Developing Data Quality and Data Sharing Tools for a Global HIV Research Consortium

Judy Lewis, PhD

Application Developer, Vanderbilt Institute for Clinical and Translational Research, VUMC

Adjoint Assistant Professor, Department of Biomedical Engineering, Vanderbilt University

IeDEA

Harmonist Team at Vanderbilt

Stephany Duda
Principal Investigator

Eva Bascompte Moragas
Hub Lead

Judy Lewis
Toolkit Lead

Jeremy Stephens
Toolkit Developer

Hilary Vansell
Grant Coordinator

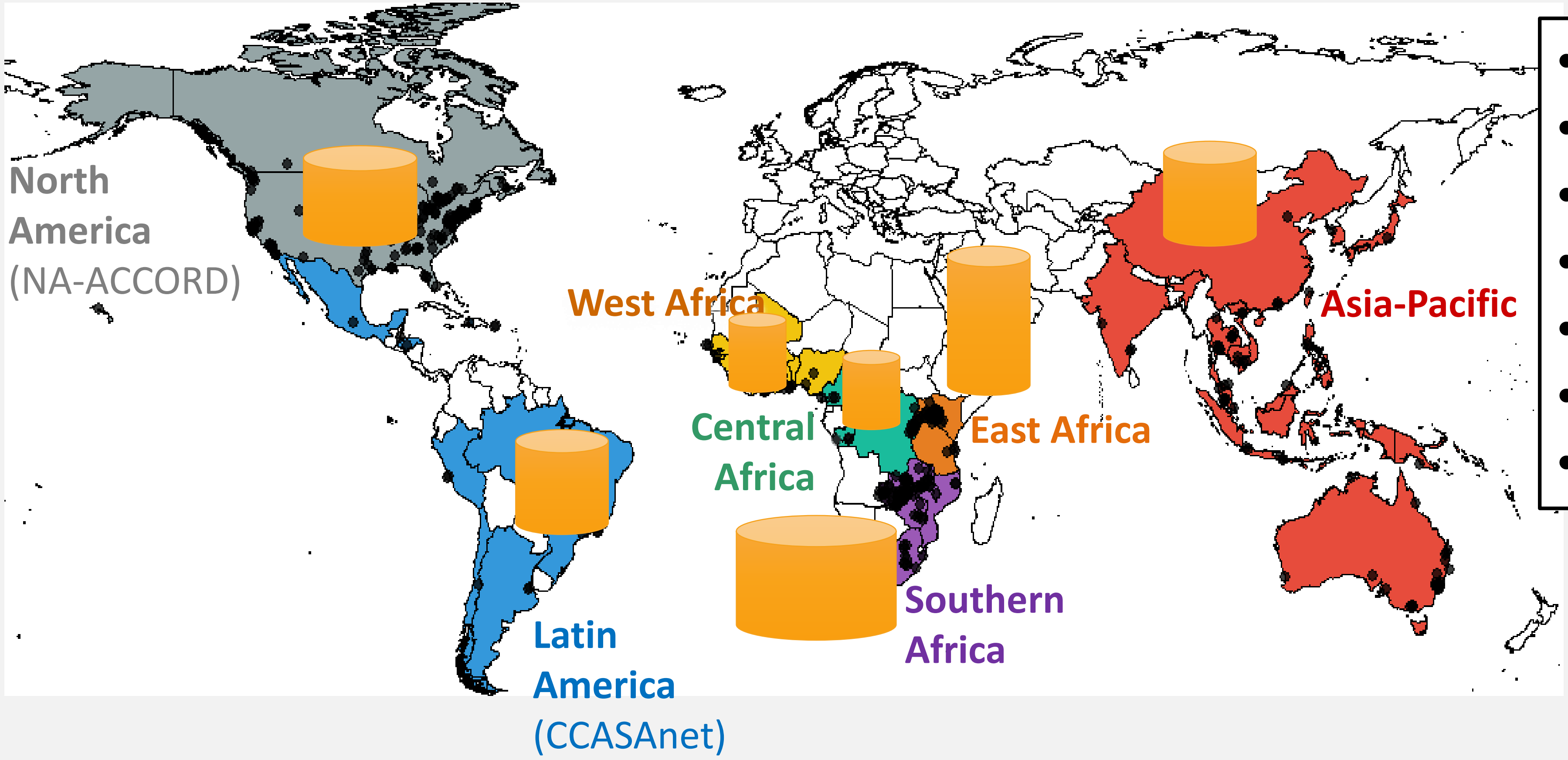
Harmonist: Developing informatics solutions to harmonize observational data in a global research consortium

Today’s Agenda

1. IeDEA research consortium
2. Challenges in IeDEA multiregional data sharing, merging, and analysis
3. Harmonist software tools: design and implementation
4. Example workflow
5. Initial feedback and results
6. Lessons learned

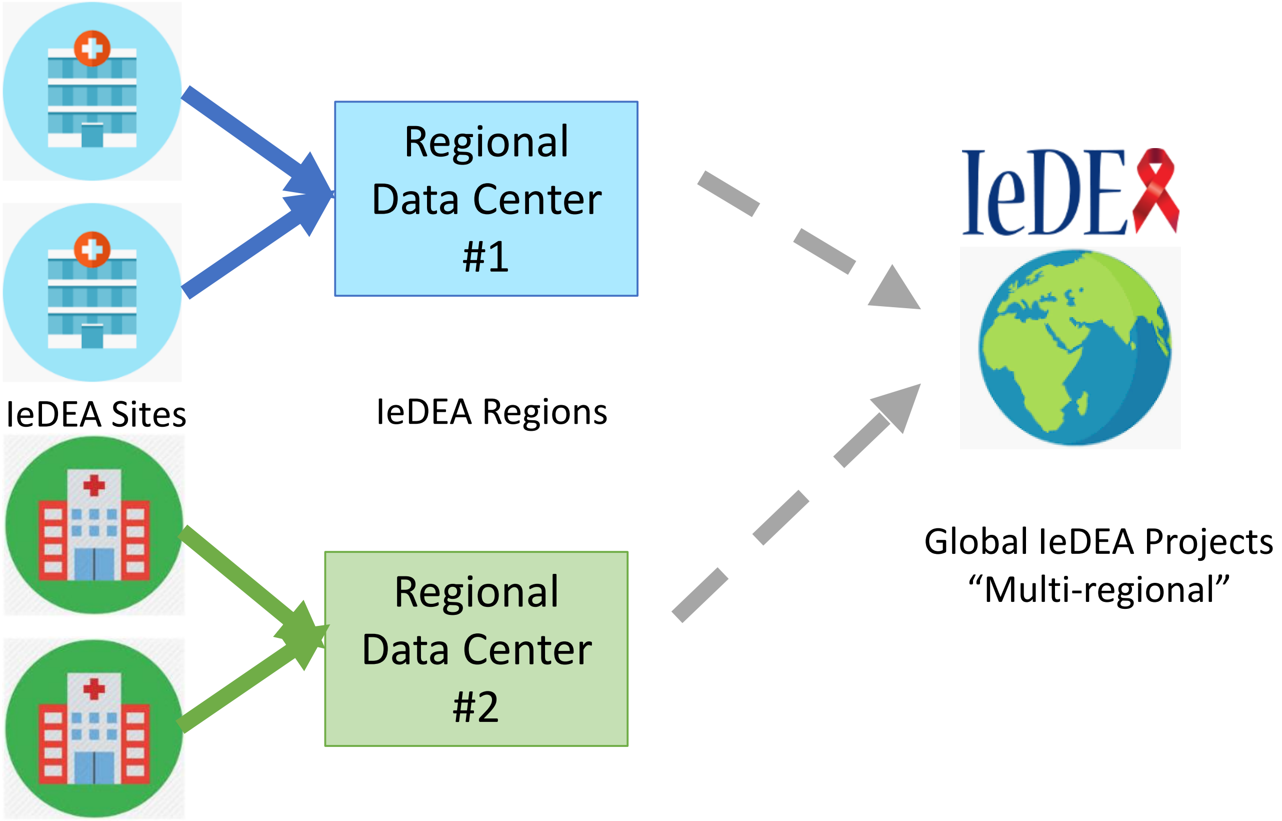
IeDEA

International epidemiologic Databases to Evaluate AIDS



- Established in 2005
- Funded by NIH
- 7 regions
- 46 countries
- 400+ clinics
- ~2 million patients
- 100’s of publications

Flow of IeDEA Data



In IeDEA

- Sites generate the data.
- Regional Data Centers combine all the data from one region.
- Researchers can get data from multiple regions for a global IeDEA project.

Data Considerations

- Data from every clinic can be different.
- Data at every Regional Data Center can be different.

Regional Data Center #1

Regional Data Center #2

- Global IeDEA data are not stored centrally – subsets of the data are merged for specific projects.
- Sites and Regions have the ultimate say in whether their data is included for a specific project.

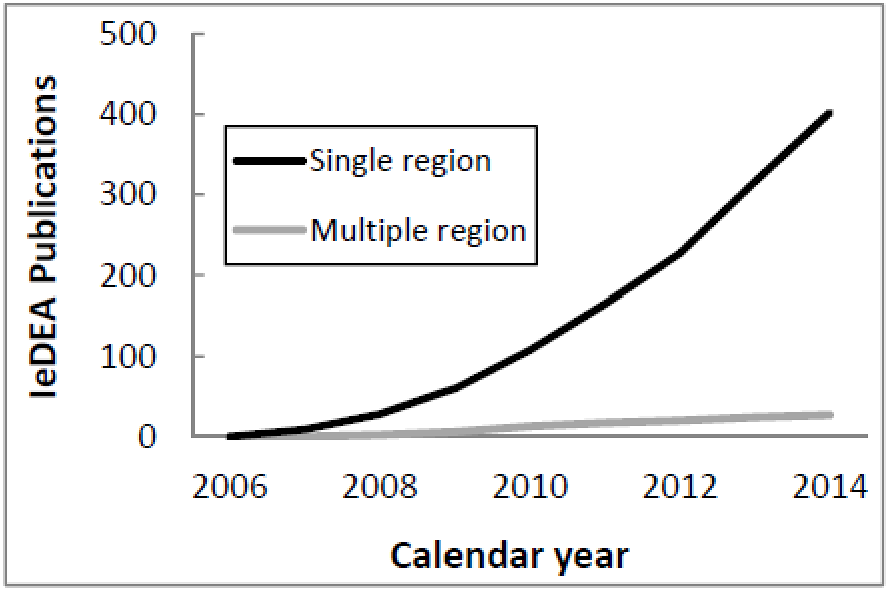
IeDEA

In the Early Days of IeDEA…

- We had no standardized way to share data for global projects.
- Multi-regional projects (projects with 3+ IeDEA regions) were very slow, in part because it was difficult to merge the data.

Cumulative number of IeDEA publications by publication year

(figure from Constantin Yiannoutsos)

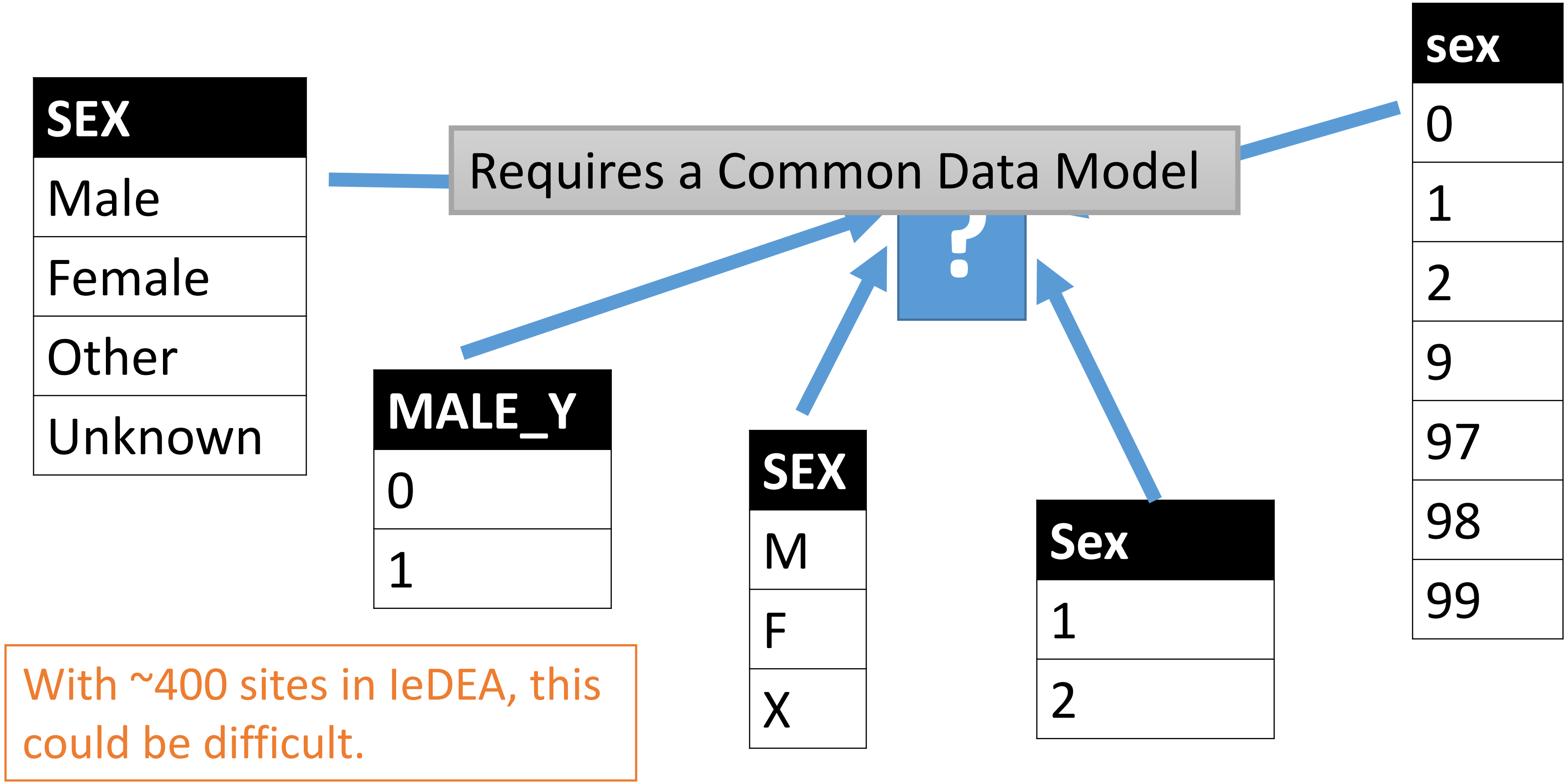


IeDEA Data Harmonization Challenges

- Data from multiple regions must be merged
  - Need **common data model** that can **evolve**, is easy to share and access
- Meaningful research requires quality data
  - Need **data quality checking** algorithms
  - Need **report generation** to summarize dataset quality and characteristics
- Datasets must be transferred from regions to investigators
  - Need secure method for **submitting and receiving datasets**
- Regions must communicate to track requests, submit votes
  - Need **project management** hub
- Computing resources vary across regions and data managers are busy
  - *Need all software tools to require **minimal user resources and maintenance***

