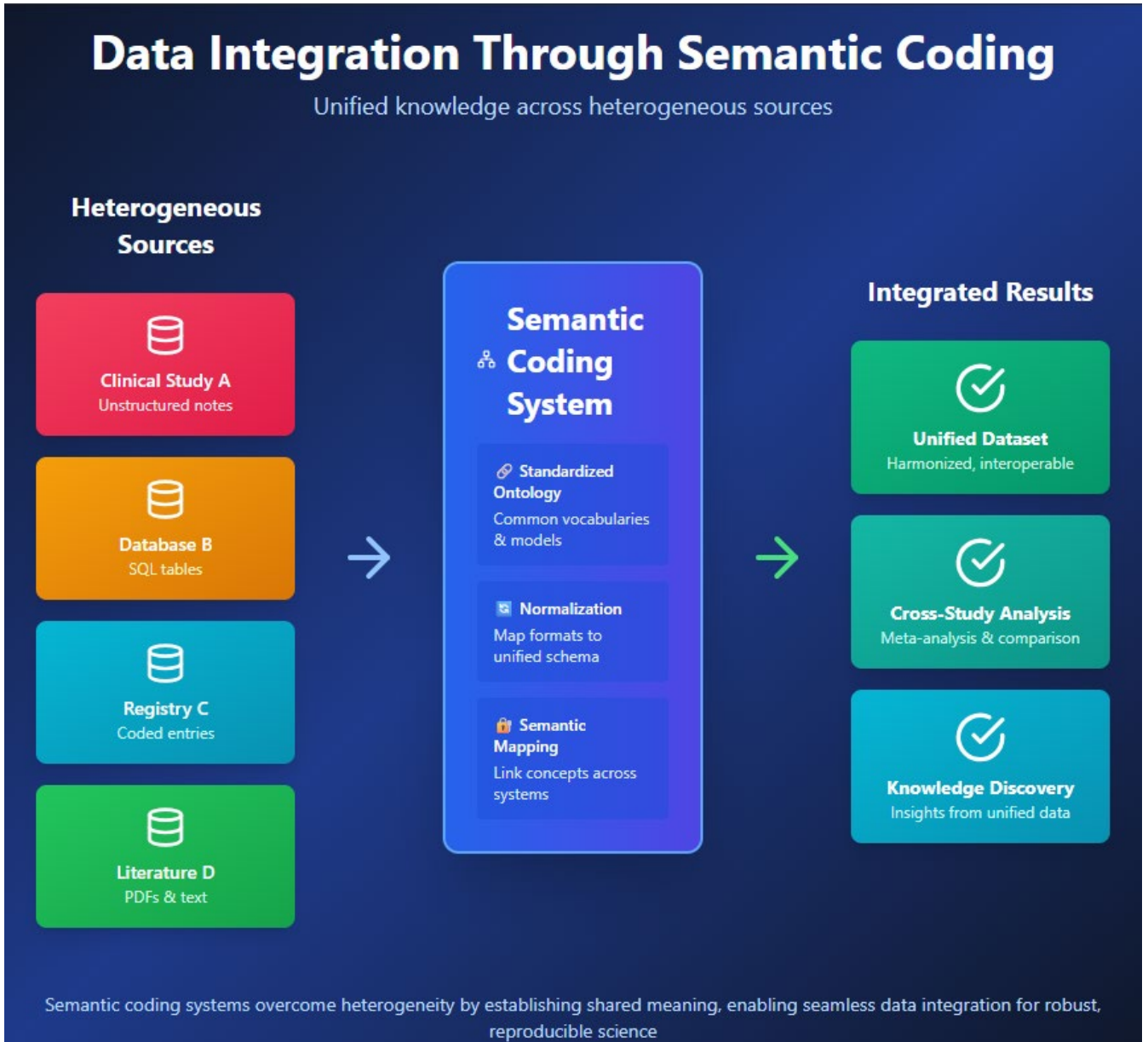
* *ODSS leads implementation of the NIH Strategic Plan for Data Science 2025-2030*

CADRs=*controlled-access data repositories*

Some GenerativeAI-leveraged Visualizations…

Claude Sonnet's rep of objective #1

Claude Haiku's rep of objective #2



Some GenerativeAI-leveraged Visualizations…

Claude Sonnet's *extended* representations of objective #1

Some GenerativeAI-leveraged Visualizations…

Claude Sonnet's *extended* representations of kidney use case *across* objectives 1-2

Prompt ⟺ Reply-text from the Large Language Model

Please generate a portrait-aspect-ratio scaled visualization including logos of a collection of medical terminologies like CPT, LOINC, MeSH, RxNorm, SNOMED-CT, and all others connected to the National LIbrary of Medicine's Unified Medical Language System or UMLS, conveying how their semantics can underly the metadata for biomedical research efforts' data collection per this learning objective: "What can medical terminologies (& other semantic codings) do for you in data collection (from capture to curation)? "

I'll research the current UMLS source vocabularies to make sure this is accurate and complete before building the visualization.
Loaded tools, used a tool
Loaded tools, used a tool
This gives me enough grounding. I'll build this as an HTML artifact (portrait aspect ratio, since it's meant as a visual poster/infographic) rather than logos-as-images (I can't embed real trademarked logos, so I'll use clean styled badges/icon representations instead — that also keeps this safely free of reproduced copyrighted marks).
UMLS Terminologies: From Capture to Curation
Interactive artifact
I built this as a portrait-format infographic poster covering your learning objective end-to-end. A quick note on the logos: I used clean, color-coded text badges instead of actual trademarked logos (SNOMED International, AMA, NLM, etc. all hold rights to their real marks), so this stays safely reusable for teaching without IP issues.
What's in it:
Header framing the learning objective directly.
The data lifecycle strip — Capture → Code → Harmonize → Integrate → Curate — so the terminologies are anchored to a concrete workflow, not just listed abstractly.
A hub-and-spoke diagram showing UMLS as the connective layer, with SNOMED CT, LOINC, MeSH, RxNorm, CPT, ICD-10-CM, HPO, Gene Ontology, MedDRA, and others radiating out — visually reinforcing that UMLS's CUIs are what tie disparate vocabularies together.
Nine terminology cards, each with what it standardizes and its primary data domain (clinical, lab, literature, drugs, procedures, genomics, safety) so a viewer can quickly map "if I'm capturing X kind of data, this is the relevant terminology."
A "why it matters" section explicitly walking through capture → structure → interoperability → metadata → curation, tying the terminologies back to the learning objective's own framing.
A few things I'd flag or that you might want to adjust:
I picked SNOMED CT, LOINC, MeSH, RxNorm, CPT, ICD-10-CM, HPO, Gene Ontology, and MedDRA as the "core nine" — UMLS actually integrates 200+ source vocabularies, so if there's a specific subset your course emphasizes (e.g., you also want NCI Thesaurus, DSM-5, or OMIM as full cards rather than just hub nodes), I can swap those in.
Want me to add a version with real question-based annotations tailored to a specific research use case (e.g., a multi-site clinical trial or an EHR-based cohort study) instead of the general framework?

Some GenerativeAI-leveraged Visualizations…

example kidney use case

APPLIED ANNOTATION · USE CASE

Two Kidney Cohorts, One Semantic Layer APOL1 · HIV Modification · Renal Adverse Events

A question-driven walkthrough of how standardized codes let you define cohorts, layer genetic risk, moderate it by HIV status, and capture renal adverse events consistently — from capture to curation.

Claude Opus's *extended* representations of kidney use case *across* objectives 1-2

Some GenerativeAI-leveraged Visualizations…

Claude Opus's *extended* representations of kidney use case *across* objectives 1-2

LAYER 1 — ADDED RISK: APOL1 GENE VARIANTS

THE GENE

APOL1 (Apolipoprotein L1) — identified by **HGNC:618**, NCBI Gene **8542**. This is the stable anchor a gene symbol alone can't guarantee (symbols get reused/renamed; the HGNC ID does not).

THE VARIANTS

Two risk alleles, coded to the base pair. The **G1** allele is two linked missense SNPs — `rs73885319` (p.Ser342Gly) and `rs60910145` (p.Ile384Met); **G2** is a 6-bp in-frame deletion, `rs71785313` (p.Asn388_Tyr389del). dbSNP rsIDs give a system-independent handle; HGVS (`p.Ser342Gly`) states the exact protein change; ClinVar records the clinical interpretation.

THE GENOTYPE

Risk is recessive — the annotation must capture zygosity, not just presence. High-risk = 2 risk alleles (G1/G1, G1/G2, or G2/G2); low-risk = 0 or 1. A flat "APOL1 variant present" flag would collapse a 3–30× risk distinction; the coding scheme has to preserve allele count.

THE PHENOTYPE

HPO / SNOMED bridge genotype to observable disease. e.g. proteinuria (`HP0 HP:0000093`), FSGS, collapsing glomerulopathy — so genomic and clinical layers join on the same concept via UMLS CUIs.

LAYER 2 — MODERATED RISK: HIV STATUS AS EFFECT MODIFIER

Q: "Does HIV status change what the APOL1 genotype means for this person?"

APOL1 high-risk genotype is often not pathogenic without a "second hit" — and **untreated HIV is one of the strongest known triggers**, with two risk alleles pushing the odds ratio for HIV-associated nephropathy (HIVAN) into the ~29–89 range in reported cohorts. So HIV is not a covariate to footnote; it is an **effect modifier** the data model must capture explicitly.

This requires **three distinct coded facts**, not one:

HIV INFECTION

Condition code — `SNOMED 86406008 / ICD-10 B20`. Presence of the diagnosis.

VIRAL CONTROL

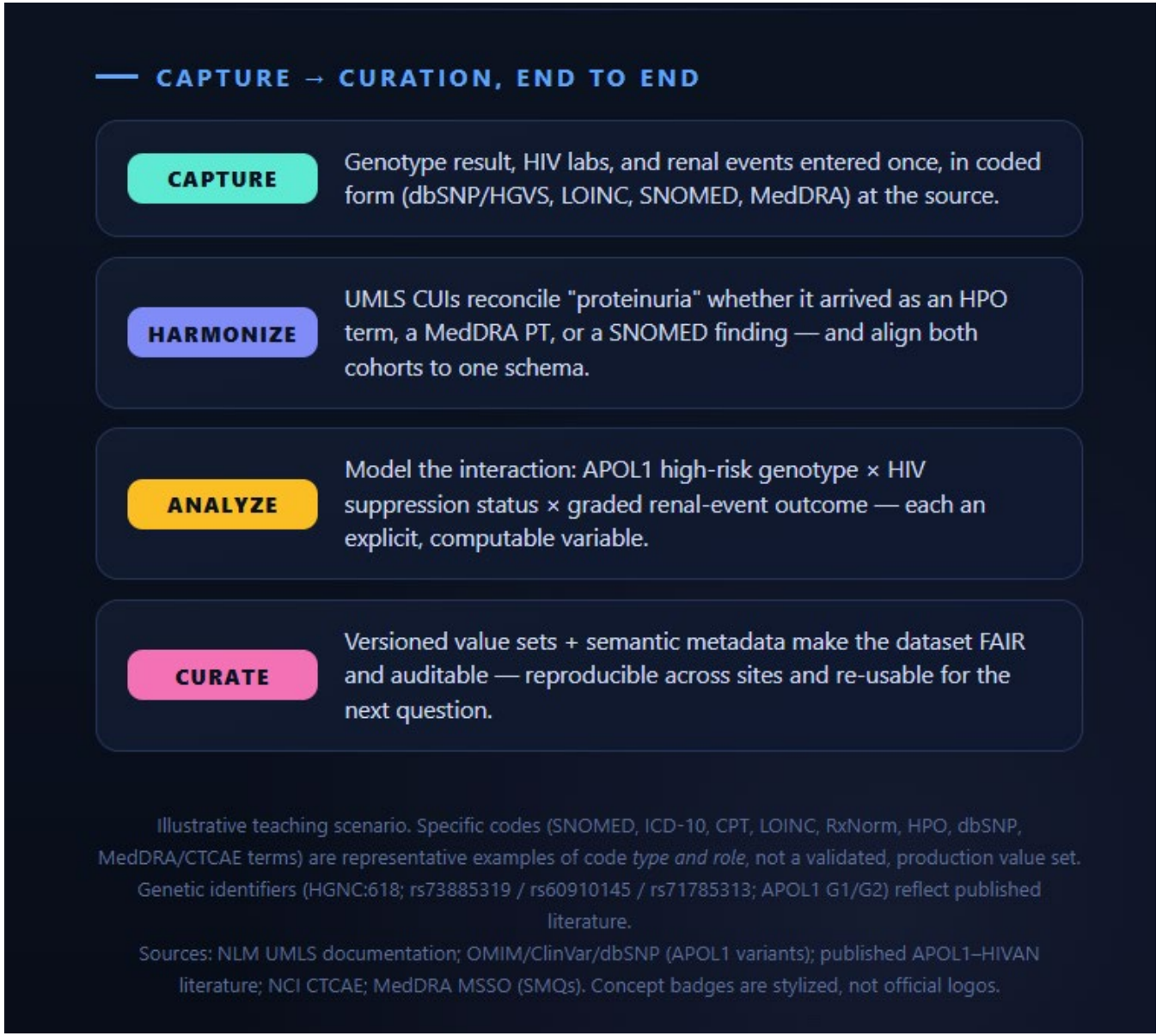
Lab observation — HIV viral load (`LOINC`) + ART drug exposure (`RxNorm`). **Treated/suppressed vs. untreated is the axis that moderates risk** — a diagnosis code alone can't express it.