Common Data Model

What happens when everyone has a different data format or coding? (ex: sex at birth)



Requires a Common Data Model

SEX
Male
Female
Other
Unknown

MALE_Y
0
1

SEX
M
F
X

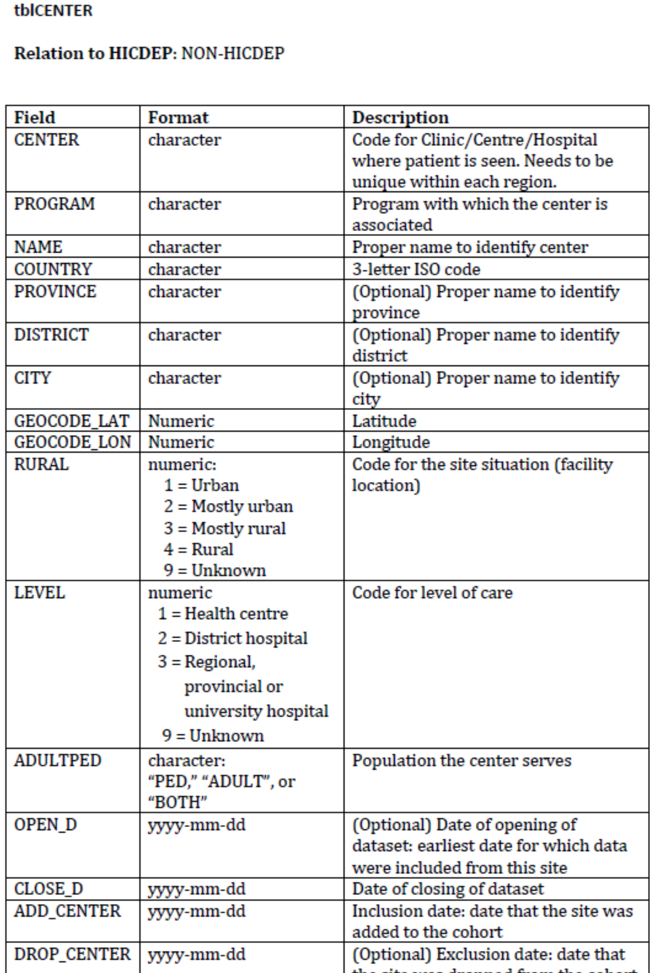
Sex
1
2

sex
0
1
2
9
97
98
99

With ~400 sites in IeDEA, this could be difficult.

IeDEA Data Exchange Standard (DES)

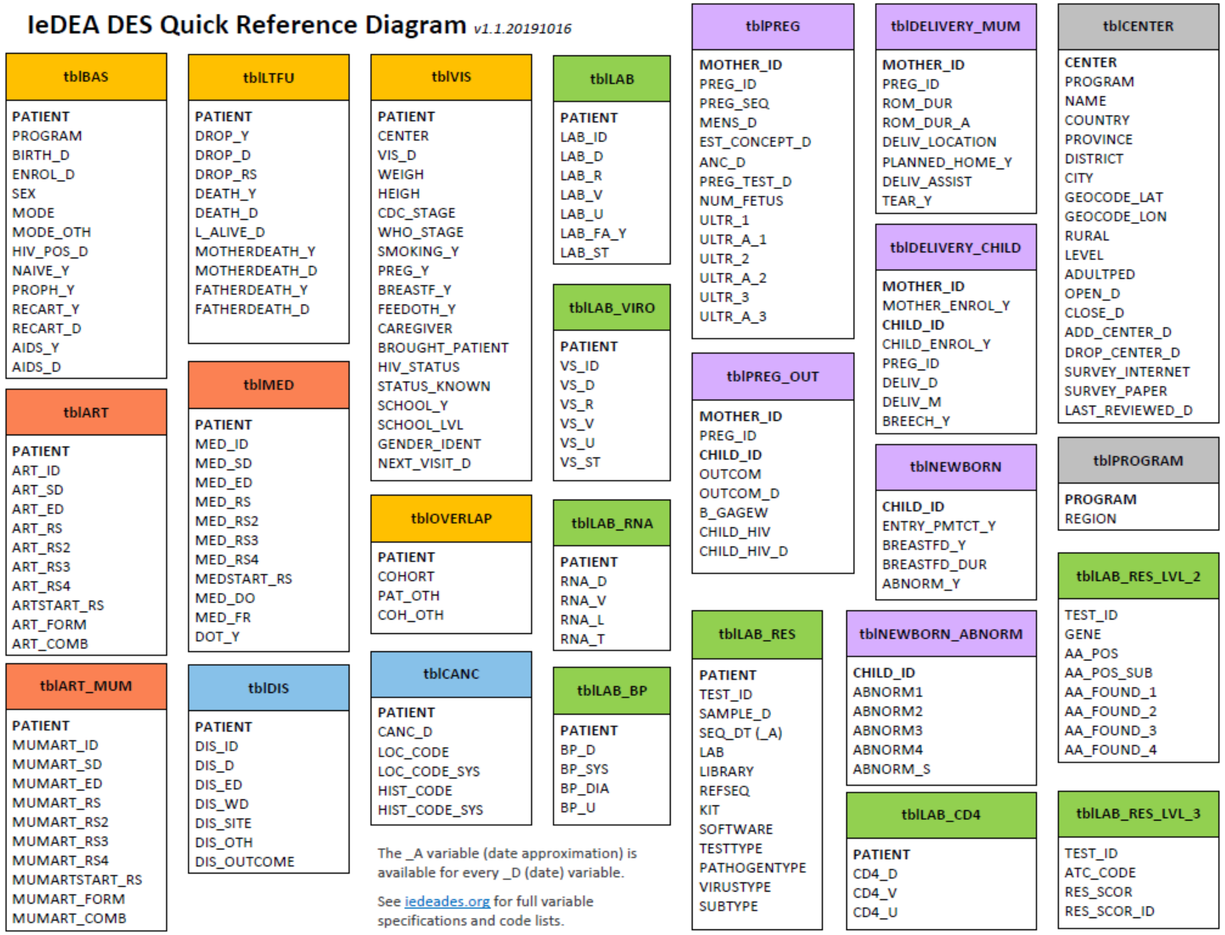
The IeDEA DES defines the variable names, variable definitions, and code lists for data sharing for global IeDEA projects.



tblCENTER

Relation to HICDEP: NON-HICDEP

Field	Format	Description
CENTER	character	Code for Clinic/Centre/Hospital where patient is seen. Needs to be unique within each region.
PROGRAM	character	Program with which the center is associated
NAME	character	Proper name to identify center
COUNTRY	character	3-letter ISO code
PROVINCE	character	(Optional) Proper name to identify province
DISTRICT	character	(Optional) Proper name to identify district
CITY	character	(Optional) Proper name to identify city
GEOCODE_LAT	Numeric	Latitude
GEOCODE_LON	Numeric	Longitude
RURAL	numeric: 1 = Urban 2 = Mostly urban 3 = Mostly rural 4 = Rural 9 = Unknown	Code for the site situation (facility location)
LEVEL	numeric 1 = Health centre 2 = District hospital 3 = Regional, provincial or university hospital 9 = Unknown	Code for level of care
ADULTPED	character: "PED," "ADULT", or "BOTH"	Population the center serves
OPEN_D	yyyy-mm-dd	(Optional) Date of opening of dataset: earliest date for which data were included from this site
CLOSE_D	yyyy-mm-dd	Date of closing of dataset
ADD_CENTER	yyyy-mm-dd	Inclusion date: date that the site was added to the cohort
DROP_CENTER	yyyy-mm-dd	(Optional) Exclusion date: date that



IeDEA DES Quick Reference Diagram *v1.1.20191016*

tblBAS
PATIENT
PROGRAM
BIRTH_D
ENROL_D
SEX
MODE
MODE_OTH
HIV_POS_D
NAIVE_Y
PROPH_Y
RECART_Y
RECART_D
AIDS_Y
AIDS_D

tblART
PATIENT
ART_ID
ART_SD
ART_ED
ART_RS
ART_RS2
ART_RS3
ART_RS4
ARTSTART_RS
ART_FORM
ART_COMB

tblART_MUM
PATIENT
MUMART_ID
MUMART_SD
MUMART_ED
MUMART_RS
MUMART_RS2
MUMART_RS3
MUMART_RS4
MUMARTSTART_RS
MUMART_FORM
MUMART_COMB

tblLTFU
PATIENT
DROP_Y
DROP_D
DROP_RS
DEATH_Y
DEATH_D
L_ALIVE_D
MOTHERDEATH_Y
MOTHERDEATH_D
FATHERDEATH_Y
FATHERDEATH_D

tblMED
PATIENT
MED_ID
MED_SD
MED_ED
MED_RS
MED_RS2
MED_RS3
MED_RS4
MEDSTART_RS
MED_DO
MED_FR
DOT_Y

tblDIS
PATIENT
DIS_ID
DIS_D
DIS_ED
DIS_WD
DIS_SITE
DIS_OTH
DIS_OUTCOME

tblVIS
PATIENT
CENTER
VIS_D
WEIGH
HEIGH
CDC_STAGE
WHO_STAGE
SMOKING_Y
PREG_Y
BREASTF_Y
FEEDOTH_Y
CAREGIVER
BROUGHT_PATIENT
HIV_STATUS
STATUS_KNOWN
SCHOOL_Y
SCHOOL_LVL
GENDER_IDENT
NEXT_VISIT_D

tblOVERLAP
PATIENT
COHORT
PAT_OTH
COH_OTH

tblCANC
PATIENT
CANC_D
LOC_CODE
LOC_CODE_SYS
HIST_CODE
HIST_CODE_SYS

The _A variable (date approximation) is available for every _D (date) variable.

See iedeades.org for full variable specifications and code lists.

tblLAB
PATIENT
LAB_ID
LAB_D
LAB_R
LAB_V
LAB_U
LAB_FA_Y
LAB_ST

tblLAB_VIRO
PATIENT
VS_ID
VS_D
VS_R
VS_V
VS_U
VS_ST

tblLAB_RNA
PATIENT
RNA_D
RNA_V
RNA_L
RNA_T

tblLAB_BP
PATIENT
BP_D
BP_SYS
BP_DIA
BP_U

tblPREG
MOTHER_ID
PREG_ID
PREG_SEQ
MENS_D
EST_CONCEPT_D
ANC_D
PREG_TEST_D
NUM_FETUS
ULTR_1
ULTR_A_1
ULTR_2
ULTR_A_2
ULTR_3
ULTR_A_3

tblPREG_OUT
MOTHER_ID
PREG_ID
CHILD_ID
OUTCOM
OUTCOM_D
B_GAGEW
CHILD_HIV
CHILD_HIV_D

tblLAB_RES
PATIENT
TEST_ID
SAMPLE_D
SEQ_DT (_A)
LAB
LIBRARY
REFSEQ
KIT
SOFTWARE
TESTTYPE
PATHOGENTYPE
VIRUSTYPE
SUBTYPE

tblDELIVERY_MUM
MOTHER_ID
PREG_ID
ROM_DUR
ROM_DUR_A
DELIV_LOCATION
PLANNED_HOME_Y
DELIV_ASSIST
TEAR_Y

tblDELIVERY_CHILD
MOTHER_ID
MOTHER_ENROL_Y
CHILD_ID
CHILD_ENROL_Y
PREG_ID
DELIV_D
DELIV_M
BREECH_Y

tblNEWBORN
CHILD_ID
ENTRY_PMTCT_Y
BREASTFD_Y
BREASTFD_DUR
ABNORM_Y

tblNEWBORN_ABNORM
CHILD_ID
ABNORM1
ABNORM2
ABNORM3
ABNORM4
ABNORM_S

tblLAB_CD4
PATIENT
CD4_D
CD4_V
CD4_U

tblCENTER
CENTER
PROGRAM
NAME
COUNTRY
PROVINCE
DISTRICT
CITY
GEOCODE_LAT
GEOCODE_LON
RURAL
LEVEL
ADULTPED
OPEN_D
CLOSE_D
ADD_CENTER_D
DROP_CENTER_D
SURVEY_INTERNET
SURVEY_PAPER
LAST_REVIEWED_D

tblPROGRAM
PROGRAM
REGION

tblLAB_RES_LVL_2
TEST_ID
GENE
AA_POS
AA_POS_SUB
AA_FOUND_1
AA_FOUND_2
AA_FOUND_3
AA_FOUND_4

tblLAB_RES_LVL_3
TEST_ID
ATC_CODE
RES_SCOR
RES_SCOR_ID

DES Growth Over Time

Change from 2015 to 2019

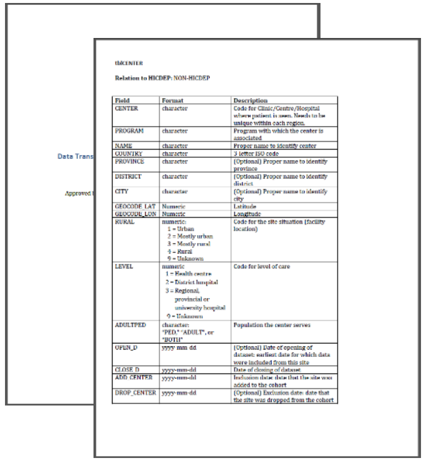
	IeDEA DES Version		
DES Feature	**2015**	**2017**	**2019**
Data Tables	9	25	29
Variables	60	215	269

New variables are related to pregnancy, mental health, substance use, hospitalizations, diagnoses, etc.