RESULTING ENTITY

HIVAN as a specific nephropathy concept (`SNOMED`), distinct from generic CKD — preserving the gene × infection interaction downstream.

Some GenerativeAI-leveraged Visualizations…

Claude Opus's *extended* representations of kidney use case *across* objectives 1-2



Some Generative AI-leveraged Visualizations…

Claude Opus's *extended* representations of kidney use case *across* objectives 1-2

Prompt ⇔ Reply-text from the Large Language Model

Sure, could you please supplement this with a real question-based annotation use case that entails individuals of two cohorts: those at risk of kidney disease and those diagnosed with kidney disease with active treatment, layering on specific use of semantics and ontologies relevant to added risk due to APOL1 gene variants, modified/moderated risk due to HIV status, and multi-system coding of kidney-related adverse events (like proteinuria or 'crashing into dialysis' kidney failure) that contrast MedDRA with use of other adverse-event-semantic coding systesm? Thanks

Let me research the specifics so the annotations are accurate — APOL1 coding, the HIV interaction, and the adverse-event coding systems that contrast with MedDRA.
Web_search
Web_search
Now let me get the specifics on APOL1 variant coding identifiers (dbSNP rsIDs, HGVS, ClinVar) and the adverse-event coding systems that contrast with MedDRA for kidney events.
Web_search
Web_search
I have what I need. Let me get the precise APOL1 variant identifiers (rsIDs/HGNC) so the genomic coding annotation is exact.
Web_search
Web_search
I now have precise, verified codes for all the layers. Let me add a second page/panel to the poster with the full use-case annotation.
UMLS Terminologies: From Capture to Curation
Interactive artifact · Version 2
I added a second portrait page to the poster carrying the full use-case annotation, structured as the question-driven walkthrough you asked for. Here's what's on it and the reasoning behind each design choice:
The two cohorts are defined the way they actually differ semantically: Cohort A (at-risk) is identified by *absence* of a diagnosis code plus presence of risk factors and surveillance labs, while Cohort B (diagnosed + treated) is identified by diagnosis codes plus treatment signals (dialysis CPT, RAAS-inhibitor RxNorm exposure). That contrast is itself a teaching point — "at risk" is a harder thing to code than "diagnosed."
Layer 1 (APOL1 added risk) walks gene → variant → genotype → phenotype, because each needs a *different* coding system and each carries a lesson. I used the verified identifiers: the G1 allele is defined by rs73885319 (p.Ser342Gly) and rs60910145 (p.Ile384Met), and G2 by the indel rs71785313 (p.NYK388K/Asn388_Tyr389del). The key annotation point is zygosity: the high-risk genotype is the presence of 2 risk alleles (G1/G1, G1/G2, or G2/G2), while 0 or 1 allele is low-risk, following autosomal recessive inheritance — so a flat "variant present" flag would destroy the signal.
Layer 2 (HIV as moderator) is framed as an effect modifier, not a covariate, because the evidence is that APOL1 risk variants are not pathogenic in the absence of a second trigger, and untreated HIV is a major one — with two risk alleles the reported odds ratio for HIVAN reaches roughly 29.2 in the United States and 89 in South Africa. That's why I split it into three coded facts (infection dx / viral control via LOINC+RxNorm / resulting HIVAN entity) — the treated-vs-untreated axis is what moderates risk, and a diagnosis code alone can't express it.
Layer 3 (MedDRA vs. other AE systems) is the contrast you specifically asked for, shown as three complementary lenses on the *same* event:

- MedDRA answers "what happened, reportably" — with Acute renal failure (SMQ) providing narrow and broad search terms to retrieve cases across a safety database.
- CTCAE adds the severity axis MedDRA lacks — it's the NCI grading system, and importantly CTCAE incorporates MedDRA terminology, with each CTCAE term mapped to a MedDRA Lowest Level Term and grouped by MedDRA System Organ Class, so they layer rather than compete.
- KDIGO/SNOMED+LOINC holds the longitudinal clinical stage — the "crashing into dialysis" trajectory that safety terms don't track.

The payoff box makes the UMLS point explicit: one CUI lets a single proteinuria event carry its MedDRA PT, CTCAE grade, and SNOMED/LOINC values simultaneously.
One important caveat I built into the footnotes: I labeled the specific SNOMED/ICD/CPT/LOINC codes as *representative examples of code type and role*, not a validated value set. The genomic identifiers (HGNC:618, the three rsIDs, G1/G2) are exact from the literature, but the clinical/billing codes are illustrative — a real study would need curated, versioned value sets (e.g., via NLM's VSAC). I'd strongly recommend not lifting those specific numeric codes into a production protocol without verification.
A couple of things you might want next: I can (1) swap the illustrative codes for a properly sourced VSAC-style value set if you tell me the target coding version, (2) turn Layer 3 into a side-by-side table showing the exact same event coded in all four systems, or (3) add a visual data-model/ER diagram showing how the genotype, HIV, and AE tables join on CUIs. Want any of those?

Data Capture: RE-tooled by IRP teams for VARIED use

Leverage NIH Informatics Resources (NIDDK's, & CBIIT & CRISPI & BRICS), via REDCap, capacity to take advantage of interoperability with CRIS/BTRIS *and* other EHRs, via FHIR / TEFCA!

BRICS Biomedical Research Informatics Computing System
An Enterprise Solution

- AI Powered Data Mapping and Transformation (DMT) Tool
- AIMS4DE: Artificial Intelligence Meaningful Search for Data Elements (Beta Version)
- AI Powered RedCap to BRICS Formatting of Data Elements (HHS access required)

NIH CRISPI Clinical Research Informatics Supporting Principal Investigators

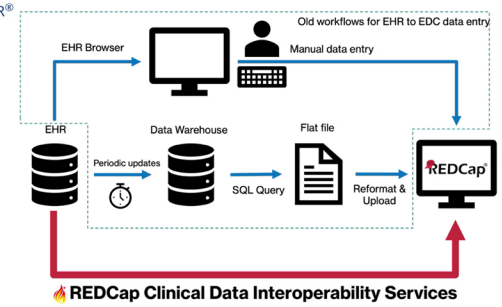
Coming Soon...

- Part 11 Compliance
- AI/LLM integration tools for IRP researchers
- REDCap+
- Federated Logons
- REDCap LEARNING LAB
- CRISPI REDCap Informational Sessions
- Enhanced Web Presence (Intranet & Public Site)
- CRISPI REDCap Roundtable

NIH National Library of Medicine
Lister Hill National Center for Biomedical Communications

LHC FHIR Tools

Tools, widgets, and services for applications using the HL7® FHIR® standard



Cheng AC, Duda SN, Taylor R, Delacqua F, Lewis AA, Bosler T, Johnson KB, Harris PA. REDCap on FHIR: Clinical Data Interoperability Services. Journal of Biomedical Informatics. 2021 Sep 1:121:103871

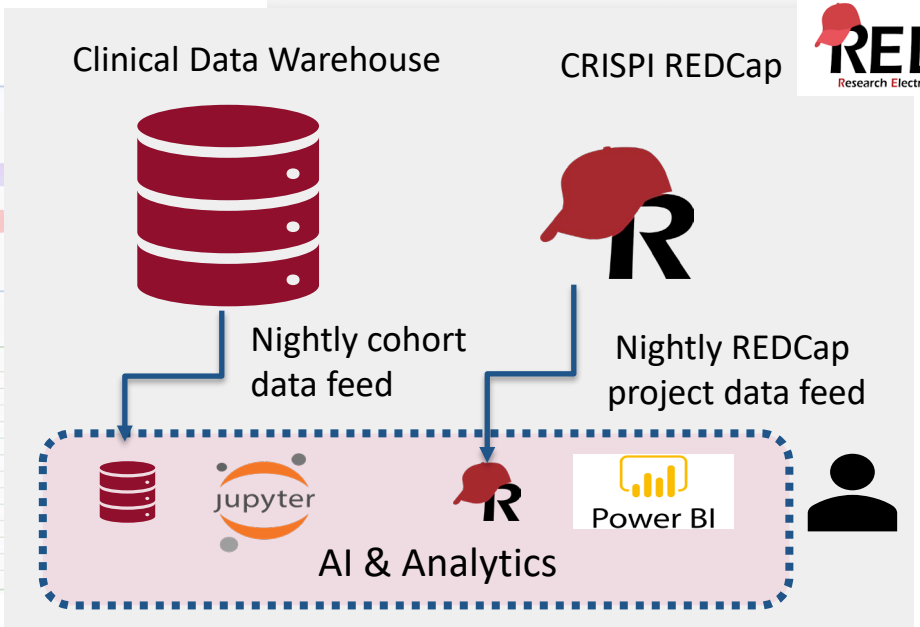
Automated Information Extraction from Clinical notes

Patient is a 65-year-old (AGE) male (SEX) with a history of type 2 diabetes mellitus (PROBLEM) and hypertension (PROBLEM). He presented to the ED complaining of chest pain (PROBLEM) and shortness of breath (PROBLEM) for 2 days. Denies fever (NEGATED), chills (NEGATED) or cough (NEGATED). Vital signs: BP 148/92 mmHg (VITAL SIGN), HR 88 bpm (VITAL SIGN), Temp 98.6 °F (VITAL SIGN), SpO2 97% (VITAL SIGN) on room air.

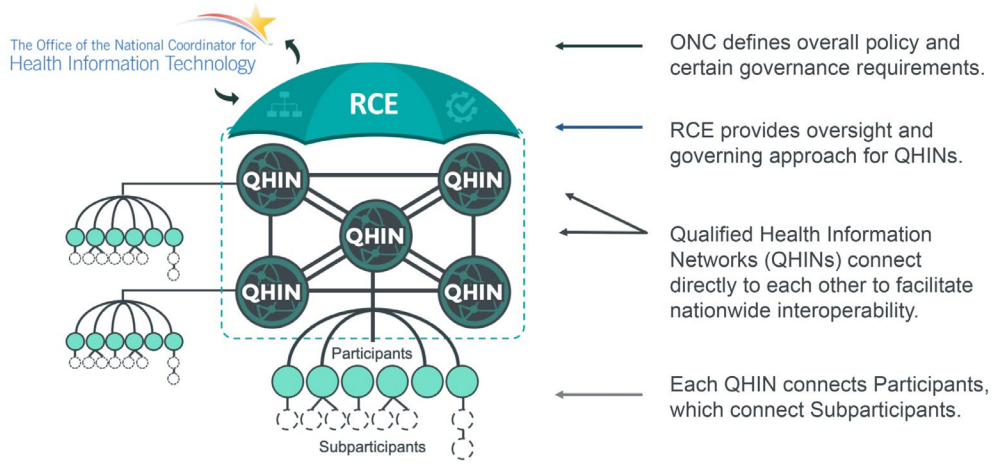
Automated Clinical Concept Extraction

Structured Output (Extracted Concepts)

Concept	Category	Status	Value / Details
65-year-old	AGE	Affirmed	65 years
Male	SEX	Affirmed	Male
Type 2 diabetes mellitus	PROBLEM	Affirmed	—
Hypertension	PROBLEM	Affirmed	—
Chest pain	PROBLEM	Affirmed	—
Shortness of breath	PROBLEM	Affirmed	Duration: 2 days
Fever	NEGATED	Negated	—
Chills	NEGATED	Negated	—
Cough	NEGATED	Negated	—
BP 148/92 mmHg	VITAL SIGN	Affirmed	148/92 mmHg
HR 88 bpm	VITAL SIGN	Affirmed	88 bpm
Temp 98.6 °F	VITAL SIGN	Affirmed	98.6 °F
SpO2 97%	VITAL SIGN	Affirmed	97% on room air



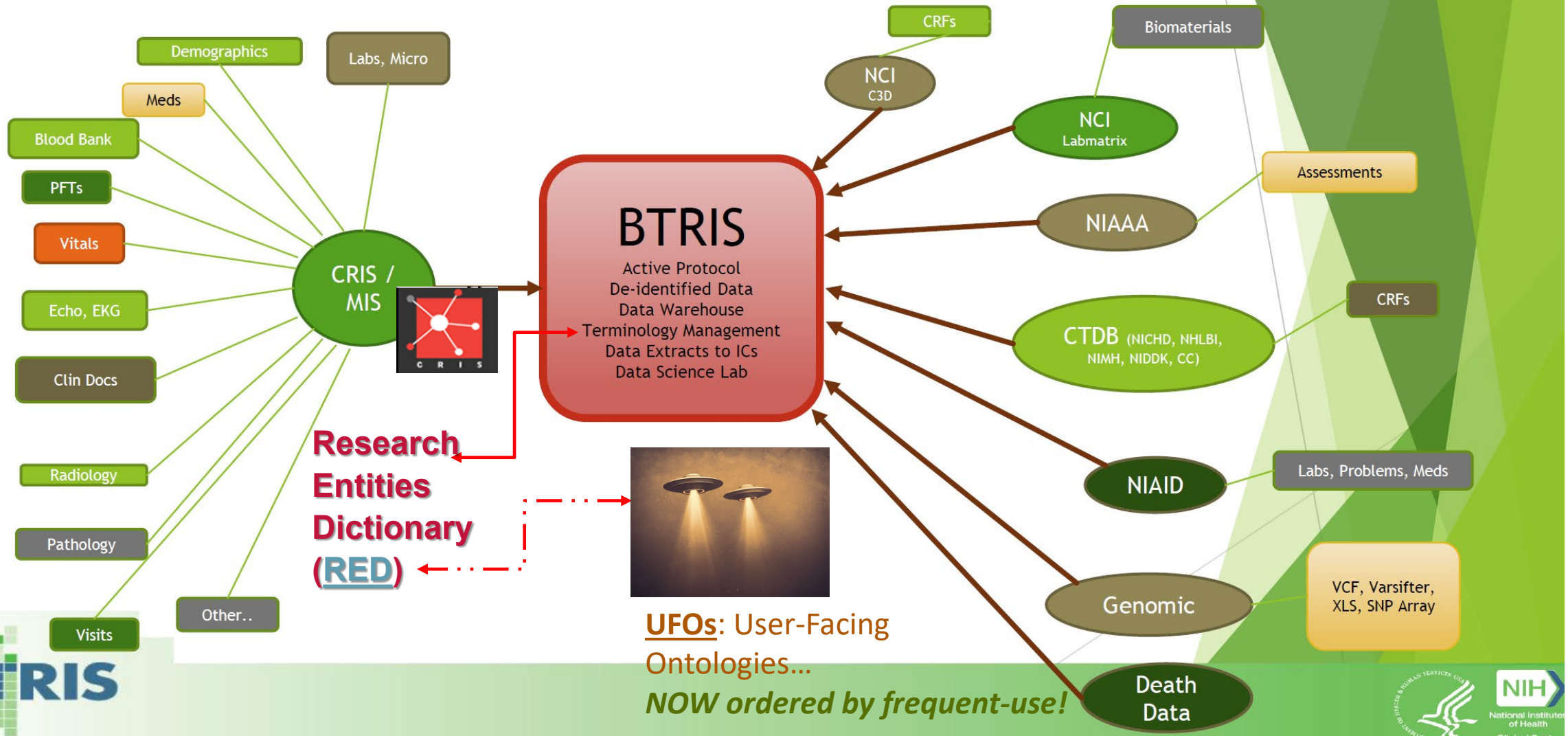
TEFCA=Trusted Exchange Framework & Common Agreement



INTRAMURAL Extra Slide: Leveraging Extant Tools When Planning Data Collection

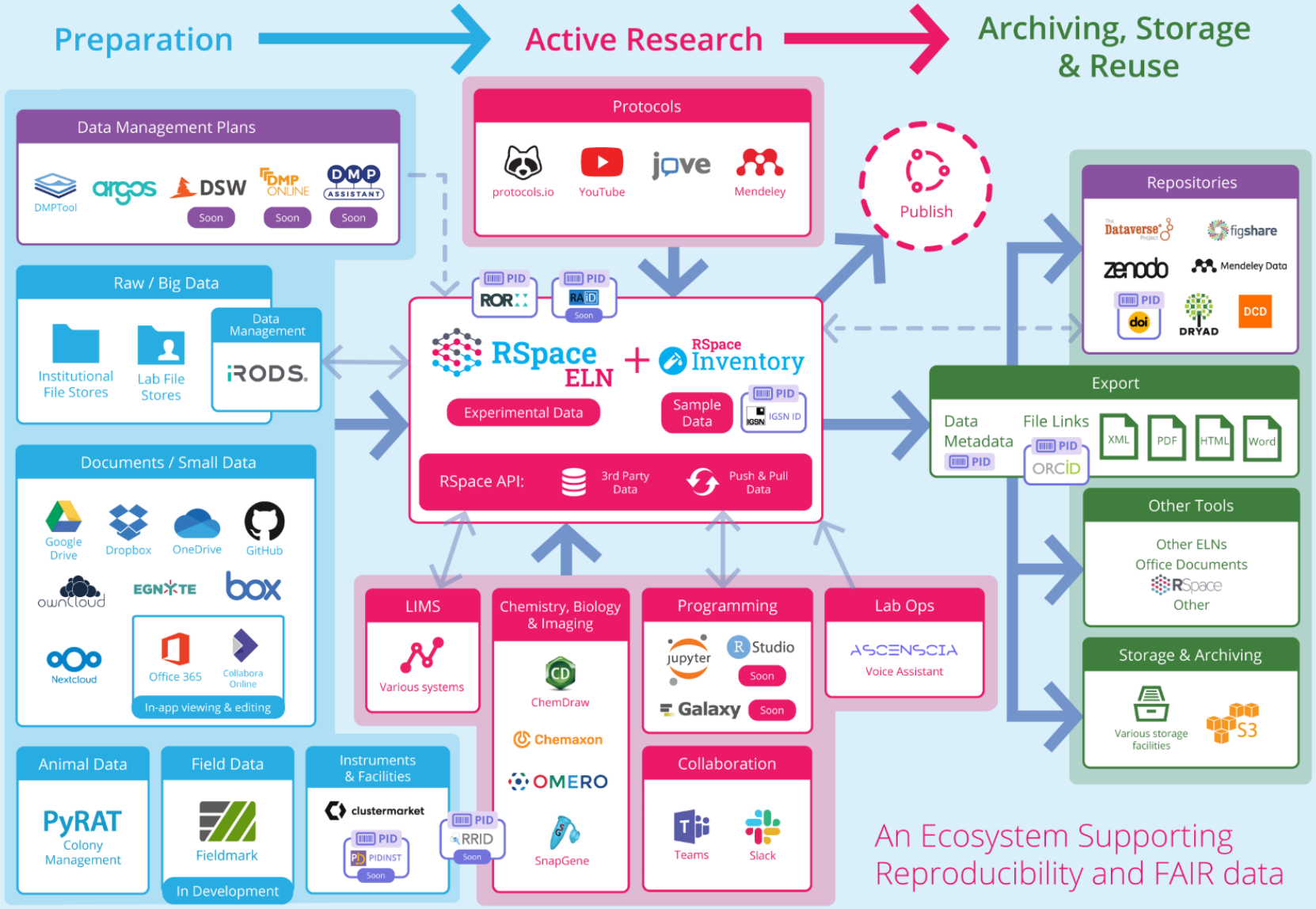
Biomedical Translational Research Informatics System (BTRIS) has over a decade's experience in leveraging terminologies in 'entity dictionaries'…recently 'ontologies'

BTRIS Data



The time is now! **Vertical interoperability** between research tools: an essential enabler for the FAIRification of data ✓
https://doi.org/10.54900/8m9by-kfy03

Slides from prior talks in webinar series, *to follow*



Jul
24
2025

NIDDK Biostats Seminar Series: Principles of Data Collection and Management

Archived

When: **Thu, Jul 24, 2025 - 2:30 pm - 4:00 pm**
Delivery: **Online**
Presented By: **Kenneth Wilkins (NIDDK), Matthew Breymaier (NIDDK), Sai Theja (NIDDK)**

Organized By: BTEP

Description

NIDDK Biostats Seminar Series: From Research Study Design to Collecting, Managing, and Analyzing Data.

Learning Objectives:

1. To delineate features of REDCap to support project management for research studies.
2. To outline steps to create detailed data collection plans which fulfill regulatory requirements.
3. To identify principled approaches to data collection and management.
4. To explain the connections between research rigor and reproducibility.

Outline:

2:30-3:00pm – Matthew Breymaier (Informatics Specialist, Office of the Clinical Director, NIDDK), Sai Theja (Senior Data Analyst, Office of the Clinical Director, NIDDK)

RedCap – functionality and basics of setup and how different types of studies can be designed in RedCap (longitudinal vs cross-sectional etc), with emphasis on NIDDK RedCap.

Other helpful resources /links within NIH SharePoint:

3:00– 4:00pm – Dr. Kenneth Wilkins (Mathematical Statistician, Biostatistics Program Office, NIDDK)

Document organization and access as part of study planning: regulatory, clinical, and case report forms

Data Management and Sharing Plans

Data Management for Reproducibility

View Details

NIH NIDDK

From Research Study Design to *Usable* Data:
Additional Slides For Reference/Further Details

ADDITIONAL SLIDES TO ROUND OUT PRACTICAL SIDES of *objectives*

Find MORE resources within our NIH SharePoint folder here

Webinar Objectives

1. **To delineate features of REDCap to support project management for research studies (e.g., how different types of studies [longitudinal vs cross-sectional etc] can be designed).**
2. **To outline steps to create detailed data collection plans which fulfill regulatory requirements.**
3. **To identify principled approaches to data collection and management.**
4. **To explain the connections between research rigor and reproducibility.**

Explore slides to address residual questions, YET feel free to follow up via email: WILKINSKJ@NIDDK.NIH.GOV

From Research Study Design to *Usable* Data: Upfront Plans per FAIR Guiding Principles

…IF you don't *FRONT-load* these decisions, issues arise too late *to address*

Get return on upfront effort: *get data that* answers your *research question*

Guiding Principles stemming from FAIR data principles, dating to 2016:

- **What does *FAIR* mean for research data management?**
  - *QUITE A BIT, yet let's give you each a starting point…*
  - *Whole courses at NIH Library on this, AND NLM has open tutorials*
  - *NIH-ecosystem-wide, ODSS has a FAIR Data & Resources Unit*

Can readily FRONT-load by RE-using existing Data Interoperability Standards:

- ***Common Data Elements (CDEs)***
- ***Common Data Models (real world data)***
- ***Data Exchange/Interchange Standards***

Findable

Accessible

Interoperable

Reusable

NIH National Library of Medicine NIH CDE Repository

NIH Endorsed CDEs *Accelerate Research*

pcornet The National Patient-Centered Clinical Research Network

cdisc

OMOP

NIH National Institutes of Health *Office of Data Science Strategy*

ODSS = Office of Data Science Strategy, https://datascience.nih.gov/ responsible for 2025-30 strategic plan