We plan to work on additional variable types (e.g., TB, cervical cancer) in 2020.

Maintaining the IeDEA DES

- Challenges with MS Word documents
  - Multiple versions, potentially conflicting edits
  - Hard to find latest version in files, email
  - Single copy is not group editable
  - Not machine-readable



Data Trans

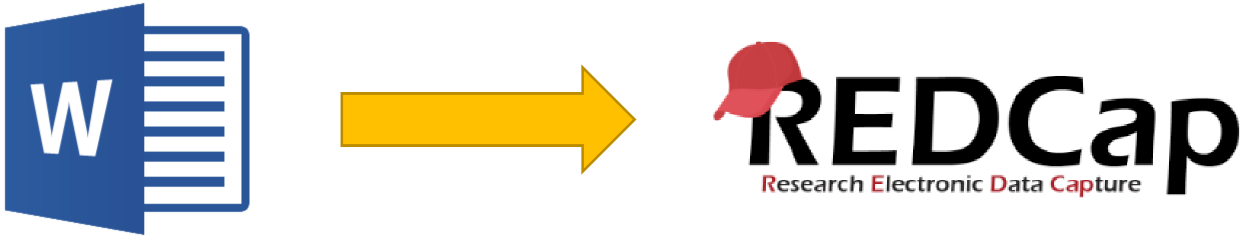
Approved

tblCENTER

Relation to HICDEP: NON-HICDEP

Field	Format	Description
CENTER	character	Code for Clinic/Centre/Hospital where patient is seen. Needs to be unique within each region.
PROGRAM	character	Program with which the center is associated
NAME	character	Proper name to identify center
COUNTRY	character	3 letter ISO code
PROVINCE	character	(Optional) Proper name to identify province
DISTRICT	character	(Optional) Proper name to identify district
CITY	character	(Optional) Proper name to identify city
GEOCODE_LAT	Numeric	Latitude
GEOCODE_LON	Numeric	Longitude
RURAL	numeric: 1 = Urban 2 = Mostly urban 3 = Mostly rural 4 = Rural 9 = Unknown	Code for the site situation (facility location)
LEVEL	numeric 1 = Health centre 2 = District hospital 3 = Regional, provincial or university hospital 9 = Unknown	Code for level of care
ADULTPED	character: "PED," "ADULT", or "BOTH"	Population the center serves
OPEN_D	yyyy mm dd	(Optional) Date of opening of dataset: earliest date for which data were included from this site
CLOSE_D	yyyy-mm-dd	Date of closing of dataset
ADD_CENTER	yyyy-mm-dd	Inclusion date: date that the site was added to the cohort
DROP_CENTER	yyyy-mm-dd	(Optional) Exclusion date: date that the site was dropped from the cohort

- Needed a machine-readable solution that was easy to edit and didn't require technical training.
- **Solution:** Use REDCap to create human-readable forms that produce machine-readable structures

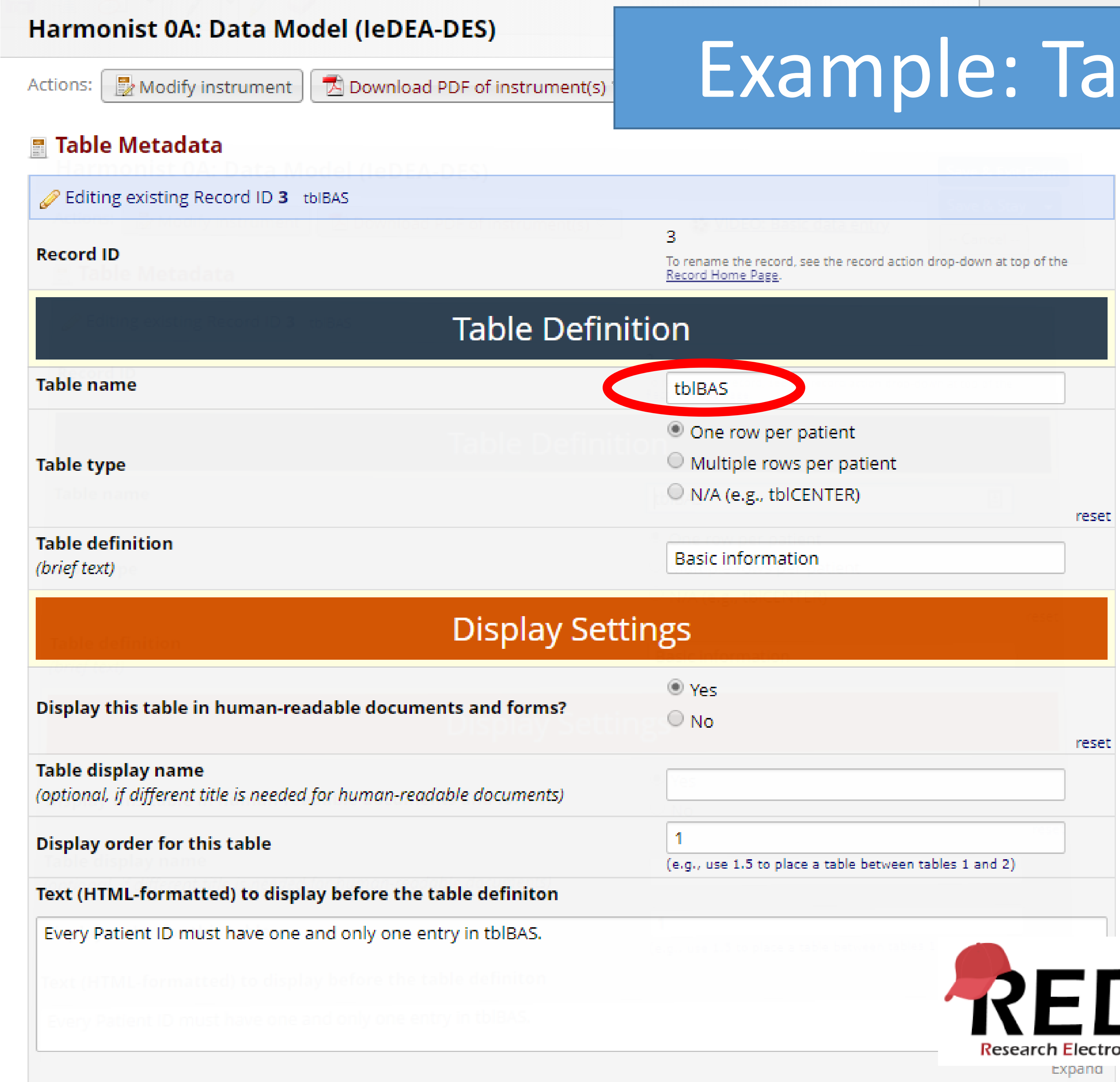


Representing the IeDEA DES in REDCap

To represent the DES in REDCap, we designed three data entry forms:

1. Information about Tables (e.g, demographics, visits, labs, meds)
2. Information about Variables
3. Information about Code Lists

Example: Tables



Harmonist 0A: Data Model (IeDEA-DES)

Actions: Modify instrument | Download PDF of instrument(s)

Table Metadata

Editing existing Record ID **3** tblBAS

Record ID	3 To rename the record, see the record action drop-down at top of the Record Home Page.

Table Definition

Table name	tblBAS
Table type	● One row per patient ○ Multiple rows per patient ○ N/A (e.g., tblCENTER) reset
Table definition *(brief text)*	Basic information

Display Settings

Display this table in human-readable documents and forms?	● Yes ○ No reset
Table display name *(optional, if different title is needed for human-readable documents)*	
Display order for this table	1 (e.g., use 1.5 to place a table between tables 1 and 2)

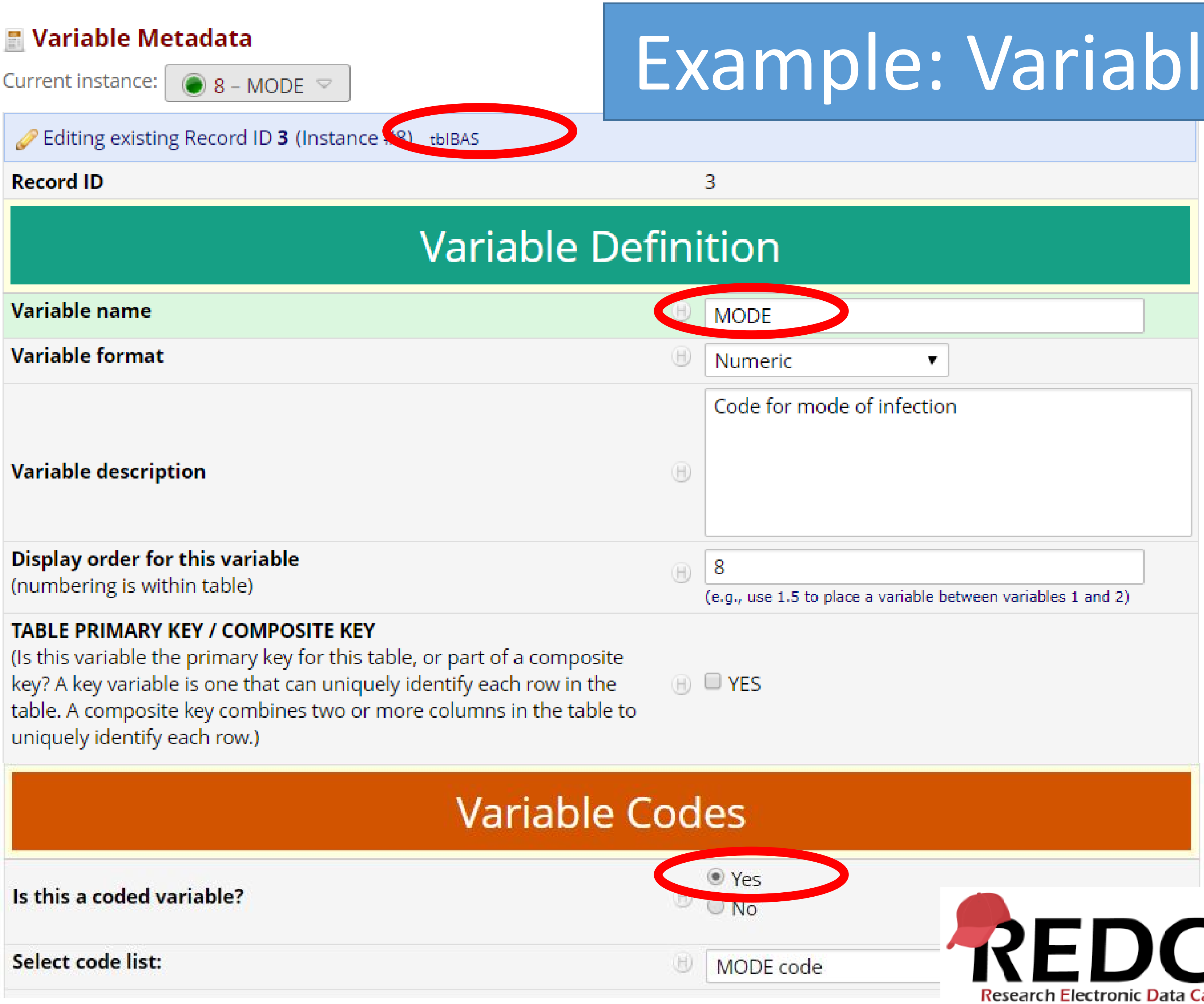
Text (HTML-formatted) to display before the table definiton

Every Patient ID must have one and only one entry in tblBAS.

Expand

REDCap
Research Electronic Data Capture

Example: Variables



Variable Metadata

Current instance: ● 8 – MODE

Editing existing Record ID **3** (Instance #8) tblBAS

Record ID	3

Variable Definition

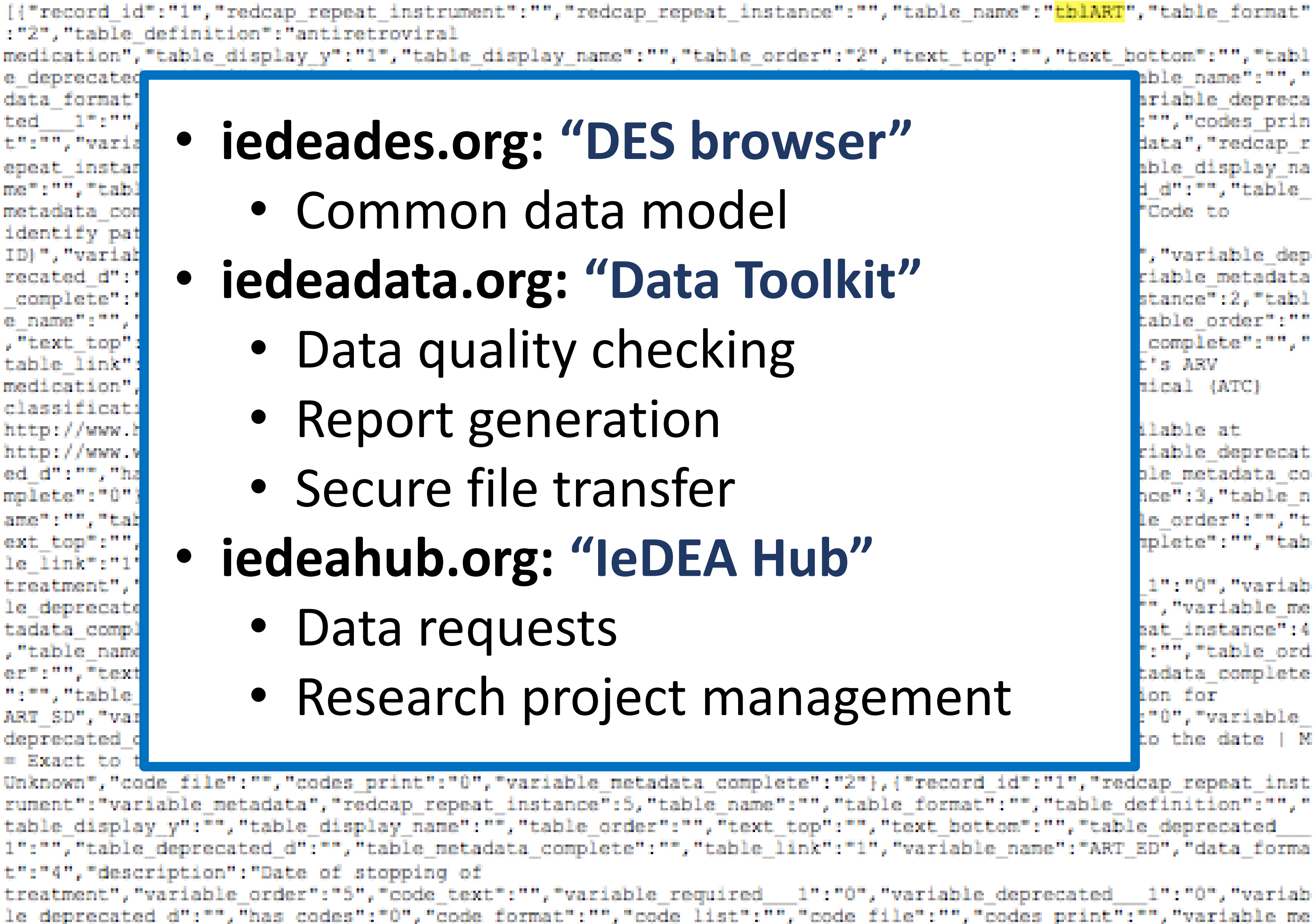
Variable name	MODE
Variable format	Numeric ▾
Variable description	Code for mode of infection
Display order for this variable (numbering is within table)	8 (e.g., use 1.5 to place a variable between variables 1 and 2)
TABLE PRIMARY KEY / COMPOSITE KEY (Is this variable the primary key for this table, or part of a composite key? A key variable is one that can uniquely identify each row in the table. A composite key combines two or more columns in the table to uniquely identify each row.)	□ YES