From Research Study Design to *Usable* Data: FAIR Guiding Principles, relative to AI

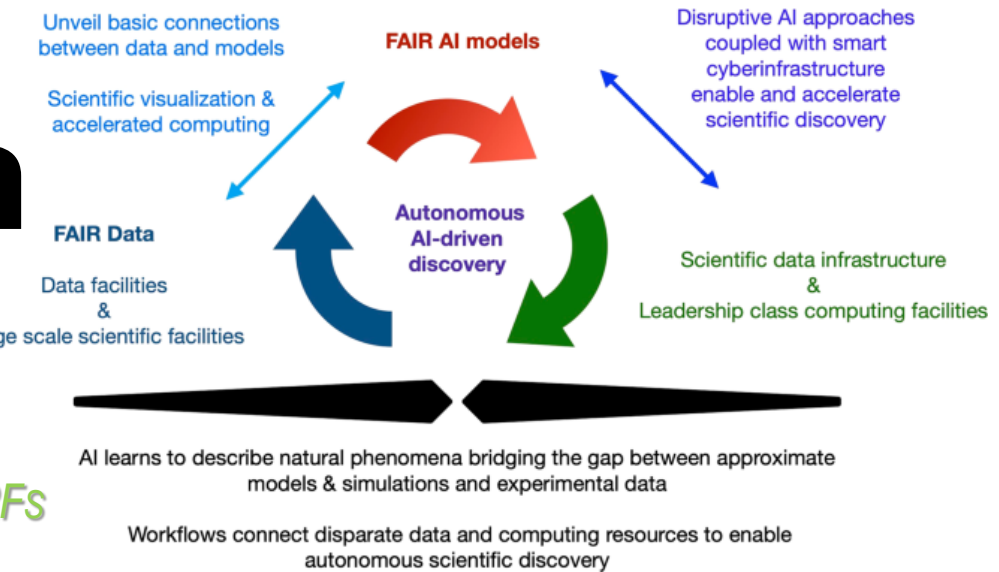
Questions to be answered: *want data to answer* your *research question?*

Guiding Principles stemming from FAIR data principles, updated to 2025

- **What does *FAIR* mean for downstream research data re-use?**
  - *Some meta-researchers say, Findable AND AI-Ready as 'FAIR'*
  - *Machine learning community proposed the Croissant Metadata standard for this*
  - *Has, in turn, led to AI models becoming FAIR=>*

scientific **data**

  - *Our focus: FINDABLE ACCESSIBLE INTEROPERABLE REUSABLE*
  - *FAIR IN REGULATORY CONTEXT COVERED BY MATT & SAI:*
    - *CLINICAL DATA INTERCHANGE STANDARDS CONSORTIUM (CDISC) & ITS*
    - *CLINICAL DATA ACQUISITION STANDARDS HARMONIZATION (CDASH) E-CRFS*



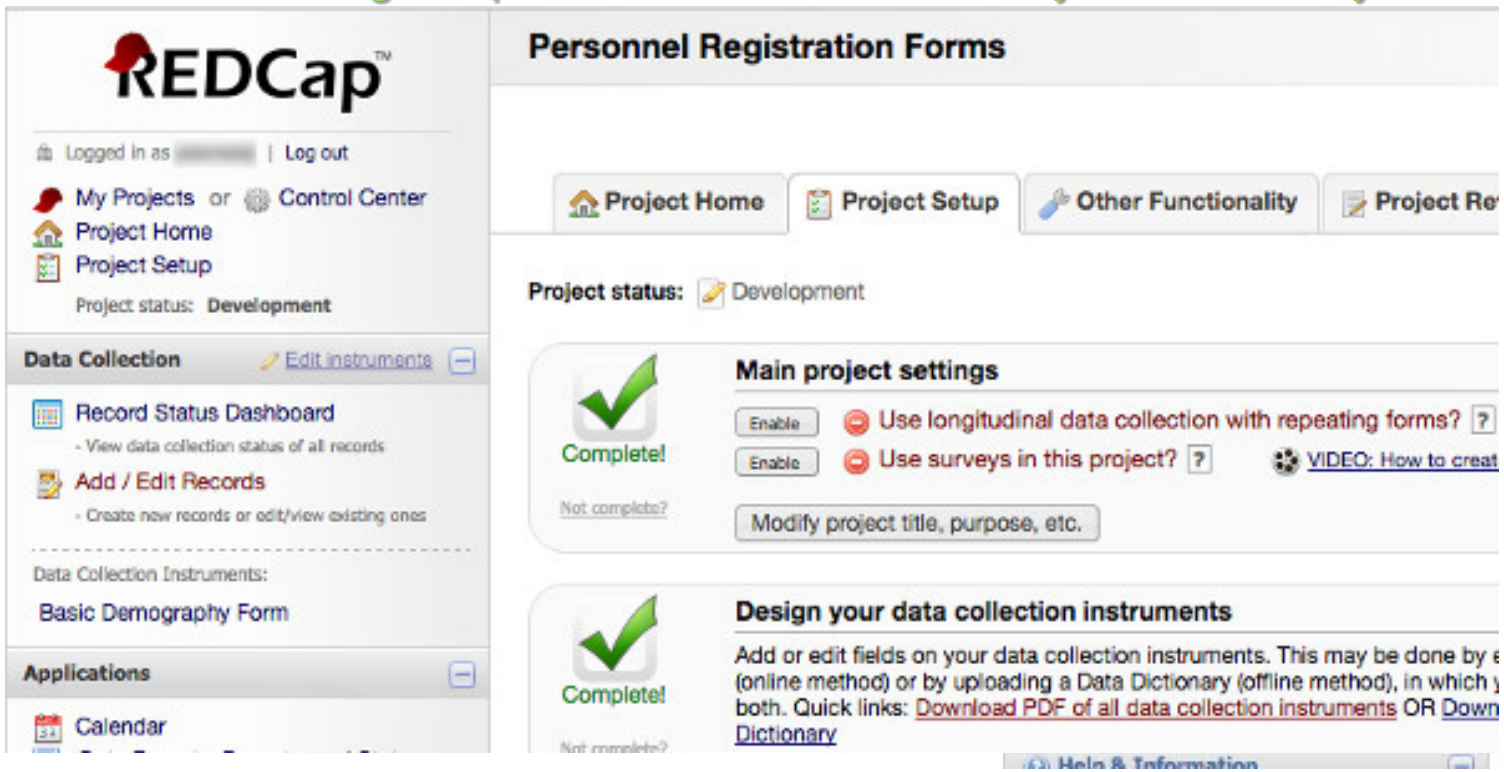
APPENDIX on REDCap: add'l slides

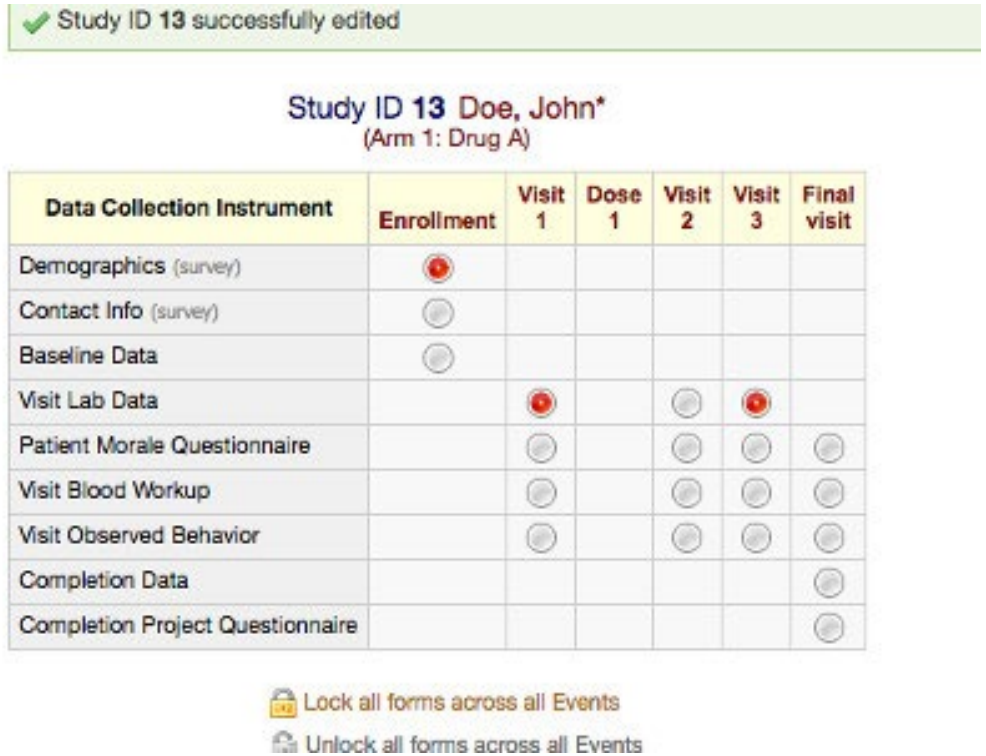
From Research Study Design to *Usable* Data: Take Home Examples on Guiding Principles

You have pre-vetted starting point* because of *Intramural Research Program REDCap

Much of our flexibility of data specifications *enabled* by how it's been configured initially

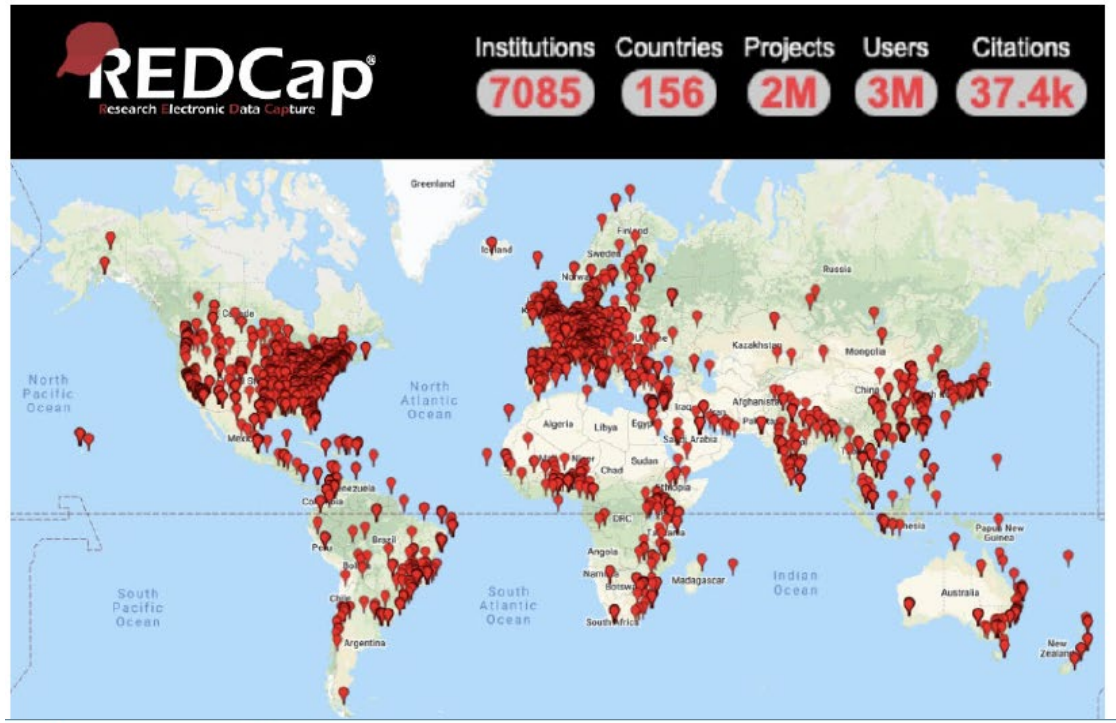
- *Some other features keep getting added centrally by a worldwide REDCap development community*
  - Can manage* operational workflows as you conduct your research, adapt CRFs for *longitudinal* reuse

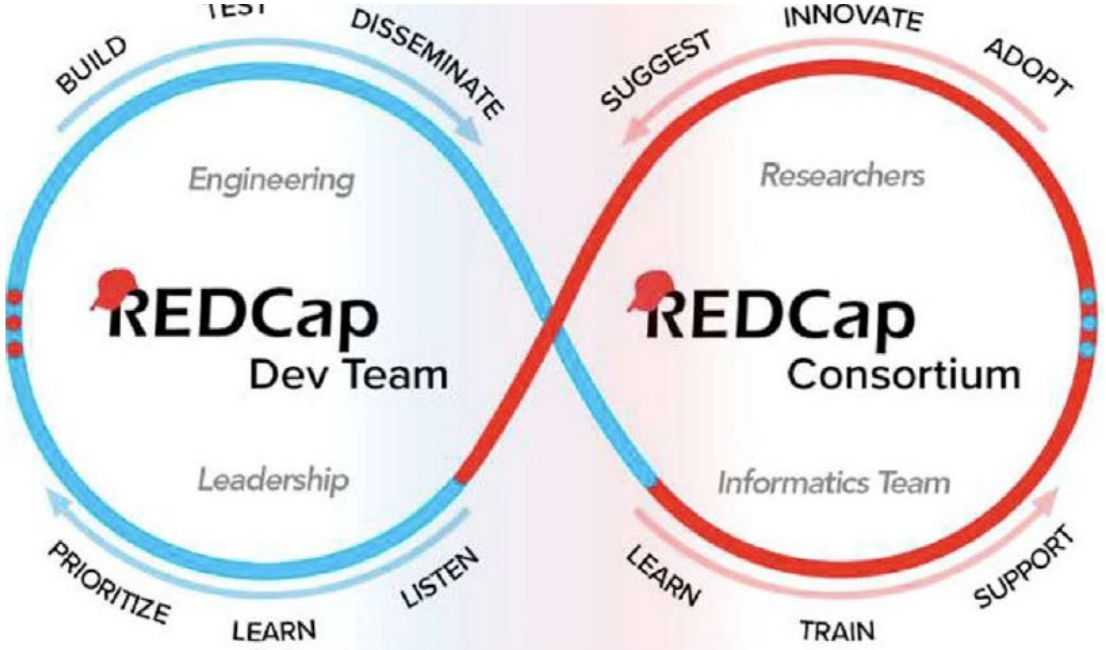




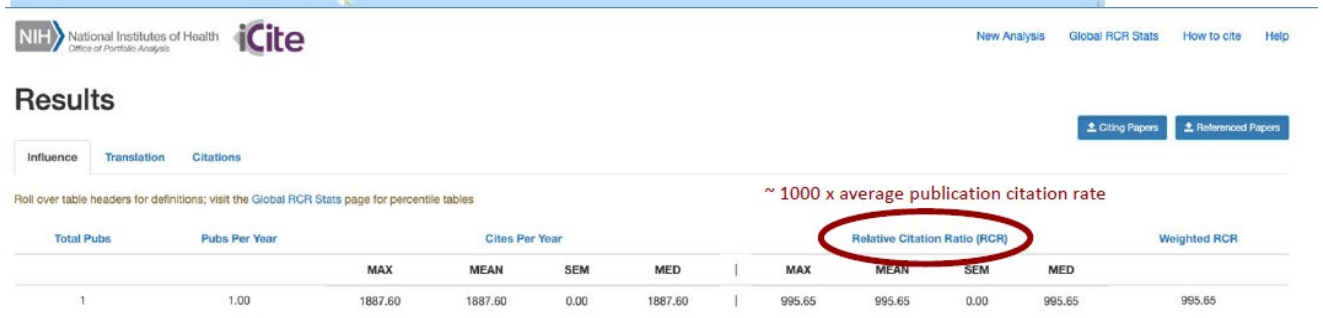
*To delineate features of REDCap to support project management for research studies/designs, as covered by Matt & Sai earlier.

You could leverage a *very* broad user community





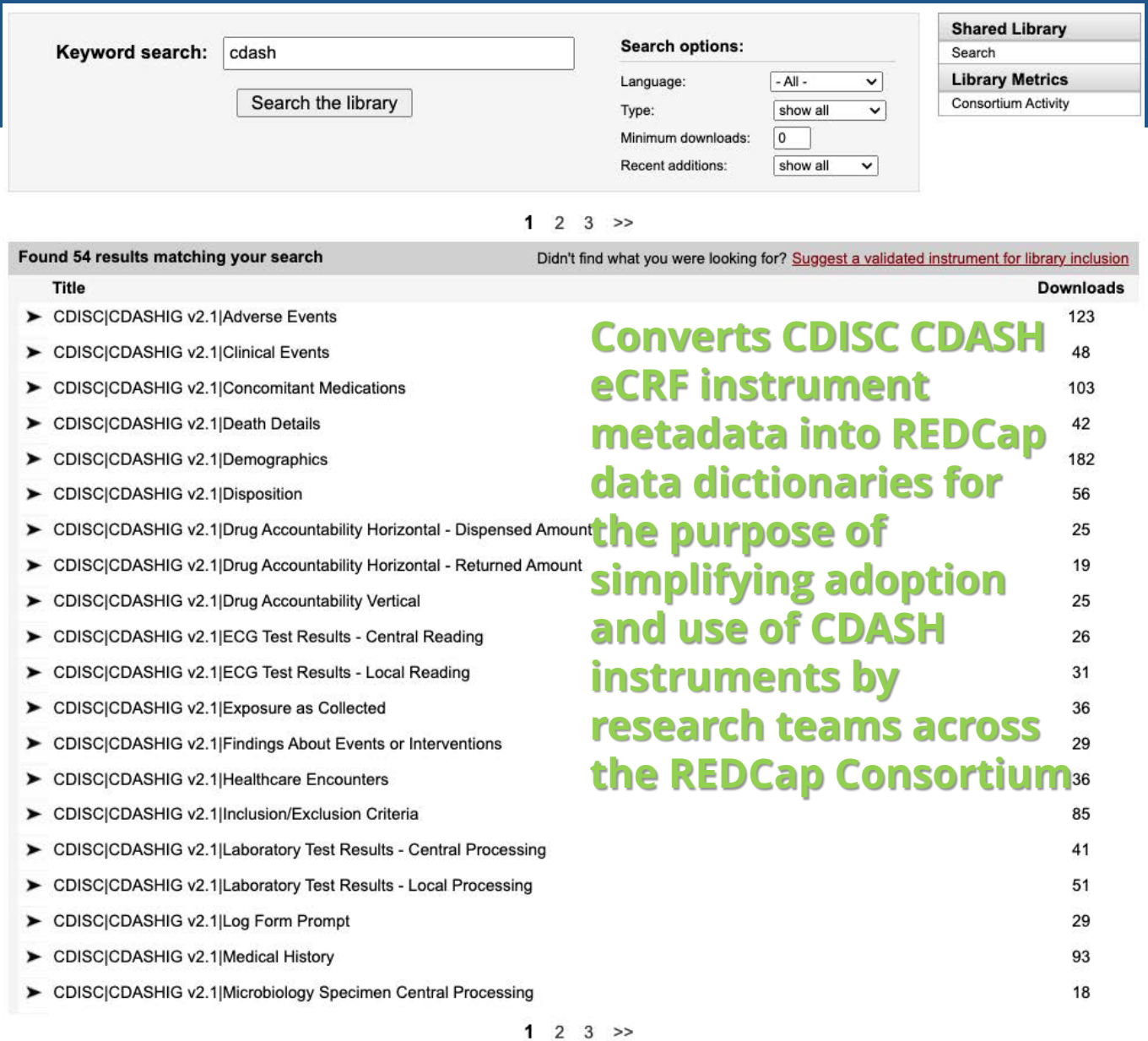
Innovation Zone



In the event of limited time skip to CDE API use

From REDCap founder Paul Harris presentation in March 2024

Leverage user community: *Regulatory* Compliance



Converts CDISC CDASH eCRF instrument metadata into REDCap data dictionaries for the purpose of simplifying adoption and use of CDASH instruments by research teams across the REDCap Consortium

Journal of the Society for Clinical Data Management

Cheng AC, et al. Creating and Disseminating CDASH Harmonization Electronic Case Report Forms on the REDCap Shared Data Instrument Library. *Journal of the Society for Clinical Data Management*. 2022; 2(1): 7, pp. 1–5. DOI: https://doi.org/10.47912/jscdm.172

ORIGINAL RESEARCH

Creating and Disseminating CDASH Harmonization Electronic Case Report Forms on the REDCap Shared Data Instrument Library

Alex C. Cheng*, Rhonda Facile†, John Owen†, Richard Marshall†, Kathleen Mellars†, Nan Kennedy*, Brenda L. Minor*, Kyle Mcguffin* and Paul Harris*

Introduction: A guiding principle behind the development and deployment of the REDCap data management platform has always included attention to workflow design that allows easy implementation of best practices for clinical and translational researchers. CDISC standards such as CDASH have helped the clinical research community improve the efficiency, actionability, and quality of their clinical trials data, but have had limited uptake among the academic institutions.

Objective: To create a scalable methodology to convert CDISC CDASH electronic case report forms (eCRFs) instrument metadata into REDCap data dictionaries for the purpose of simplifying adoption and use of CDASH instruments by research teams across the REDCap Consortium.

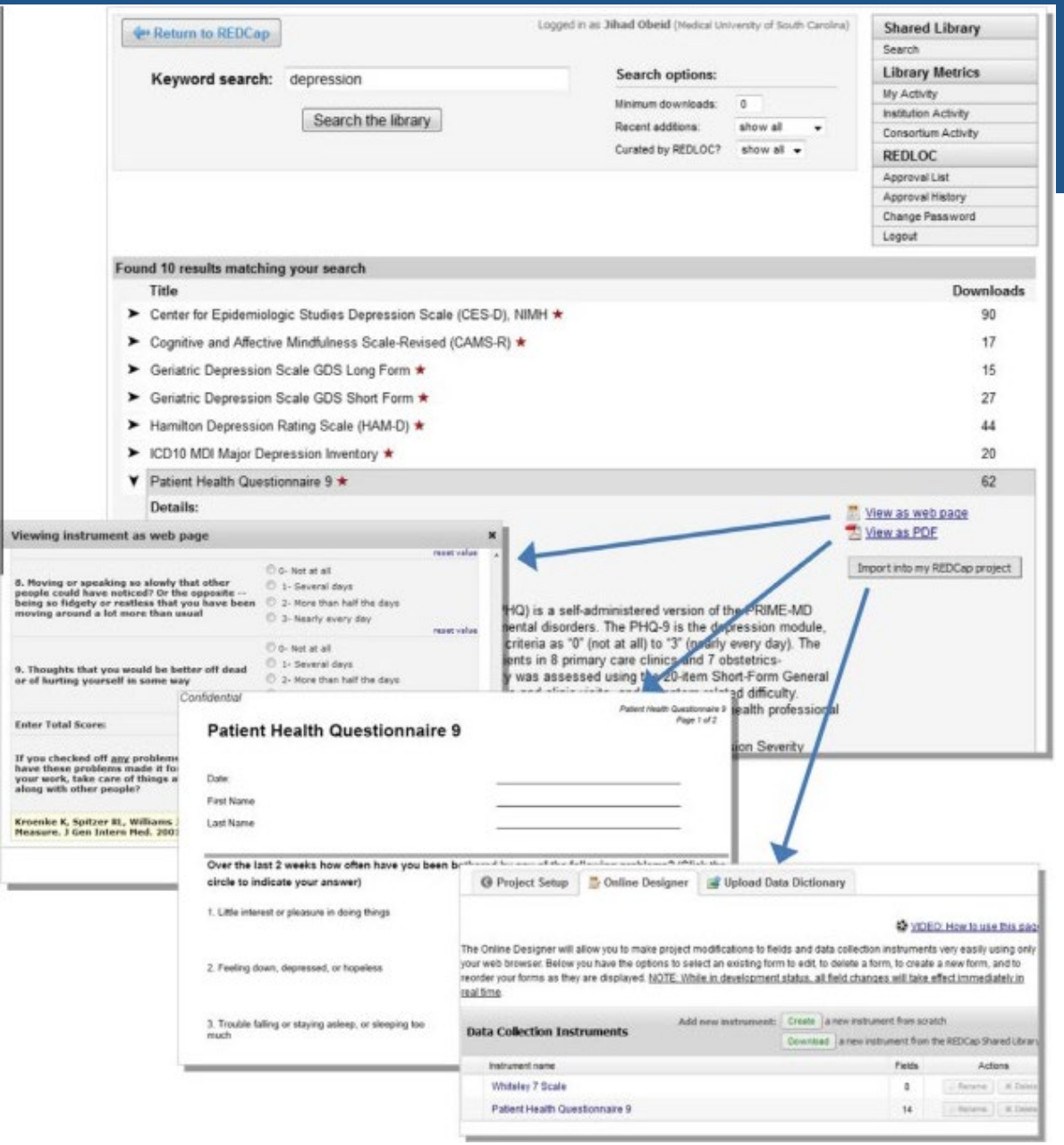
Implementation: We have used our replicable methods to translate metadata from 34 CDASH Foundational eCRFs and 20 CDASH Crohn's Disease eCRFs into REDCap eCRF metadata and have made these instruments available in the REDCap Shared Data Instrument Library for widespread sharing and uptake across the REDCap Consortium. Users can import the standardized eCRFs directly into their REDCap projects for immediate use in clinical trial data collection.