Variable Codes

Is this a coded variable?	● Yes ○ No
Select code list:	MODE code

REDCap
Research Electronic Data Capture

IeDEA DES in REDCap: Machine-Readable Foundation for Harmonist Tools

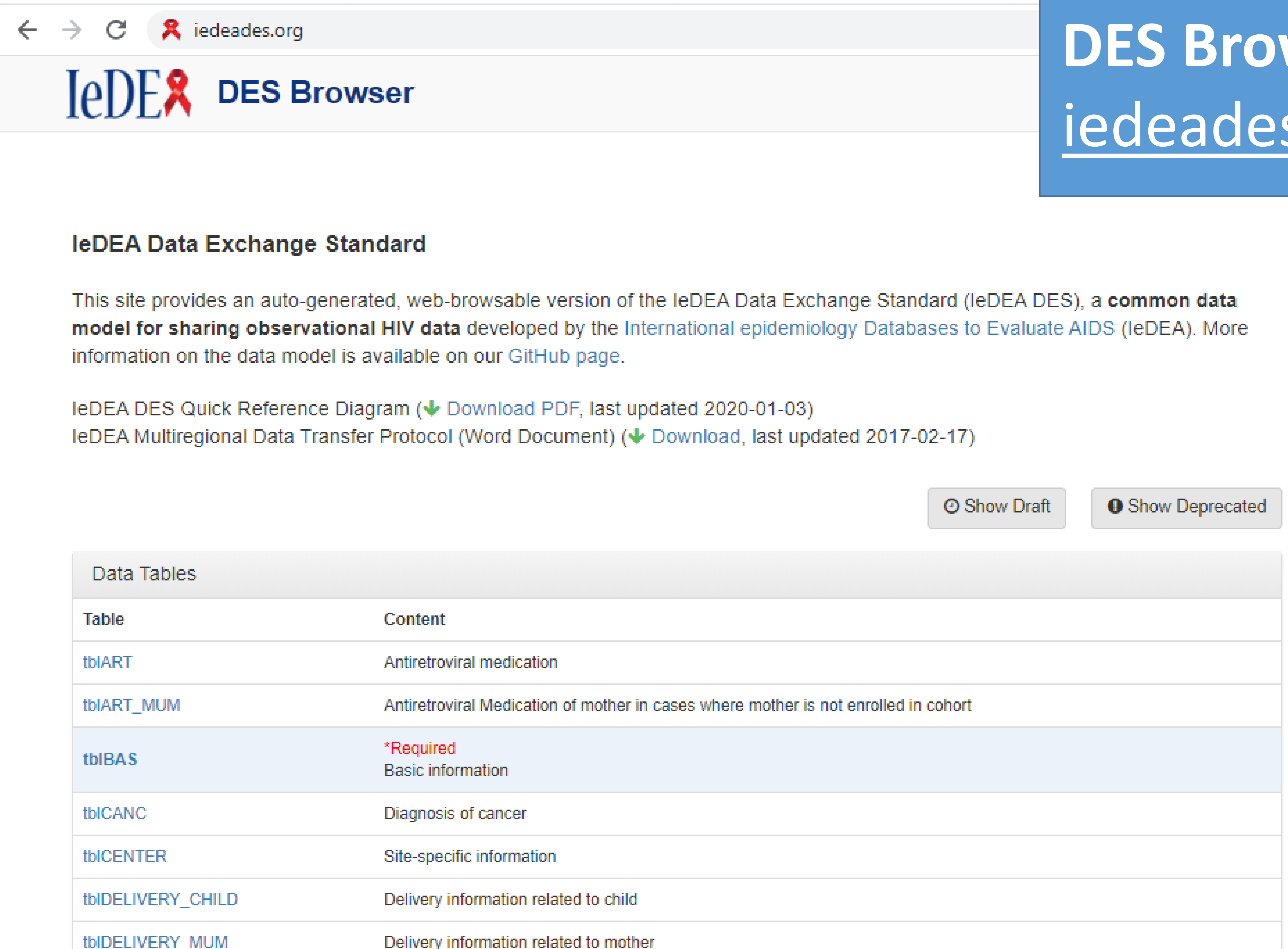


[{"record_id":"1","redcap_repeat_instrument":"","redcap_repeat_instance":"","table_name":"tblART","table_format"
:"2","table_definition":"antiretroviral
medication","table_display_y":"1","table_display_name":"","table_order":"2","text_top":"","text_bottom":"","tabl

Unknown","code_file":"","codes_print":"0","variable_metadata_complete":"2"},{"record_id":"1","redcap_repeat_inst
rument":"variable_metadata","redcap_repeat_instance":5,"table_name":"","table_format":"","table_definition":"","
table_display_y":"","table_display_name":"","table_order":"","text_top":"","text_bottom":"","table_deprecated___
1":"","table_deprecated_d":"","table_metadata_complete":"","table_link":"1","variable_name":"ART_ED","data_forma
t":"4","description":"Date of stopping of
treatment","variable_order":"5","code_text":"","variable_required___1":"0","variable_deprecated___1":"0","variab
le_deprecated_d":"","has_codes":"0","code_format":"","code_list":"","code_file":"","codes_print":"","variable_me

- **iedeades.org: “DES browser”**
  - Common data model
- **iedeadata.org: “Data Toolkit”**
  - Data quality checking
  - Report generation
  - Secure file transfer
- **iedeahub.org: “IeDEA Hub”**
  - Data requests
  - Research project management

DES Browser
iedeades.org



IeDEA Data Exchange Standard

This site provides an auto-generated, web-browsable version of the IeDEA Data Exchange Standard (IeDEA DES), a **common data model for sharing observational HIV data** developed by the International epidemiology Databases to Evaluate AIDS (IeDEA). More information on the data model is available on our GitHub page.

IeDEA DES Quick Reference Diagram (Download PDF, last updated 2020-01-03)
IeDEA Multiregional Data Transfer Protocol (Word Document) (Download, last updated 2017-02-17)

Data Tables

Table	Content
tblART	Antiretroviral medication
tblART_MUM	Antiretroviral Medication of mother in cases where mother is not enrolled in cohort
tblBAS	*Required Basic information
tblCANC	Diagnosis of cancer
tblCENTER	Site-specific information
tblDELIVERY_CHILD	Delivery information related to child
tblDELIVERY_MUM	Delivery information related to mother

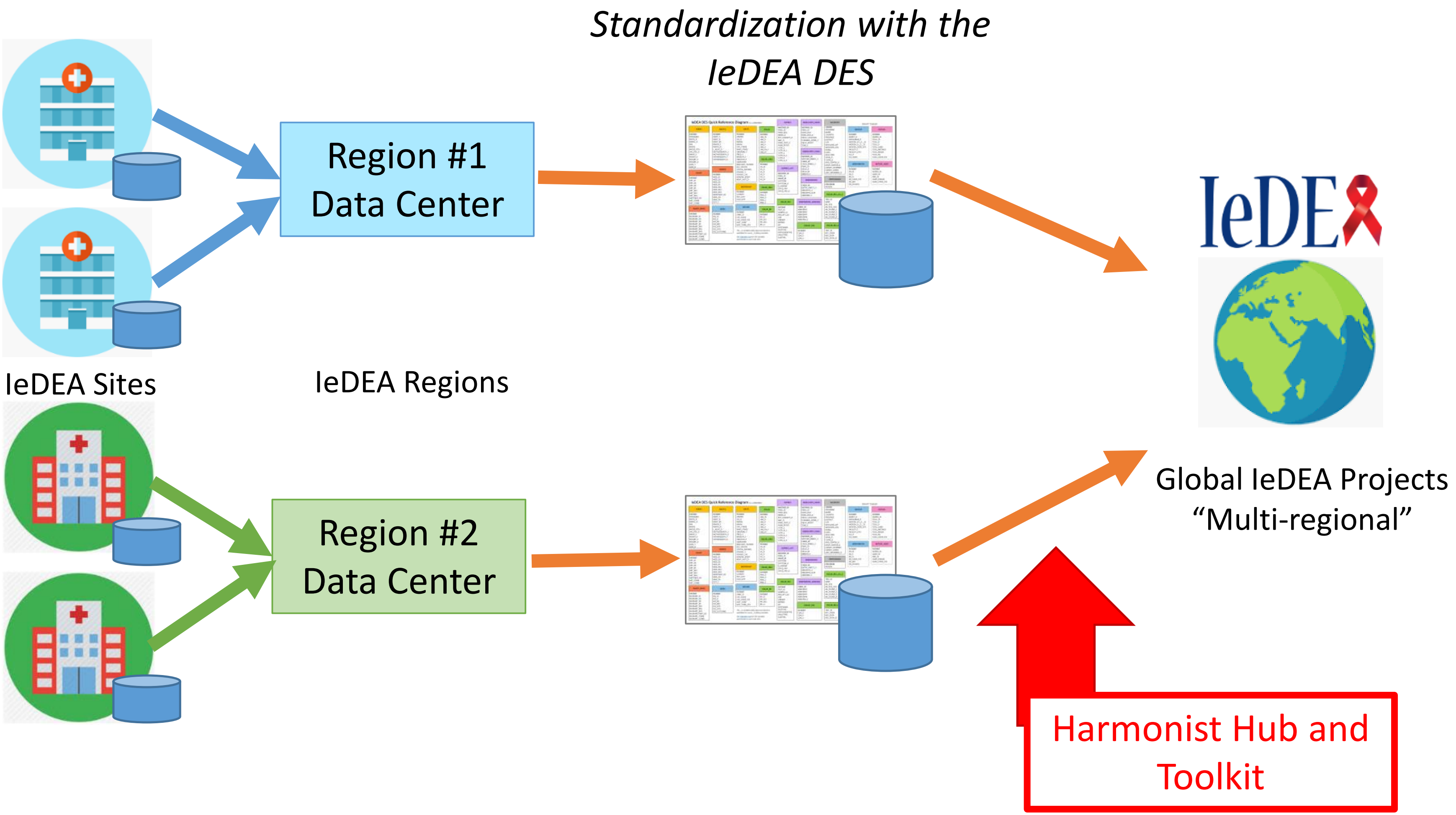
Data Toolkit

IeDEA Harmonist Data Toolkit

- Collaborative project with all seven IeDEA regions
- Web application
- Developed with open source tools (R, Shiny, REDCap)
- Designed to evolve with data exchange standard
- Features:
  - Ensures datasets conform to common data model
  - Performs data quality checks
  - Generates reproducible reports
  - Submits approved datasets to secure cloud storage

Workflow/Hub

Flow of IeDEA Data for Global Projects



Standardization with the IeDEA DES

Region #1 Data Center

Region #2 Data Center

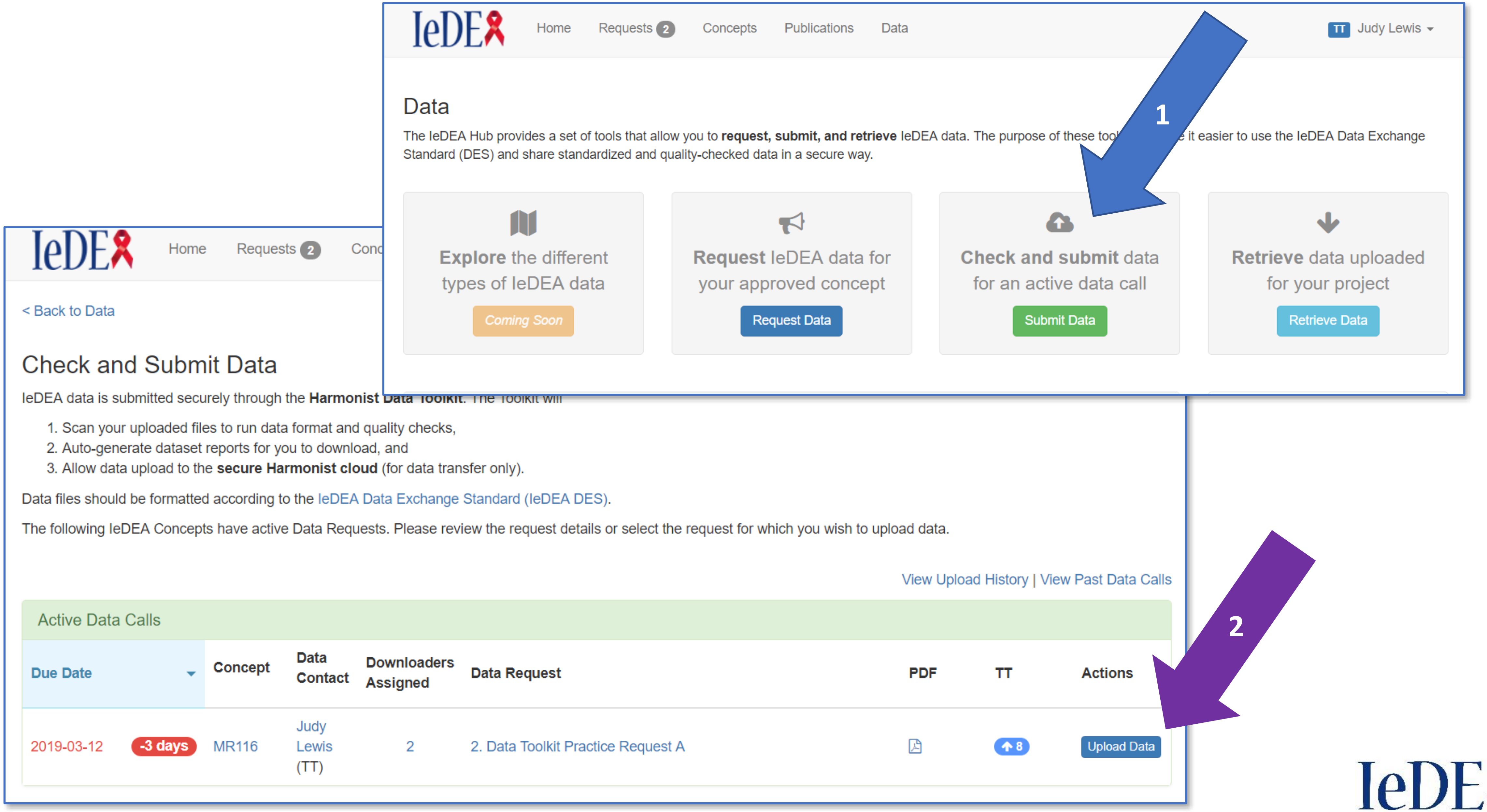
IeDEA Sites

IeDEA Regions

Global IeDEA Projects "Multi-regional"

Harmonist Hub and Toolkit

Workflow Begins in IeDEA Hub



Data

The IeDEA Hub provides a set of tools that allow you to **request, submit, and retrieve** IeDEA data. The purpose of these too e it easier to use the IeDEA Data Exchange Standard (DES) and share standardized and quality-checked data in a secure way.

Explore the different types of IeDEA data

Coming Soon

Request IeDEA data for your approved concept

Request Data

Check and submit data for an active data call

Submit Data

Retrieve data uploaded for your project

Retrieve Data

< Back to Data

Check and Submit Data

IeDEA data is submitted securely through the **Harmonist Data Toolkit**. The Toolkit will

1. Scan your uploaded files to run data format and quality checks,
2. Auto-generate dataset reports for you to download, and
3. Allow data upload to the **secure Harmonist cloud** (for data transfer only).

Data files should be formatted according to the IeDEA Data Exchange Standard (IeDEA DES).

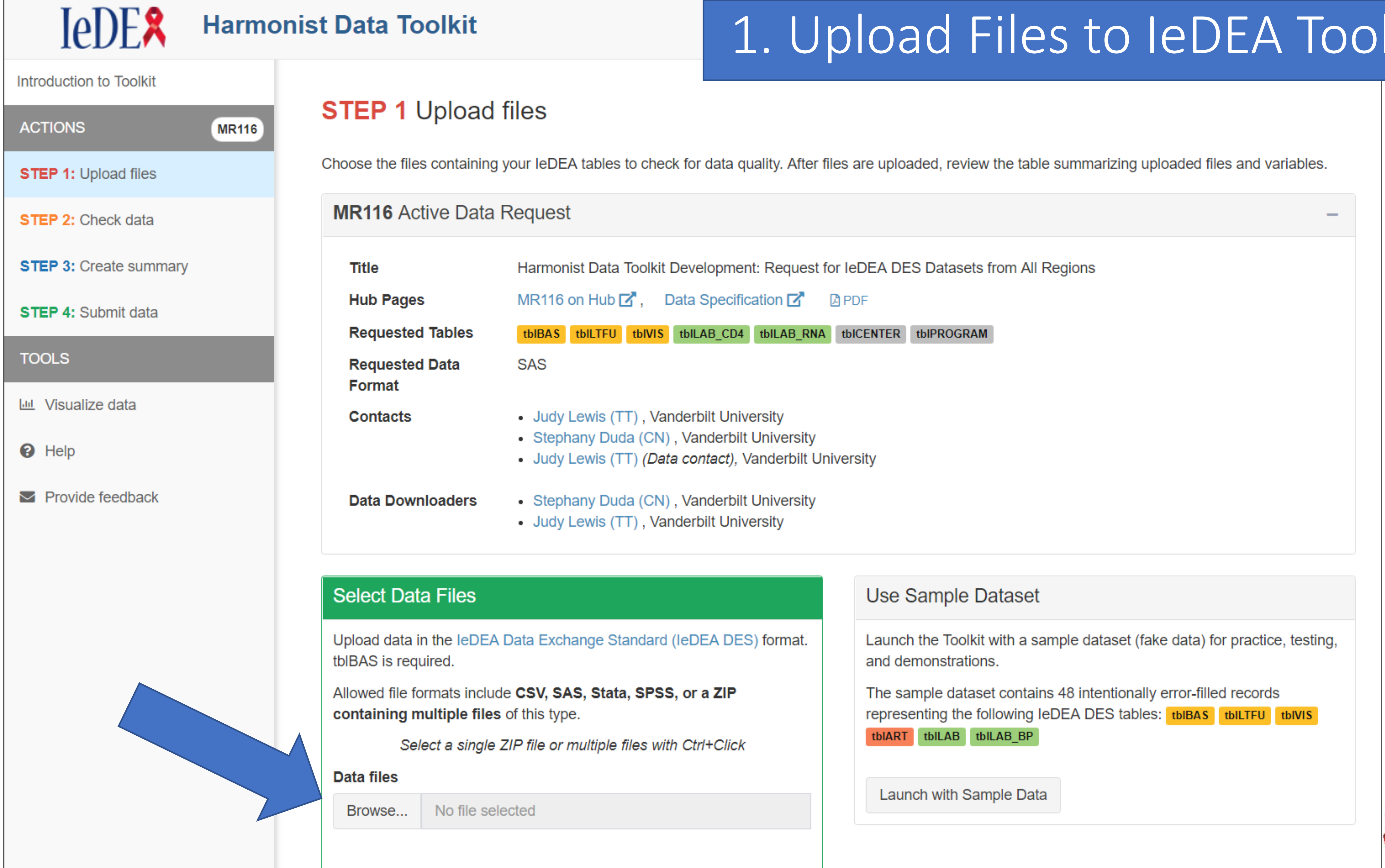
The following IeDEA Concepts have active Data Requests. Please review the request details or select the request for which you wish to upload data.

View Upload History | View Past Data Calls

Active Data Calls

Due Date		Concept	Data Contact	Downloaders Assigned	Data Request	PDF	TT	Actions
2019-03-12	-3 days	MR116	Judy Lewis (TT)	2	2. Data Toolkit Practice Request A		↑ 8	Upload Data

1. Upload Files to IeDEA Toolkit



IeDEA Harmonist Data Toolkit

Introduction to Toolkit

ACTIONS MR116

STEP 1: Upload files

STEP 2: Check data

STEP 3: Create summary

STEP 4: Submit data

TOOLS

Visualize data

Help

Provide feedback

STEP 1 Upload files

Choose the files containing your IeDEA tables to check for data quality. After files are uploaded, review the table summarizing uploaded files and variables.

MR116 Active Data Request

Title	Harmonist Data Toolkit Development: Request for IeDEA DES Datasets from All Regions
Hub Pages	MR116 on Hub , Data Specification PDF
Requested Tables	tblBAS tblLTFU tblVIS tblLAB_CD4 tblLAB_RNA tblCENTER tblPROGRAM
Requested Data Format	SAS
Contacts	• Judy Lewis (TT) , Vanderbilt University • Stephany Duda (CN) , Vanderbilt University • Judy Lewis (TT) *(Data contact)*, Vanderbilt University
Data Downloaders	• Stephany Duda (CN) , Vanderbilt University • Judy Lewis (TT) , Vanderbilt University

Select Data Files

Upload data in the IeDEA Data Exchange Standard (IeDEA DES) format. tblBAS is required.

Allowed file formats include **CSV, SAS, Stata, SPSS, or a ZIP containing multiple files** of this type.

Select a single ZIP file or multiple files with Ctrl+Click

Data files

Browse... No file selected

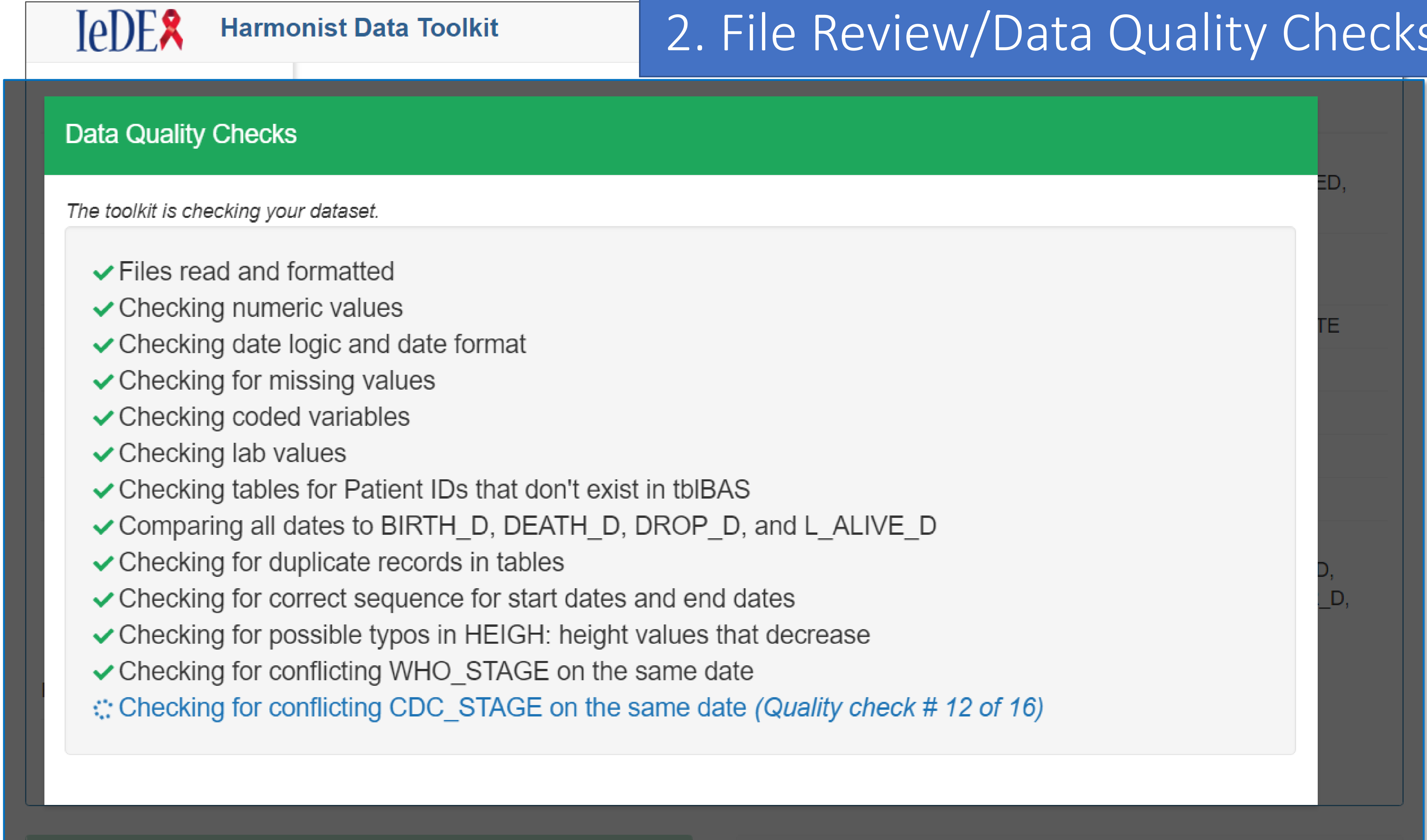
Use Sample Dataset

Launch the Toolkit with a sample dataset (fake data) for practice, testing, and demonstrations.