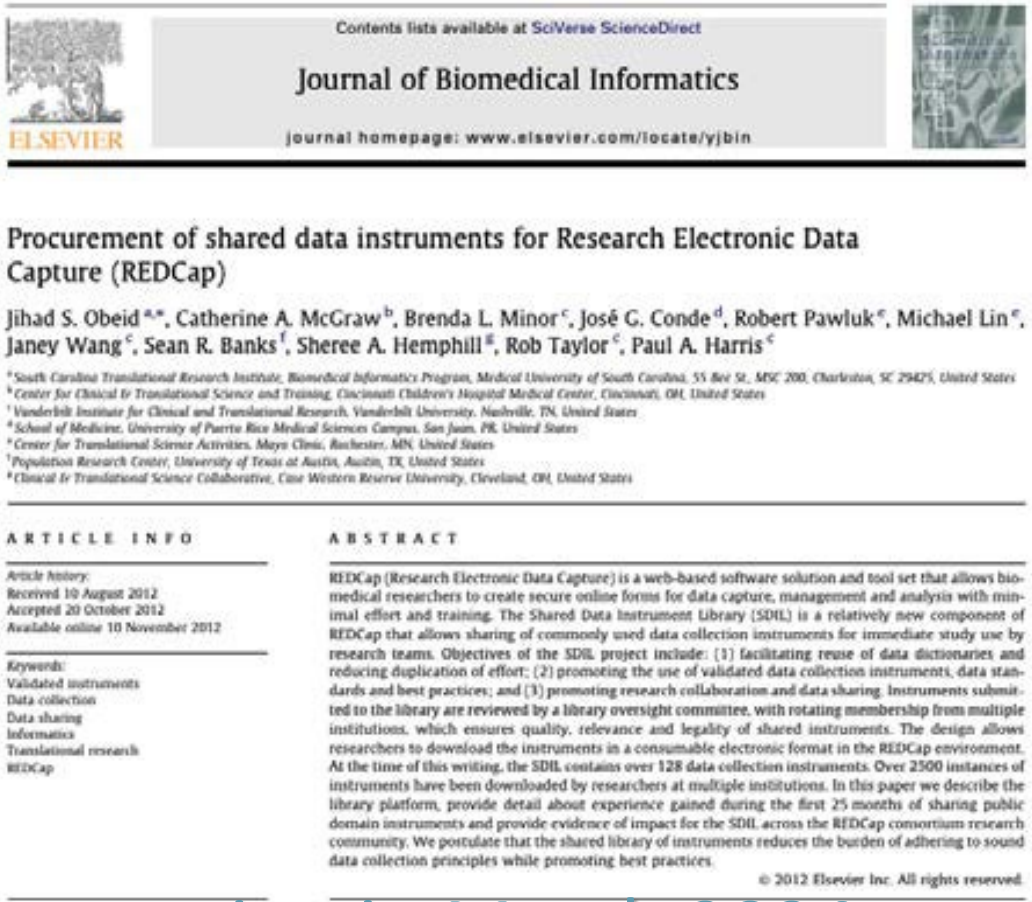
Conclusion: Disseminating CDISC standards through the REDCap community will increase the accessibility of these standards for academic medical centers. Having academic clinical researchers using CDISC standards may lead to more research datasets that interoperate with pharmaceutical sponsored trials, and more discoveries from secondary use of clinical research data.

From REDCap founder Paul Harris presentation in March 2024

Leverage user community: *Ease of* CRF development (1)

EVEN *IF* You don't need FDA-ready CDASH e-CRFs via REDCap, re-*use* prior instruments!





Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Procurement of shared data instruments for Research Electronic Data Capture (REDCap)

Jihad S. Obeid [a,*], Catherine A. McGraw [b], Brenda L. Minor [c], José G. Conde [d], Robert Pawluk [e], Michael Lin [e], Janey Wang [c], Sean R. Banks [f], Sheree A. Hemphill [g], Rob Taylor [c], Paul A. Harris [c]

[a] South Carolina Translational Research Institute, Biomedical Informatics Program, Medical University of South Carolina, 55 Bee St., MSC 200, Charleston, SC 29425, United States
[b] Center for Clinical & Translational Science and Training, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[c] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University, Nashville, TN, United States
[d] School of Medicine, University of Puerto Rico Medical Sciences Campus, San Juan, PR, United States
[e] Center for Translational Science Activities, Mayo Clinic, Rochester, MN, United States
[f] Population Research Center, University of Texas at Austin, Austin, TX, United States
[g] Clinical & Translational Science Collaborative, Case Western Reserve University, Cleveland, OH, United States

ARTICLE INFO

Article history:
Received 10 August 2012
Accepted 20 October 2012
Available online 10 November 2012

Keywords:
Validated instruments
Data collection
Data sharing
Informatics
Translational research
REDCap

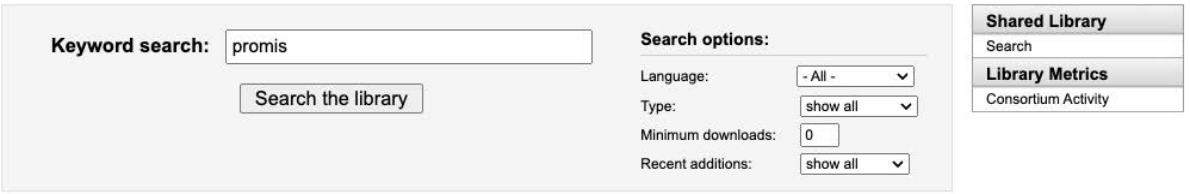
ABSTRACT

REDCap (Research Electronic Data Capture) is a web-based software solution and tool set that allows biomedical researchers to create secure online forms for data capture, management and analysis with minimal effort and training. The Shared Data Instrument Library (SDIL) is a relatively new component of REDCap that allows sharing of commonly used data collection instruments for immediate study use by research teams. Objectives of the SDIL project include: (1) facilitating reuse of data dictionaries and reducing duplication of effort; (2) promoting the use of validated data collection instruments, data standards and best practices; and (3) promoting research collaboration and data sharing. Instruments submitted to the library are reviewed by a library oversight committee, with rotating membership from multiple institutions, which ensures quality, relevance and legality of shared instruments. The design allows researchers to download the instruments in a consumable electronic format in the REDCap environment. At the time of this writing, the SDIL contains over 128 data collection instruments. Over 2500 instances of instruments have been downloaded by researchers at multiple institutions. In this paper we describe the library platform, provide detail about experience gained during the first 25 months of sharing public domain instruments and provide evidence of impact for the SDIL across the REDCap consortium research community. We postulate that the shared library of instruments reduces the burden of adhering to sound data collection principles while promoting best practices.

From REDCap founder Paul Harris presentation in March 2024

Leverage user community: *Ease of* CRF development (2)

PROMIS® (Patient-Reported Outcomes Measurement Information System), part of HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 >>

Found 1869 results matching your search — Didn't find what you were looking for? Suggest a validated instrument for library inclusion

Title	Downloads
PROMIS-29+2 Profile/Battery v2.1 (PROPr) [Battery]	204
PROMIS - 29 Profile/Battery v2.0 [Battery]	499
PROMIS - 43 Profile v2.0 (Spanish)	15
PROMIS - 43 Profile/Battery v2.0 [Battery]	36
PROMIS - 57 Profile v2.0 (Spanish)	19
PROMIS - 57 Profile/Battery v2.0 [Battery]	62
PROMIS Item Bank v1.0 - Trastornos del Sueño - Cuestionario Abreviado 6a	6
PROMIS Item Bank v1.0 - Trastornos del Sueño - Cuestionario Abreviado 8a	4
PROMIS Item Bank v1.1 - Trastorno Emocional - Enojo	6
PROMIS Item Bank v1.1- Efectos del Dolor	16
PROMIS Item Bank v1.2 - Capacidad de Funcionamiento Físico	15
PROMIS Item Bank v2.0 - Aislamiento social - Cuestionario abreviado 4a	18

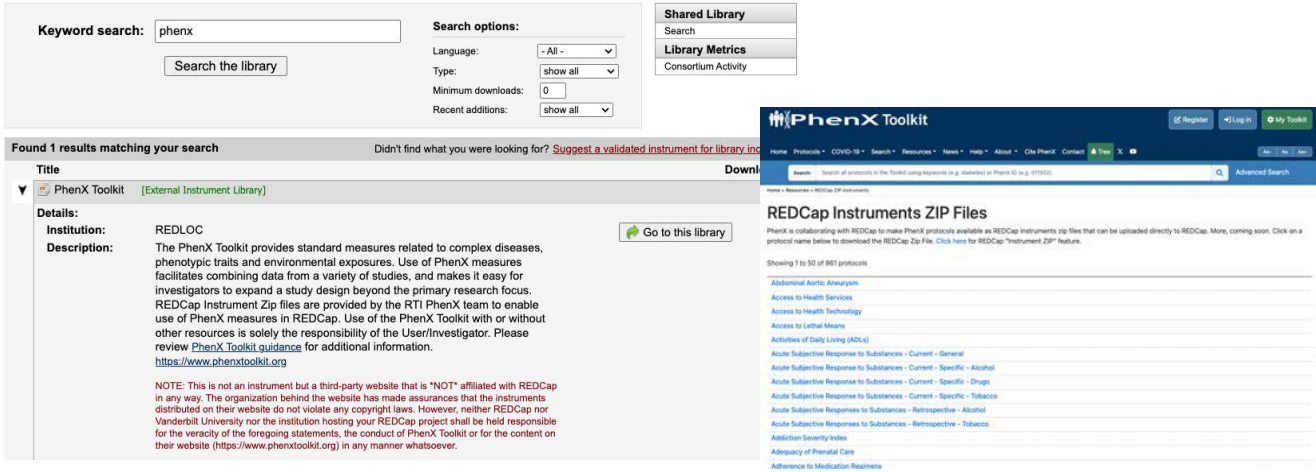
PROMIS, Patient-Reported Outcomes Measurement Information System, & the PROMIS logo are marks owned by the U. S. Department of Health & Human Services.