The sample dataset contains 48 intentionally error-filled records representing the following IeDEA DES tables: tblBAS tblLTFU tblVIS tblART tblLAB tblLAB_BP

Launch with Sample Data

2. File Review/Data Quality Checks

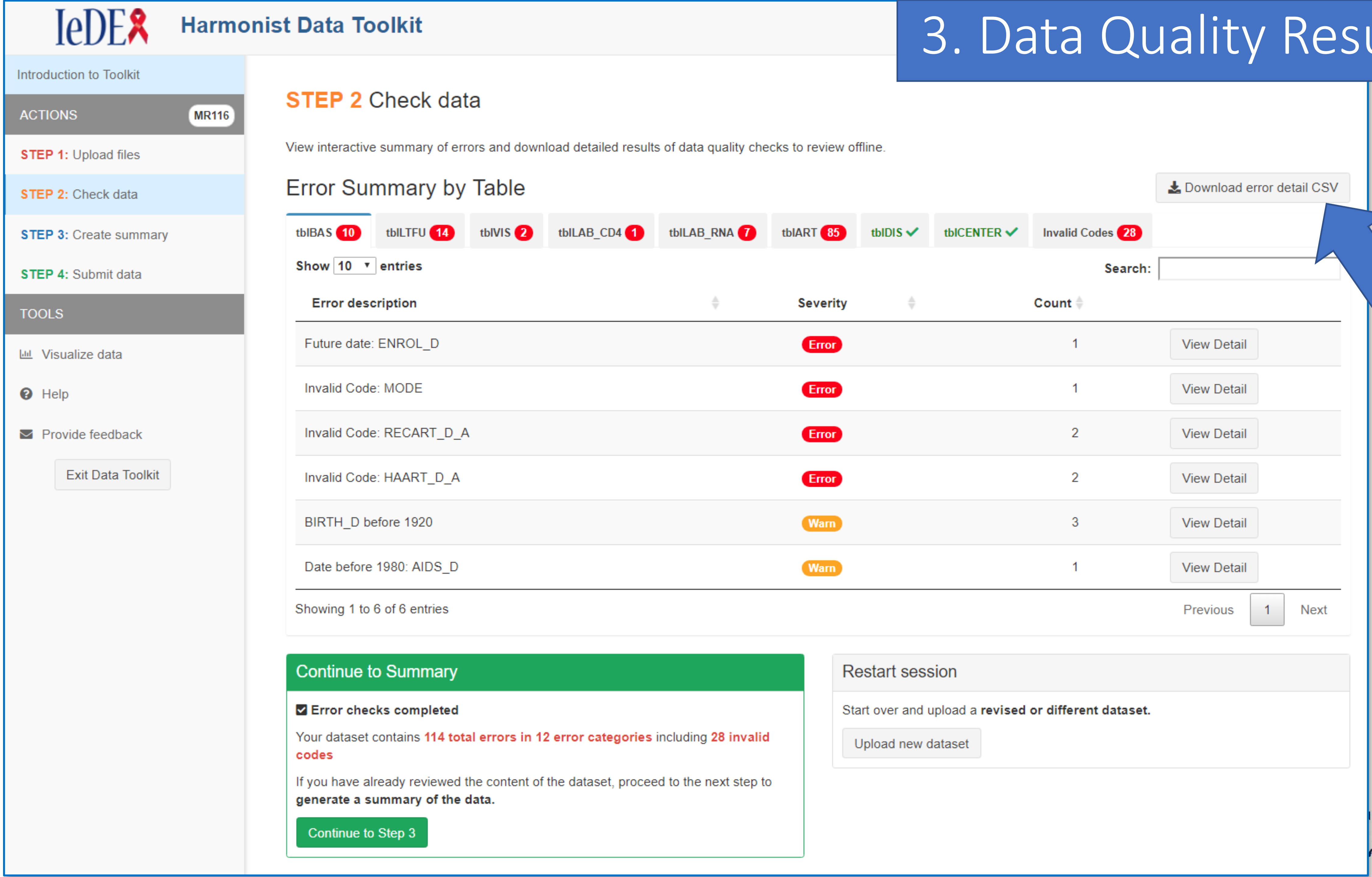


Data Quality Checks

The toolkit is checking your dataset.

- ✓ Files read and formatted
- ✓ Checking numeric values
- ✓ Checking date logic and date format
- ✓ Checking for missing values
- ✓ Checking coded variables
- ✓ Checking lab values
- ✓ Checking tables for Patient IDs that don't exist in tblBAS
- ✓ Comparing all dates to BIRTH_D, DEATH_D, DROP_D, and L_ALIVE_D
- ✓ Checking for duplicate records in tables
- ✓ Checking for correct sequence for start dates and end dates
- ✓ Checking for possible typos in HEIGH: height values that decrease
- ✓ Checking for conflicting WHO_STAGE on the same date
- Checking for conflicting CDC_STAGE on the same date *(Quality check # 12 of 16)*

3. Data Quality Results



STEP 2 Check data

View interactive summary of errors and download detailed results of data quality checks to review offline.

Error Summary by Table

Download error detail CSV

tblBAS 10 | tblLTFU 14 | tblVIS 2 | tblLAB_CD4 1 | tblLAB_RNA 7 | tblART 85 | tblDIS ✓ | tblCENTER ✓ | Invalid Codes 28

Show 10 entries Search:

Error description	Severity	Count	
Future date: ENROL_D	Error	1	View Detail
Invalid Code: MODE	Error	1	View Detail
Invalid Code: RECART_D_A	Error	2	View Detail
Invalid Code: HAART_D_A	Error	2	View Detail
BIRTH_D before 1920	Warn	3	View Detail
Date before 1980: AIDS_D	Warn	1	View Detail

Showing 1 to 6 of 6 entries

Previous 1 Next

Continue to Summary

☑ **Error checks completed**

Your dataset contains **114 total errors in 12 error categories** including **28 invalid codes**

If you have already reviewed the content of the dataset, proceed to the next step to **generate a summary of the data.**

Continue to Step 3

Restart session

Start over and upload a **revised or different dataset.**

Upload new dataset

4. Reproducible Reports

Harmonist Data Toolkit Report

Dataset submitted from IeDEA Region: Harmonist Test

Report date: 2019-03-07

Dataset Summary

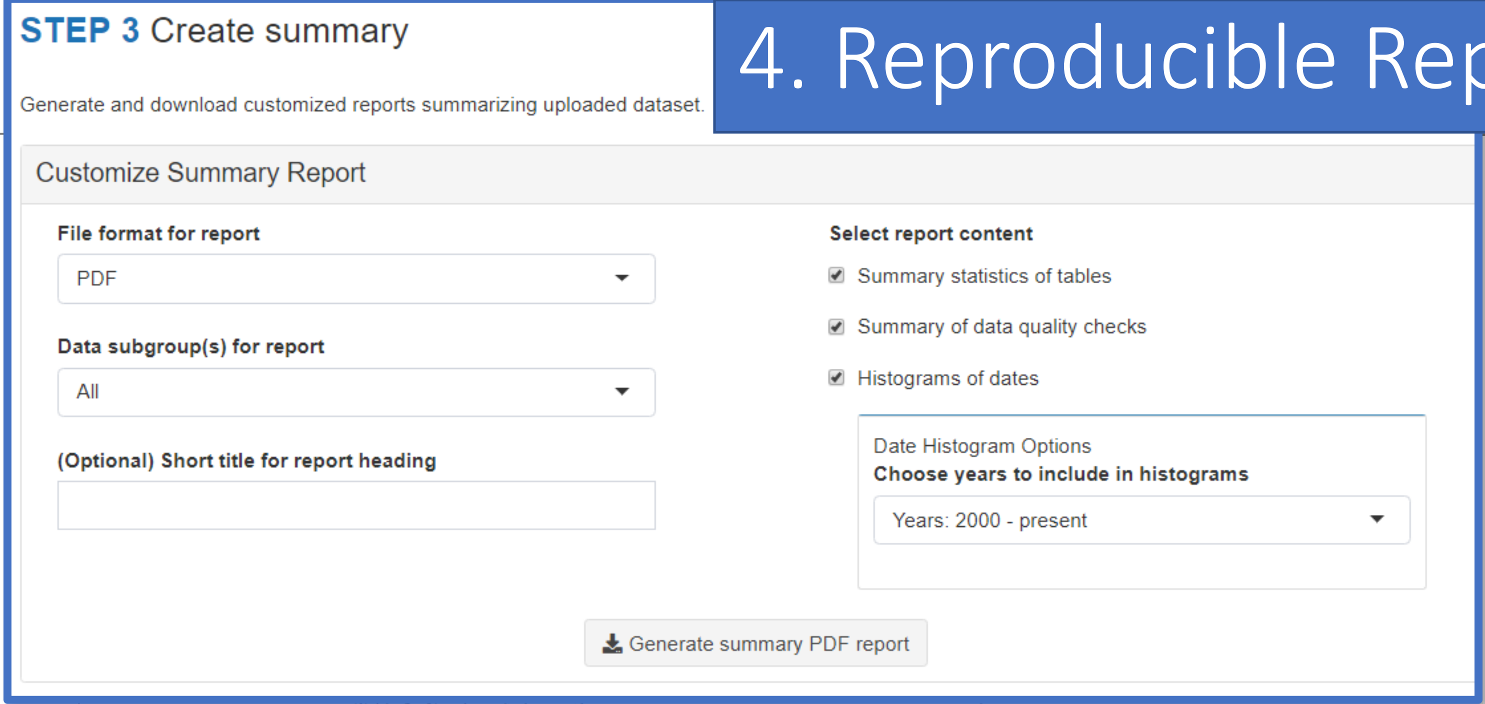
Total number of patients in dataset: 28089

Table 1: Table Summary

			Age at Enrollment					
Table	Records	Patients	0-4	5-9	10-14	15-19	20-24	Adults 25+
tblBAS	28089	28089	0	0	0	971	3342	23776
tblLTFU	28089	28089	0	0	0	971	3342	23776
tblVIS	1084237	26820	0	0	0	963	3270	22587
tblLAB_CD4	298041	27304	0	0	0	945	3273	23086
tblLAB_RNA	159234	15524	0	0	0	471	2010	13043
tblART	140435	25840	0	0	0	889	2971	21980
tblDIS	3750	3215	0	0	0	100	345	2770

Table 2: SITE in Dataset

SITE	Patients	tblLTFU	tblVIS	tblLAB_CD4	tblLAB_RNA	tblART	tblDIS
Hogwarts	594	594	537	557	461	536	36
Hufflepuff	11763	11763	11744	11447	239	11530	1933
Muggleton	5248	5248	5240	5145	4898	4397	650
Potterburg	3208	3208	3208	3168	3116	2723	69
Ravenclaw	1079	1079	0	983	983	904	0
Slytherin	458	458	413	429	308	426	38
Snapetown	4099	4099	4099	3971	3912	3800	345
Wizardville	1640	1640	1579	1604	1607	1524	144



STEP 3 Create summary

Generate and download customized reports summarizing uploaded dataset.

Customize Summary Report

File format for report

PDF

Data subgroup(s) for report

All

(Optional) Short title for report heading

Select report content

☑ Summary statistics of tables

☑ Summary of data quality checks

☑ Histograms of dates

Date Histogram Options

Choose years to include in histograms

Years: 2000 - present

Generate summary PDF report

Table 5: Number of observations per year

Variable	< 2011	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
Enrolled	14728	2517	2527	2859	2391	1439	1332	295	0	0	28088
Visits	411772	115301	117341	124501	118630	120098	71094	5500	0	0	1084237
Deaths	1064	249	232	234	170	216	141	10	0	0	2316
Transfers Out	193	19	19	48	94	145	139	14	0	0	671
Viral Load	59549	15201	15740	16776	16726	17731	15799	1708	0	0	159230
CD4	122521	30414	30781	30465	33865	26465	20152	3378	0	0	298041

Histograms of important dates by SITE

SITE: Hogwarts

(Note: 23 observations before year 2000 and 1 observation after year 2019 are hidden)

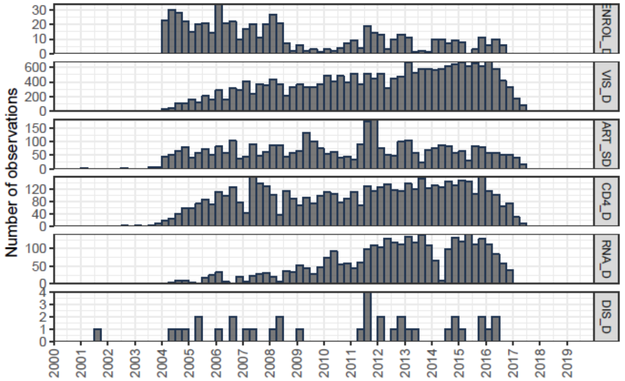
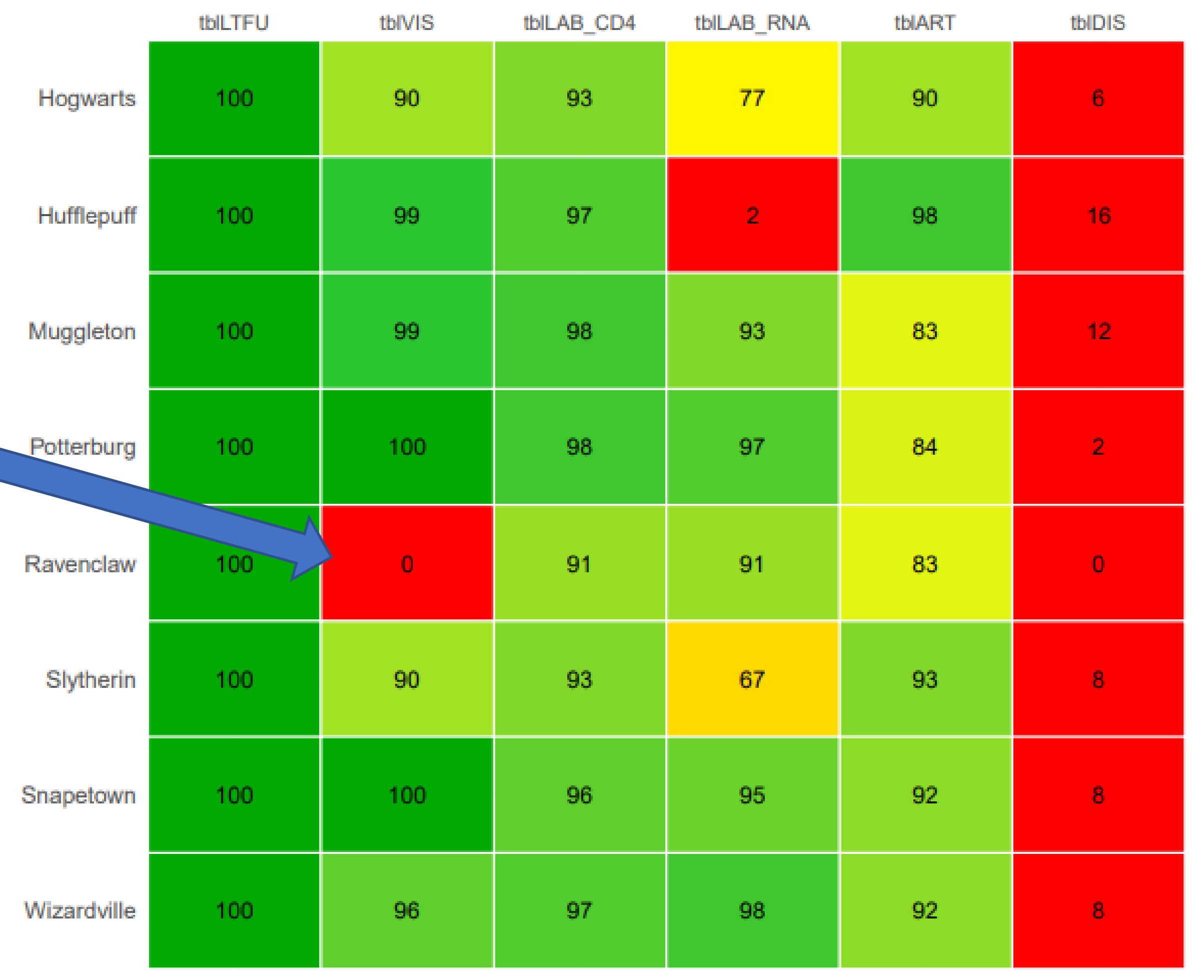