Differences between PROMIS Measures
Computer Adaptive Tests (CATs) versus Short Forms
•Many domains offer a **computer adaptive test (CAT)** and one or more short forms. Select that type of measure that fits your needs and resources. **Learn more>>**

- CATs
  - Tailored selection of items for each respondent
  - Requires administration technology
  - High measurement precision across a wide range of symptom/function severity
- Short forms
  - All respondents answer all questions
  - No special administration technology needed
  - Degree of measurement precision varies

Example *Reproducible-by-design* CRF development via REDCap, re-*use* participant-reported outcomes instruments via PROMIS *OR* others (NIH Phenotype eXplorer Toolkit, if not NIH Toolbox®)

PhenX Toolkit



From REDCap founder Paul Harris presentation in March 2024

Leverage user community: *Ease of* CRF development (3)

NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

NIH NATIONAL CANCER INSTITUTE

Manage | View | Collaborate | Data Interchange | Personalize

>>Favorites>>NCI Standard CRFs

isplay Values | View | Delivery Options | Add to Guest User Cart

mplate Rows 1..130 of 130 1 Request

iew	1 Form Name	Public ID	Version	Owned By
	Adverse Event/Serious Adverse Event CTCAE v3 NCI Standard Template	2748814	3.00	NCI Standards
	Adverse Event/Serious Adverse Event CTCAE v4 NCI Standard Template	3265657	3.00	NCI Standards
	Adverse Event/Serious Adverse Event CTCAE v4.0 CDISC Aligned NCI Standard Template	6988567	1.00	NCI Standards
	Adverse Event/Serious Adverse Event CTCAE v5.0 CDISC Aligned NCI Standard Template	6984728	1.00	NCI Standards
	Adverse Event/Serious Adverse Event CTCAE v5.0 NCI Standard Template	6081561	1.00	NCI Standards
	Brief Pain Inventory-BPI CDISC Aligned NCI Standard Template	7093187	2.00	NCI Standards
	Concomitant Medication CDISC Aligned NCI Standard Template	6984232	1.00	NCI Standards
	Concomitant Medication NCI Standard Template	2867231	2.00	NCI Standards
	Consent CDISC Aligned NCI Standard Template	6936999	1.00	NCI Standards

Extended examples of *Reproducible-by-design* CRF development via REDCap, re-*use of longstanding Common Data Elements*, at least for *NCI's Adverse Events data elements*

[for more on NCI CDEs, see its (NIH) Enterprise Vocabulary Services or EVS, esp. its Cancer Data Standards Registry or caDSR sites]

From REDCap founder Paul Harris presentation in March 2024

From Research Study Design to *Usable* Data: Take Home on "Using REDCap for FAIR data"

BONUS pre-vetted starting points* within *Intramural Research Program REDCap

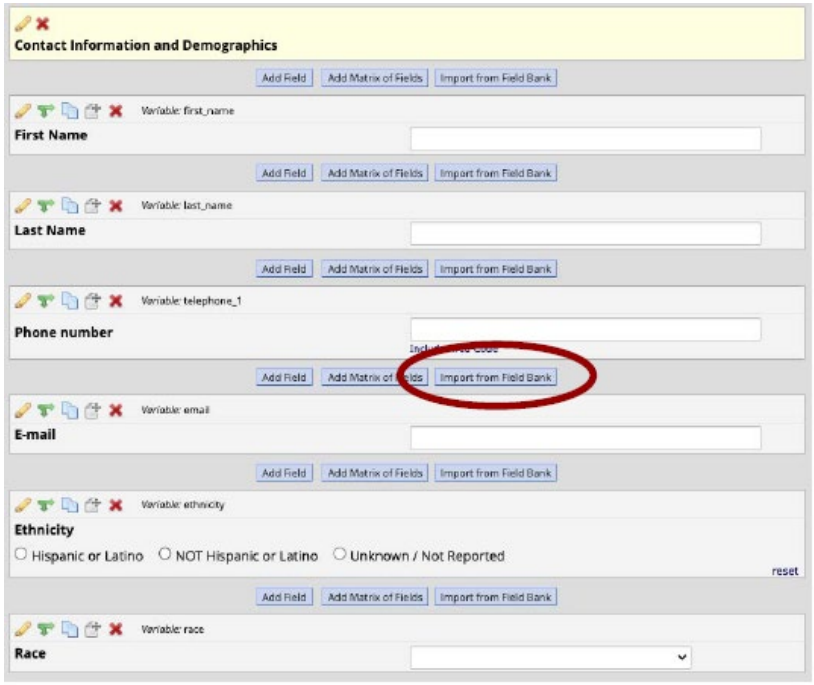
Inherent capacity to semantically search up COMMON DATA ELEMENTS *within* REDCap

- *Some that—since 2020 NIH-wide criteria were set—are designated as NIH-ENDORSED!*

[*if you don't' use REDCap*]

[*try CDEMapper toolkit*]

- Example of demographics:
  - NAME
  - TEL#
  - EMAIL
  - ETHNICITY*
  - RACE*



National Institutes of Health
Endorsed CDEs Accelerate Research

REDCap
Research Electronic Data Capture

*In the case of 'Demographics' be aware of Federal Government changes in capturing Race/Ethnicity as specified in Statistical Policy Directive #15's revision: https://spd15revision.gov/

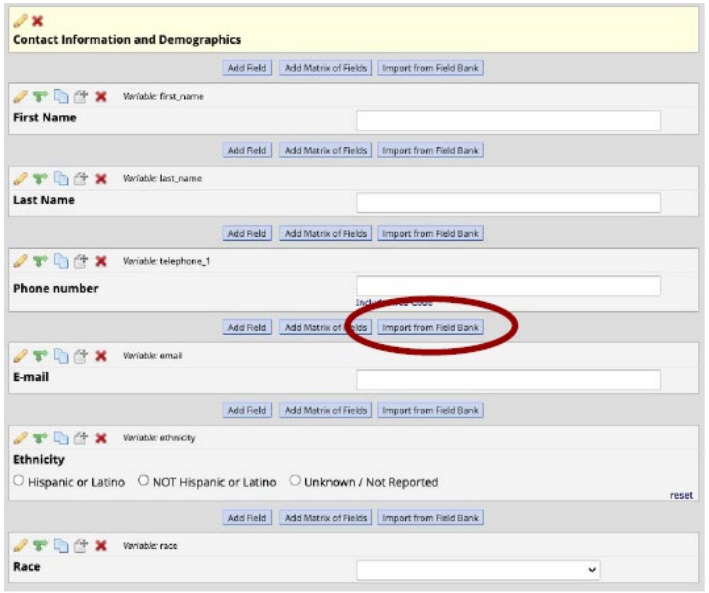
In the event of NOT limited time, leading here, skip back to take-homes

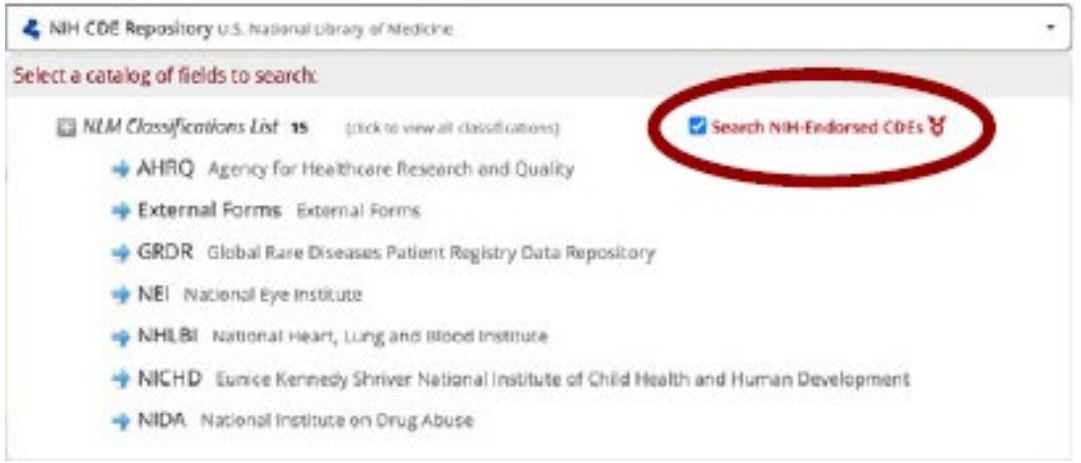
From Research Study Design to *Usable* Data: Take Home on "Using REDCap for FAIR data"

You have pre-vetted starting point* because of *Intramural Research Program REDCap

Inherent capacity to semantically search up COMMON DATA ELEMENTS *within* REDCap

- *Some that—since 2020 NIH-wide criteria were set—are NIH-ENDORSED!*
- Example of demographics:





NIH National Institutes of Health
Endorsed CDEs Accelerate Research

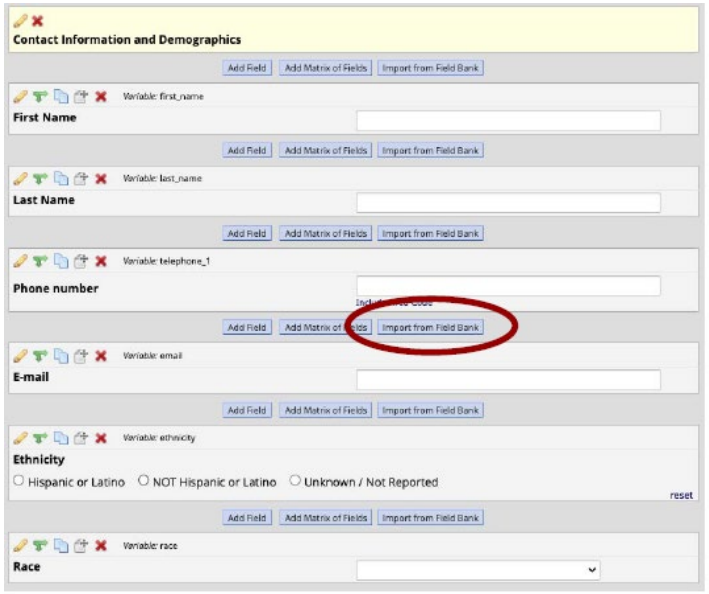
REDCap
Research Electronic Data Capture

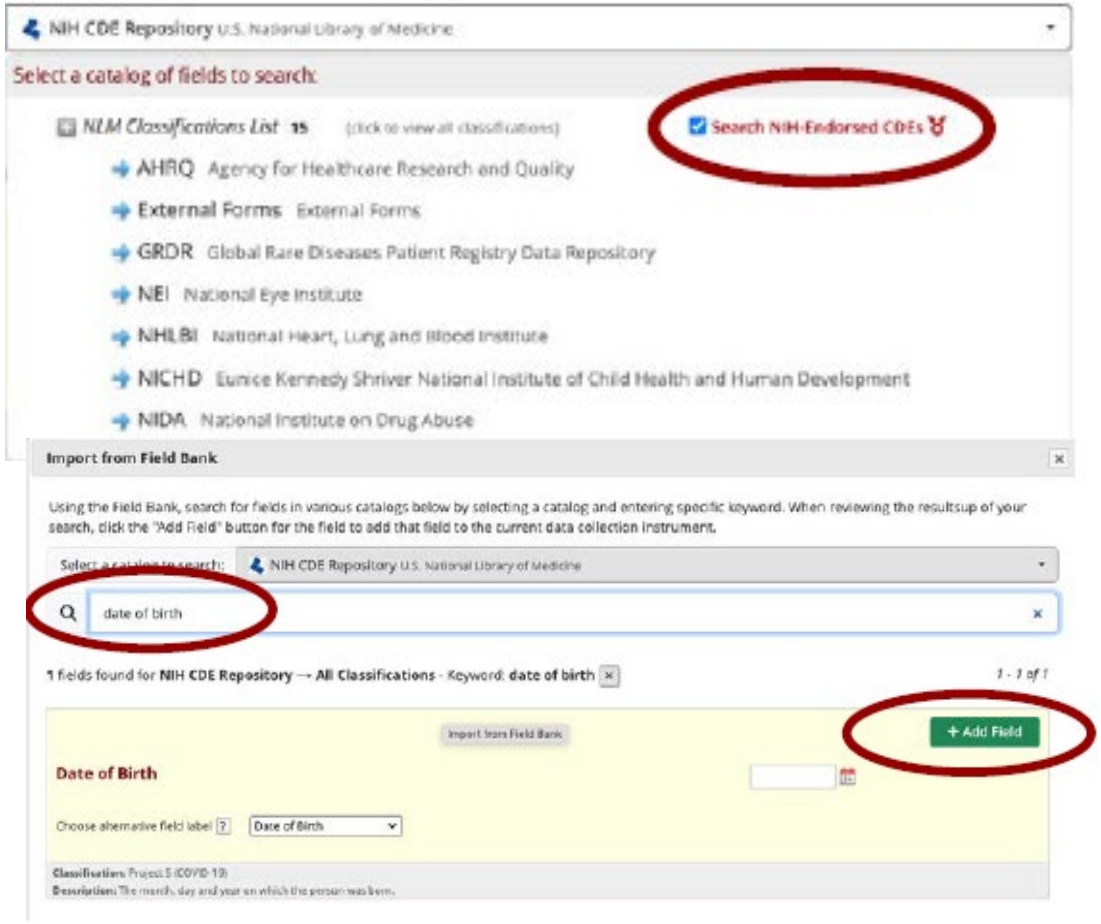
From Research Study Design to *Usable* Data: Take Home on "Using REDCap for FAIR data"