Table 4: Summary statistics from uploaded tables

Value	Count	Percent
Sex		
Male	17453	62.1
Female	10636	37.9
Missing	0	0.0
Deceased		
No	25773	91.8
Yes	2316	8.2
Missing	0	0.0
Treatment Naive at Enrollment		
No	2940	10.5
Yes	25149	89.5
Missing	0	0.0
Receive Antiretroviral Therapy (ART)		
Yes	25840	92.0
Missing	2249	8.0

Details by Site: Spot gaps in data reporting



	tblLTFU	tblVIS	tblLAB_CD4	tblLAB_RNA	tblART	tblDIS
Hogwarts	100	90	93	77	90	6
Hufflepuff	100	99	97	2	98	16
Muggleton	100	99	98	93	83	12
Potterburg	100	100	98	97	84	2
Ravenclaw	100	0	91	91	83	0
Slytherin	100	90	93	67	93	8
Snapetown	100	100	96	95	92	8
Wizardville	100	96	97	98	92	8

Figure 1: Percent of Patients from tblBAS Included

Data Quality Summary

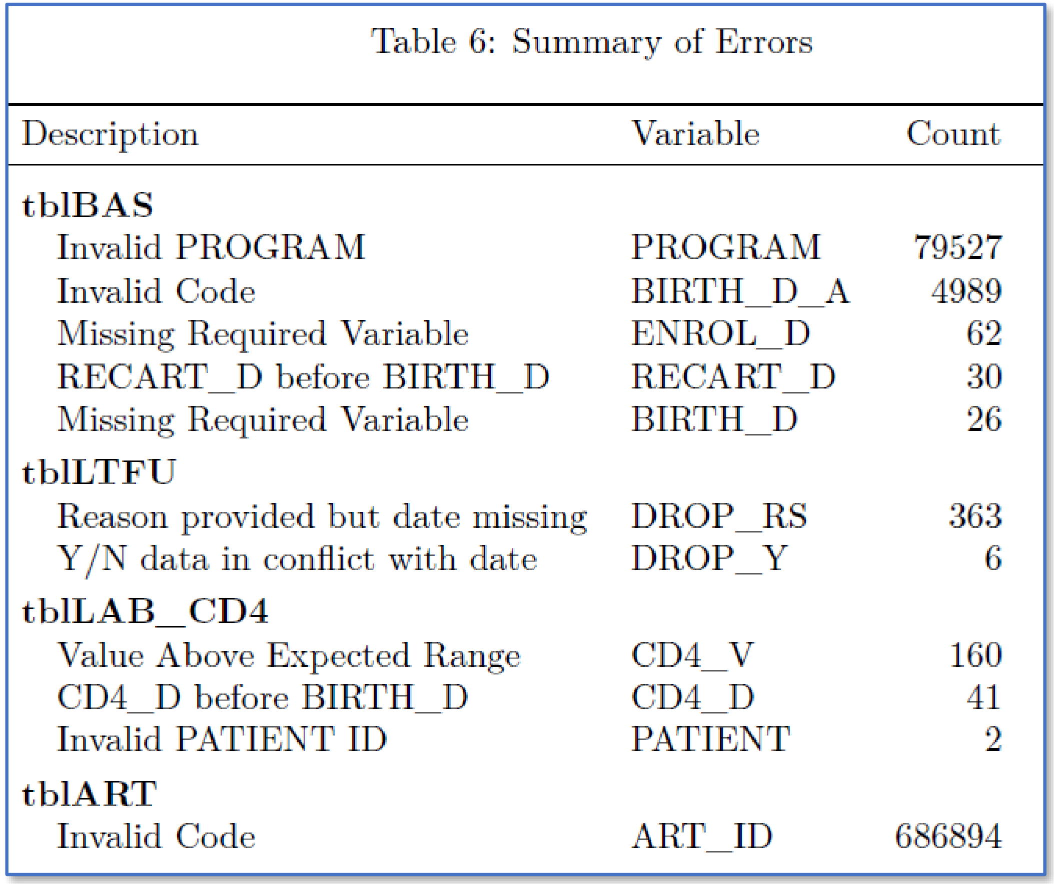
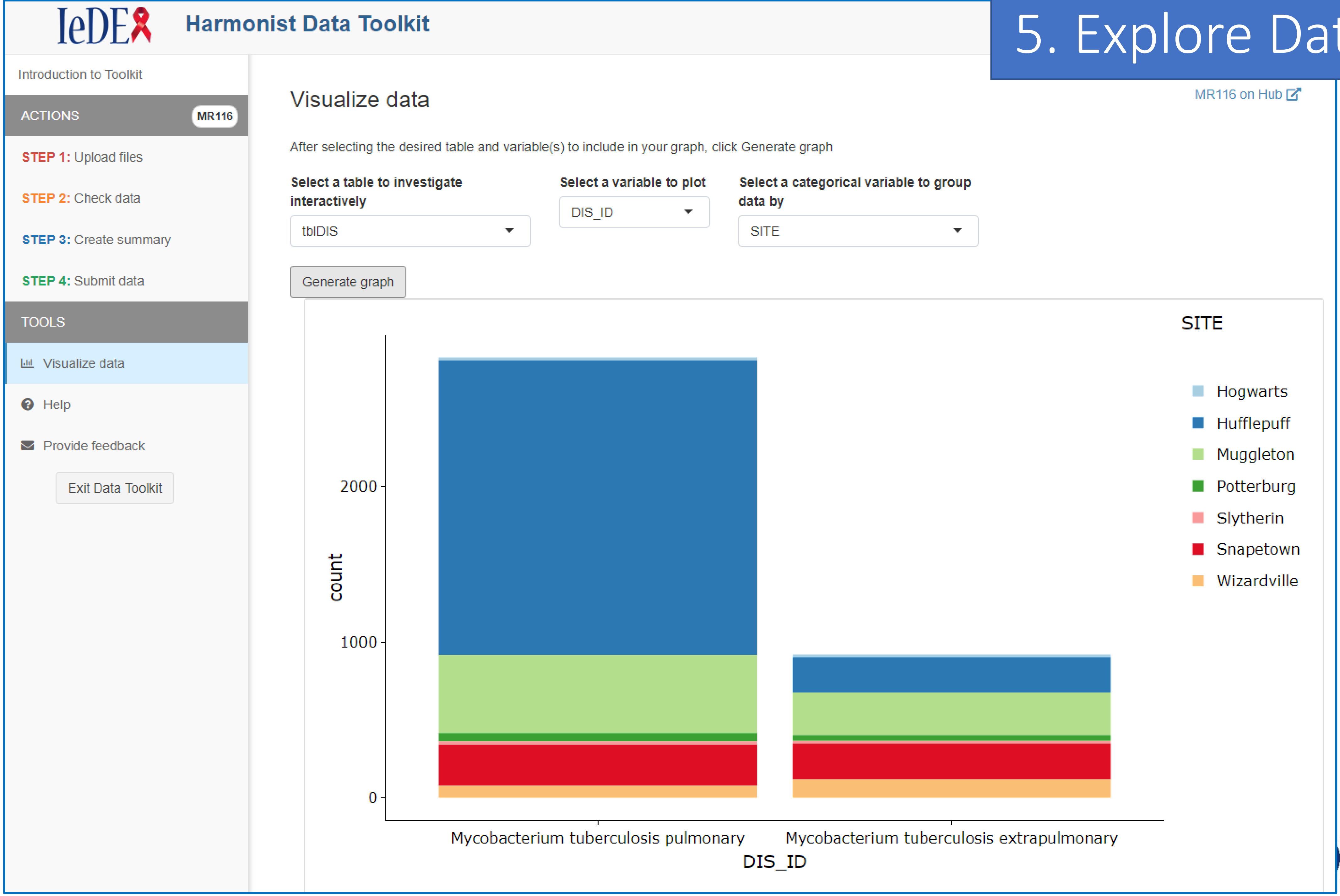


Table 6: Summary of Errors

Description	Variable	Count
tblBAS		
Invalid PROGRAM	PROGRAM	79527
Invalid Code	BIRTH_D_A	4989
Missing Required Variable	ENROL_D	62
RECART_D before BIRTH_D	RECART_D	30
Missing Required Variable	BIRTH_D	26
tblLTFU		
Reason provided but date missing	DROP_RS	363
Y/N data in conflict with date	DROP_Y	6
tblLAB_CD4		
Value Above Expected Range	CD4_V	160
CD4_D before BIRTH_D	CD4_D	41
Invalid PATIENT ID	PATIENT	2
tblART		
Invalid Code	ART_ID	686894

5. Explore Dataset



Visualize data

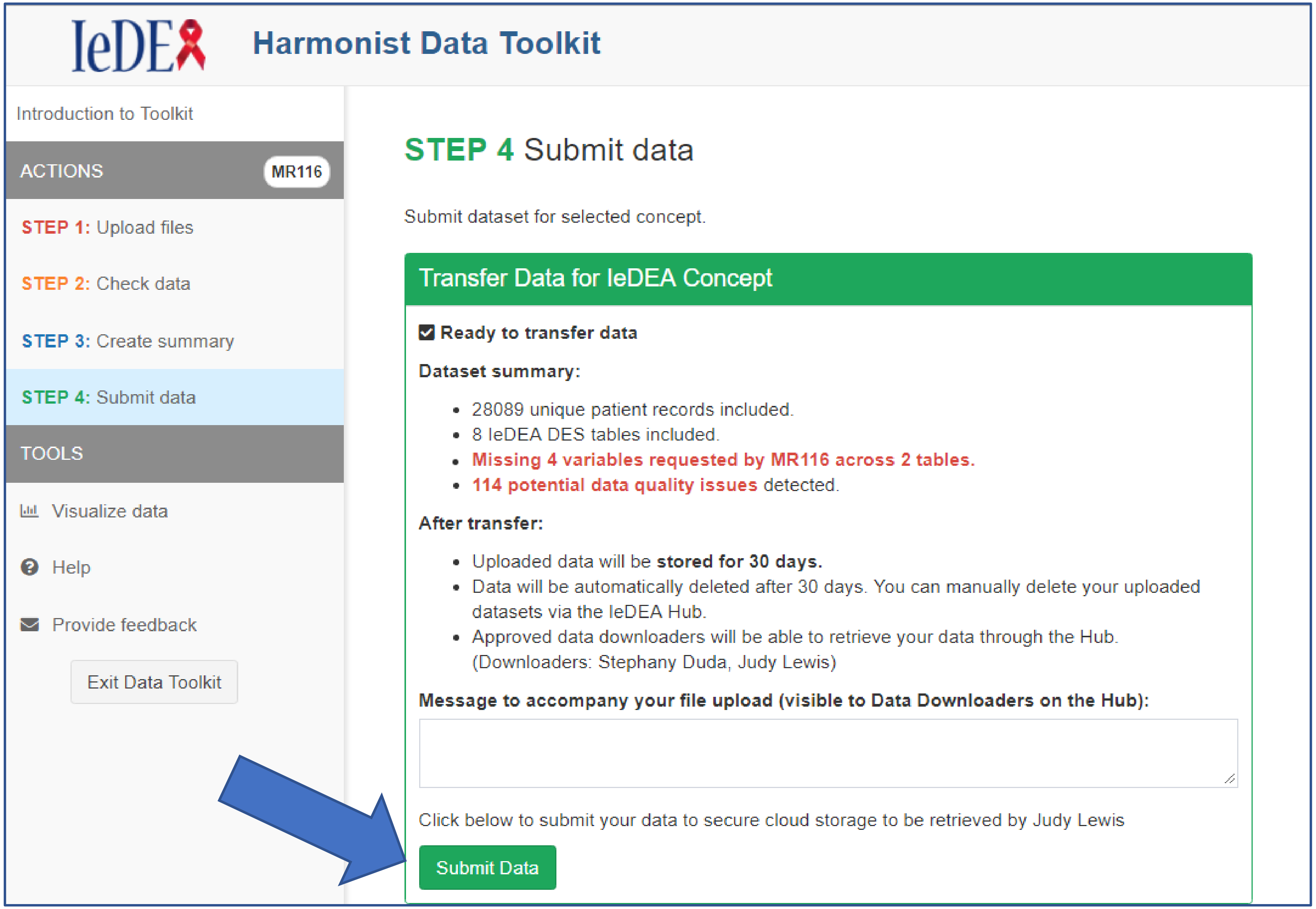
After selecting the desired table and variable(s) to include in your graph, click Generate graph

Select a table to investigate interactively: tblDIS

Select a variable to plot: DIS_ID

Select a categorical variable to group data by: SITE

6. Transfer Data



STEP 4 Submit data

Submit dataset for selected concept.

Transfer Data for IeDEA Concept

☑ **Ready to transfer data**

Dataset summary:

- 28089 unique patient records included.
- 8 IeDEA DES tables included.
- **Missing 4 variables requested by MR116 across 2 tables.**
- **114 potential data quality issues** detected.