You have pre-vetted starting point* because of *Intramural Research Program REDCap

Inherent capacity to semantically search up COMMON DATA ELEMENTS *within* REDCap

- *Some that—since 2020 NIH-wide criteria were set—are NIH-ENDORSED!*
- Example of demographics:





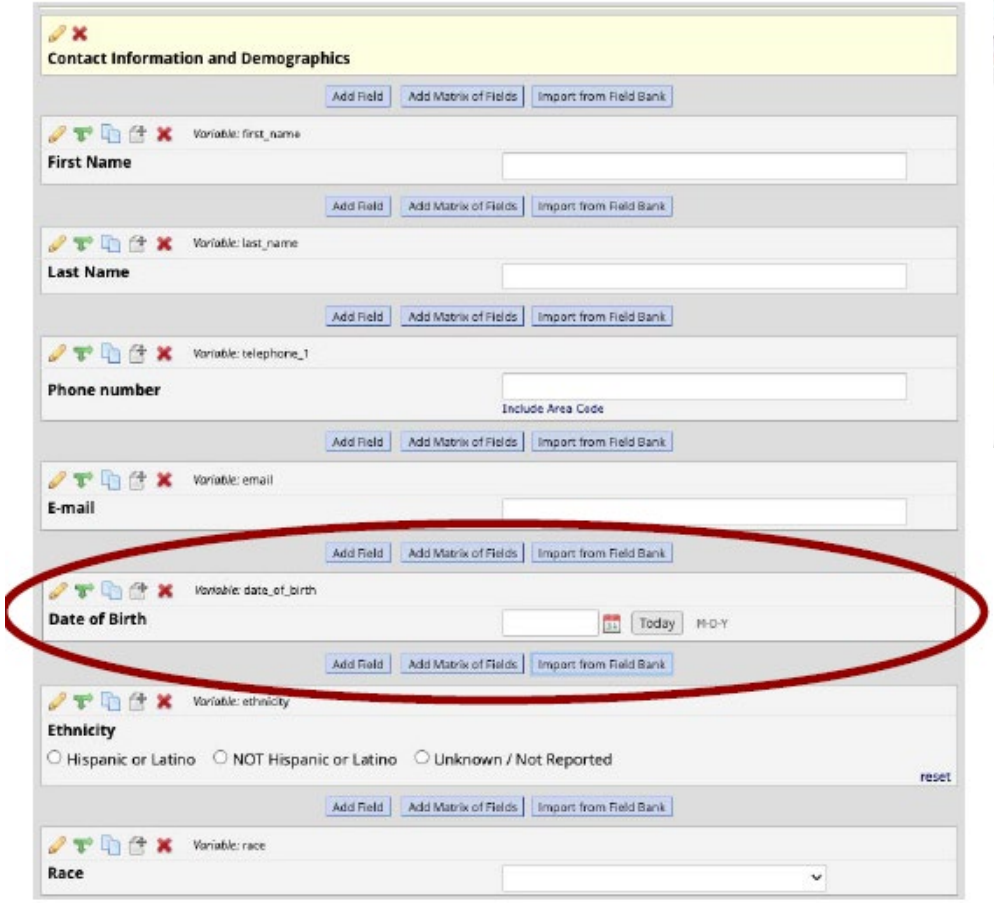
REDCap
Research Electronic Data Capture

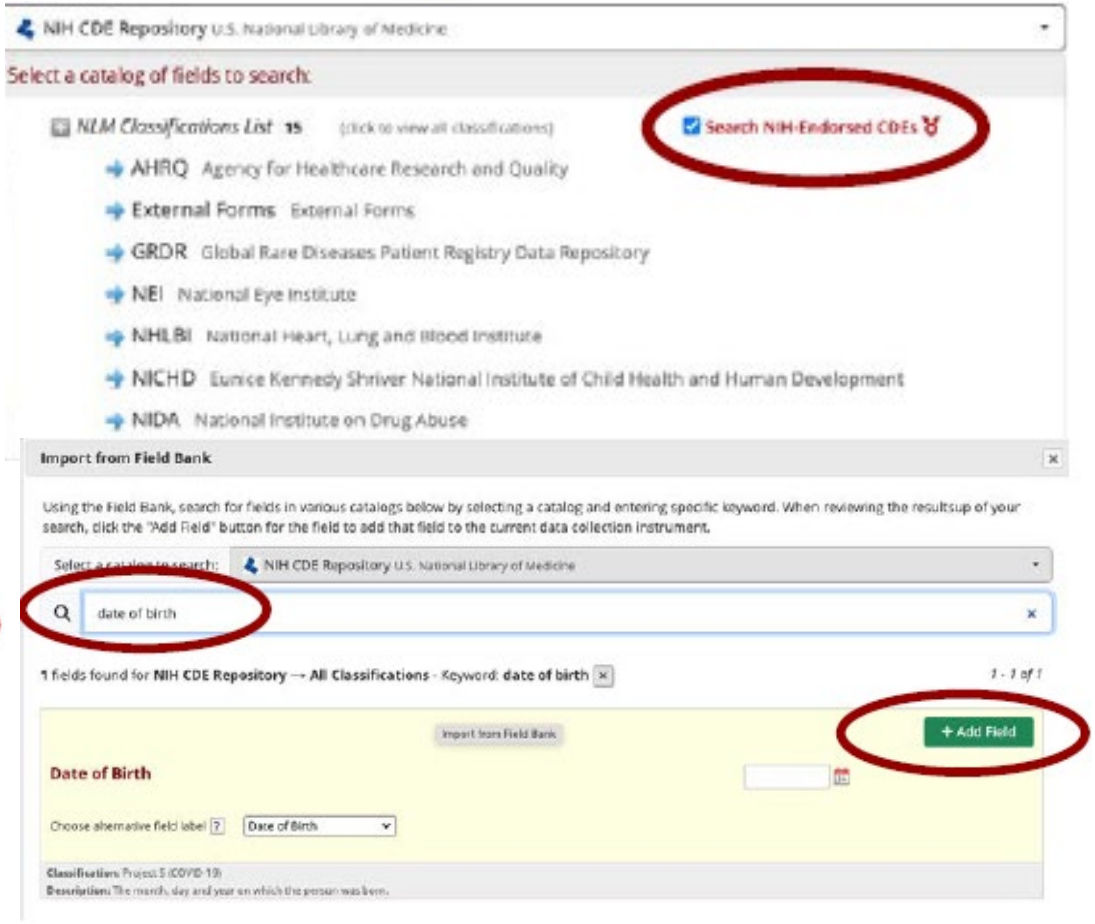
From Research Study Design to *Usable* Data: Take Home on "Using REDCap for FAIR data"

You have pre-vetted starting point* because of *Intramural Research Program REDCap

Inherent capacity to semantically search up COMMON DATA ELEMENTS *within* REDCap

- *Some that—since 2020 NIH-wide criteria were set—are NIH-ENDORSED!*
- Demographics:





REDCap
Research Electronic Data Capture

From Research Study Design to *Usable* Data: Take Home on "Using REDCap for FAIR data"

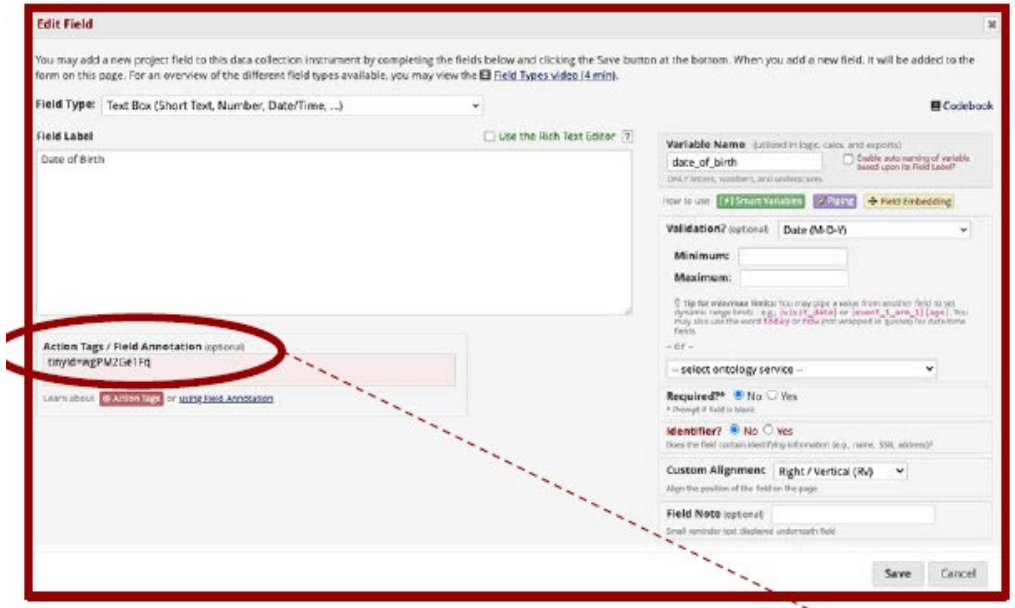
You have pre-vetted starting point* because of *Intramural Research Program REDCap

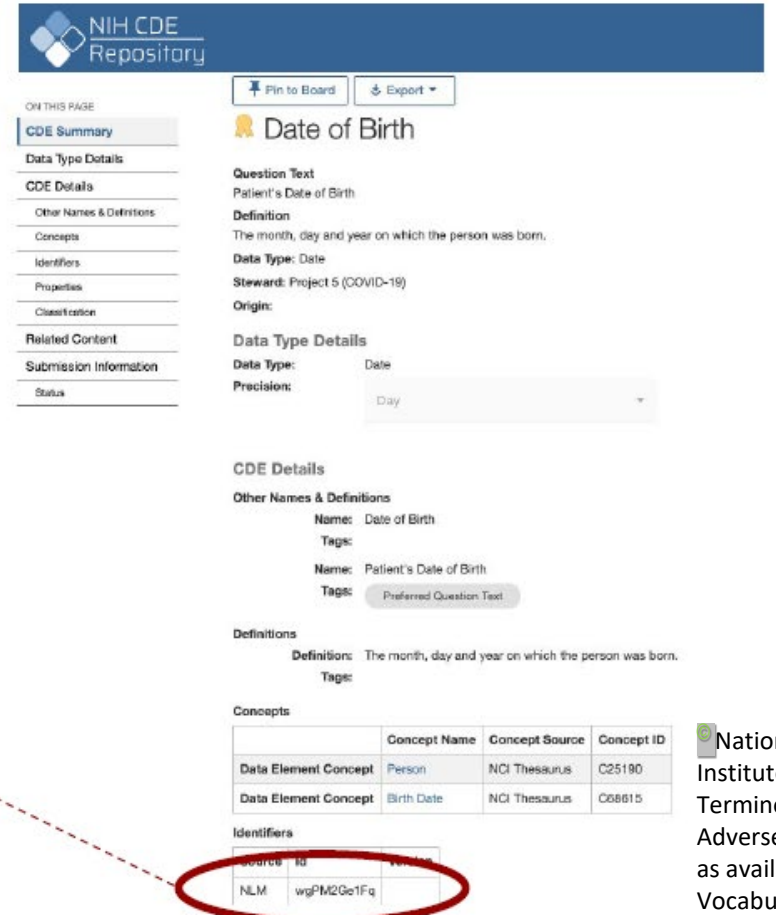
Inherent capacity to semantically search up COMMON DATA ELEMENTS *within* REDCap

- *Some that—since 2020 NIH-wide criteria were set—are NIH-ENDORSED!*
- Demographics:
- *Adverse Events*

CTCAE©

REDCap
Research Electronic Data Capture





In the event of NOT limited time, leading here, skip back to take-homes

*In the case of 'Demographics' be aware of Federal Government changes in capturing Race/Ethnicity as specified in Statistical Policy Directive #15's revision: https://spd15revision.gov/

© National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) as available in Enterprise Vocabulary Services (EVS)