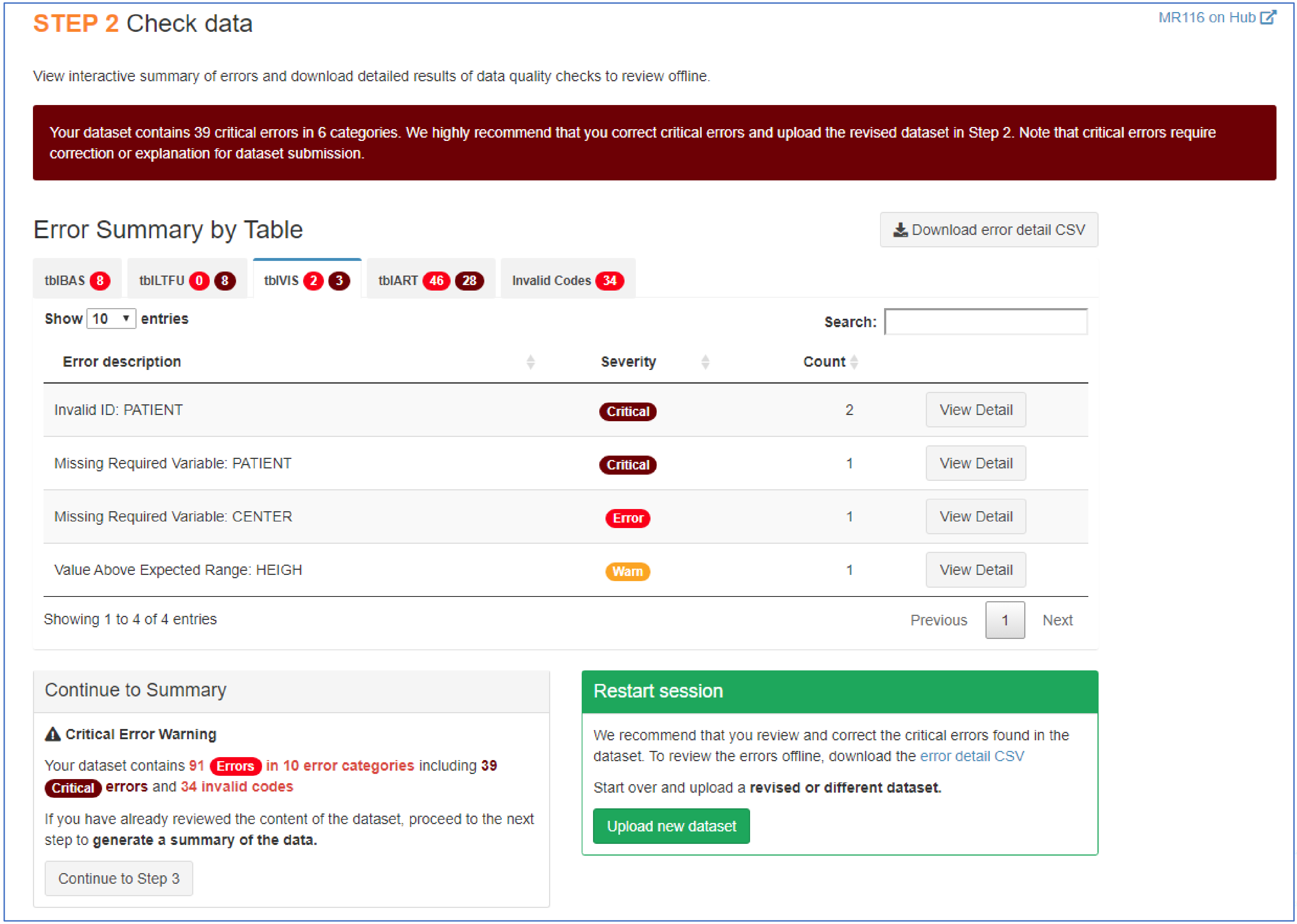
After transfer:

- Uploaded data will be **stored for 30 days.**
- Data will be automatically deleted after 30 days. You can manually delete your uploaded datasets via the IeDEA Hub.
- Approved data downloaders will be able to retrieve your data through the Hub. (Downloaders: Stephany Duda, Judy Lewis)

Message to accompany your file upload (visible to Data Downloaders on the Hub):

Click below to submit your data to secure cloud storage to be retrieved by Judy Lewis

Sometimes Datasets Include Critical Errors…



STEP 2 Check data

MR116 on Hub

View interactive summary of errors and download detailed results of data quality checks to review offline.

Your dataset contains 39 critical errors in 6 categories. We highly recommend that you correct critical errors and upload the revised dataset in Step 2. Note that critical errors require correction or explanation for dataset submission.

Error Summary by Table

Download error detail CSV

tblBAS 8 | tblLTFU 0 8 | tblVIS 2 3 | tblART 46 28 | Invalid Codes 34

Show 10 entries

Search:

Error description	Severity	Count	
Invalid ID: PATIENT	Critical	2	View Detail
Missing Required Variable: PATIENT	Critical	1	View Detail
Missing Required Variable: CENTER	Error	1	View Detail
Value Above Expected Range: HEIGH	Warn	1	View Detail

Showing 1 to 4 of 4 entries

Previous 1 Next

Continue to Summary

Critical Error Warning

Your dataset contains 91 Errors in 10 error categories including 39 Critical errors and 34 invalid codes

If you have already reviewed the content of the dataset, proceed to the next step to **generate a summary of the data.**

Continue to Step 3

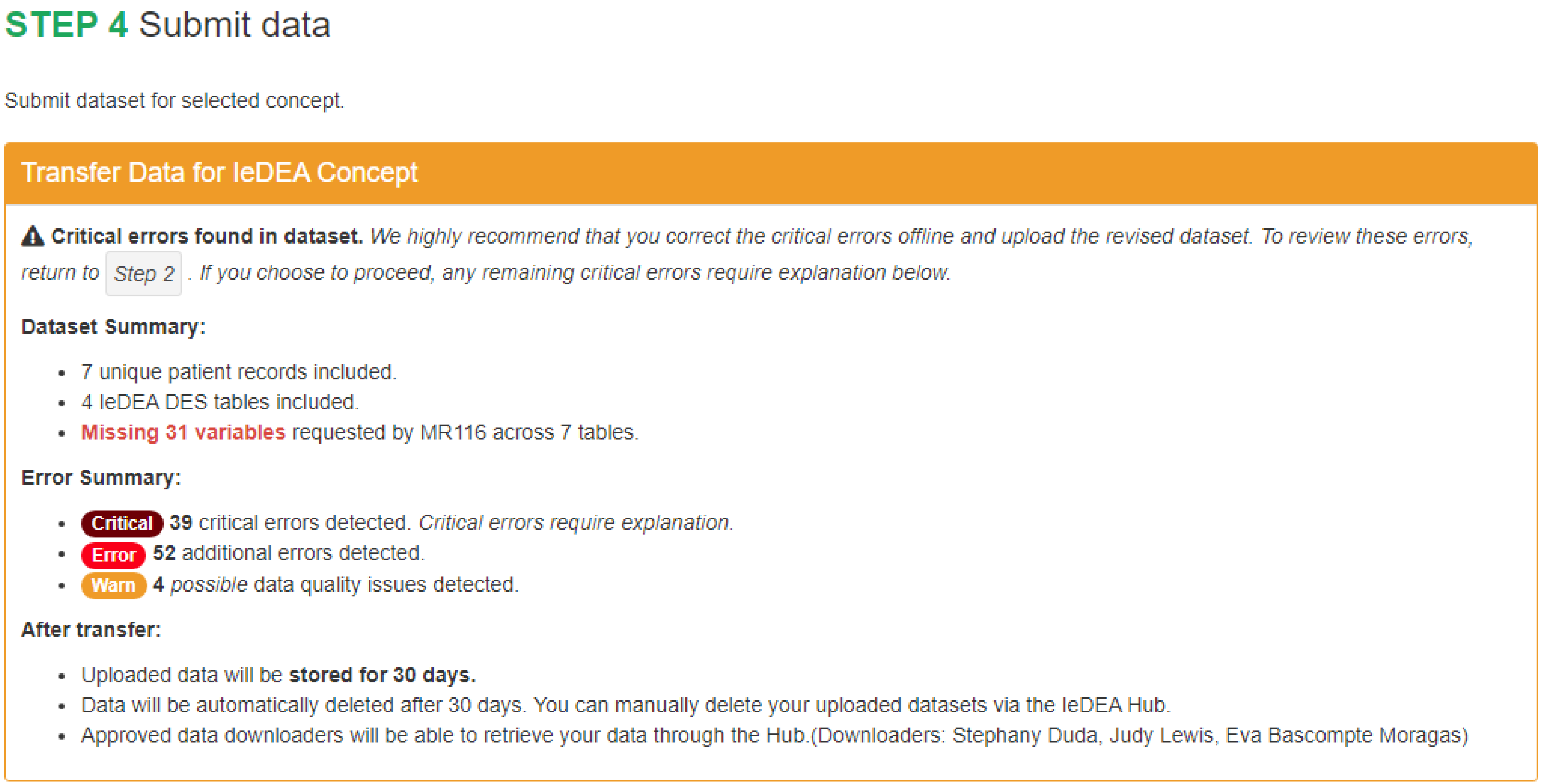
Restart session

We recommend that you review and correct the critical errors found in the dataset. To review the errors offline, download the error detail CSV

Start over and upload a **revised or different dataset.**

Upload new dataset

Researchers are strongly encouraged to revise data before submitting.



STEP 4 Submit data

Submit dataset for selected concept.

Transfer Data for IeDEA Concept

⚠ **Critical errors found in dataset.** *We highly recommend that you correct the critical errors offline and upload the revised dataset. To review these errors, return to* Step 2 *. If you choose to proceed, any remaining critical errors require explanation below.*

Dataset Summary:

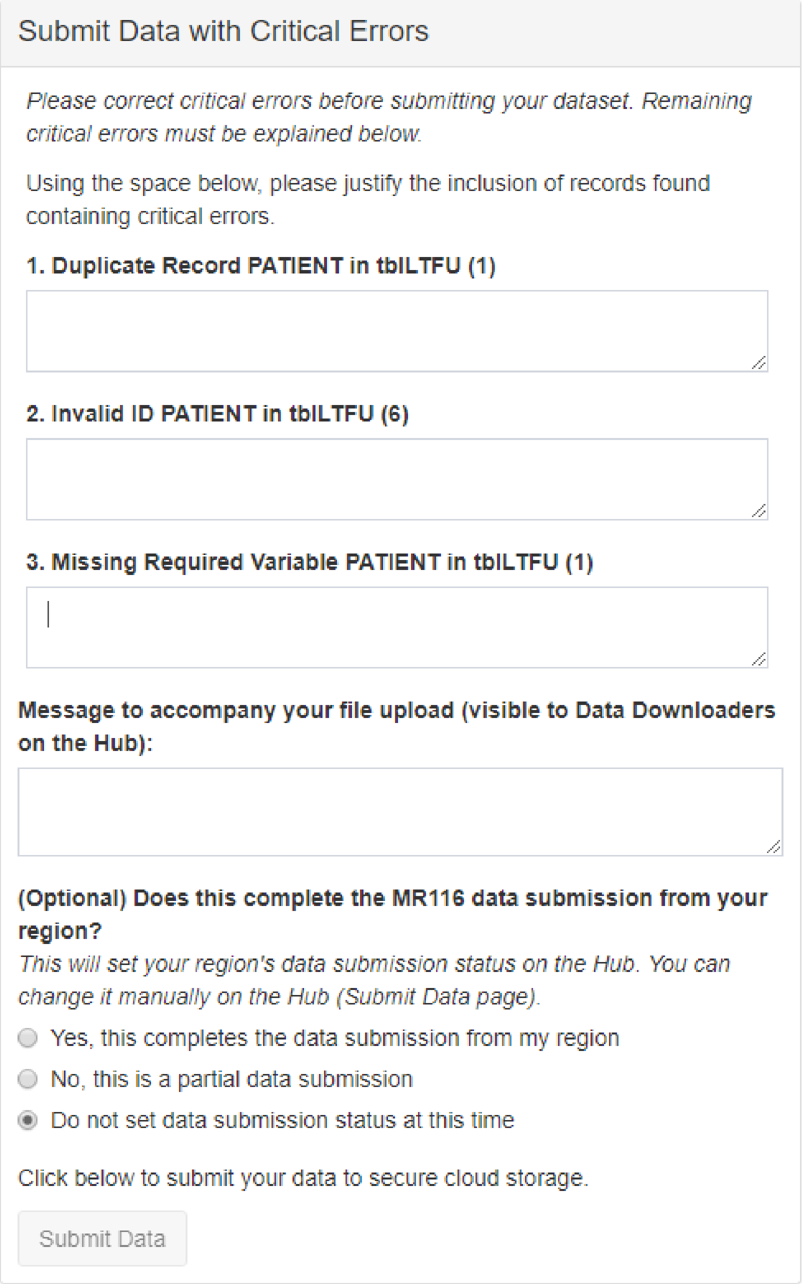
- 7 unique patient records included.
- 4 IeDEA DES tables included.
- **Missing 31 variables** requested by MR116 across 7 tables.

Error Summary:

- Critical **39** critical errors detected. *Critical errors require explanation.*
- Error **52** additional errors detected.
- Warn **4** *possible* data quality issues detected.

After transfer:

- Uploaded data will be **stored for 30 days.**
- Data will be automatically deleted after 30 days. You can manually delete your uploaded datasets via the IeDEA Hub.
- Approved data downloaders will be able to retrieve your data through the Hub.(Downloaders: Stephany Duda, Judy Lewis, Eva Bascompte Moragas)



Submit Data with Critical Errors

Please correct critical errors before submitting your dataset. Remaining critical errors must be explained below.

Using the space below, please justify the inclusion of records found containing critical errors.

1. Duplicate Record PATIENT in tblLTFU (1)

2. Invalid ID PATIENT in tblLTFU (6)

3. Missing Required Variable PATIENT in tblLTFU (1)

Message to accompany your file upload (visible to Data Downloaders on the Hub):

(Optional) Does this complete the MR116 data submission from your region?

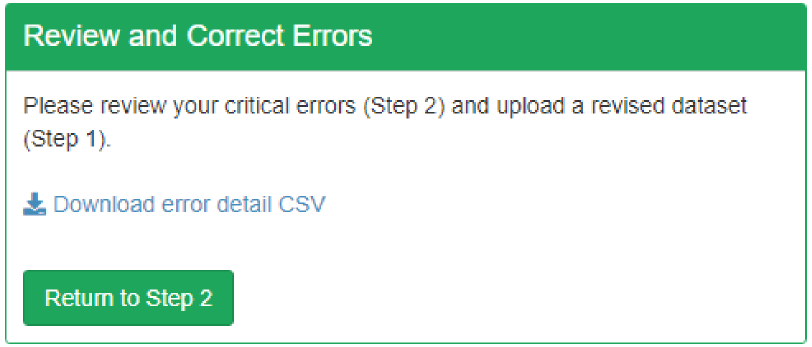
This will set your region's data submission status on the Hub. You can change it manually on the Hub (Submit Data page).

○ Yes, this completes the data submission from my region

○ No, this is a partial data submission

● Do not set data submission status at this time

Click below to submit your data to secure cloud storage.

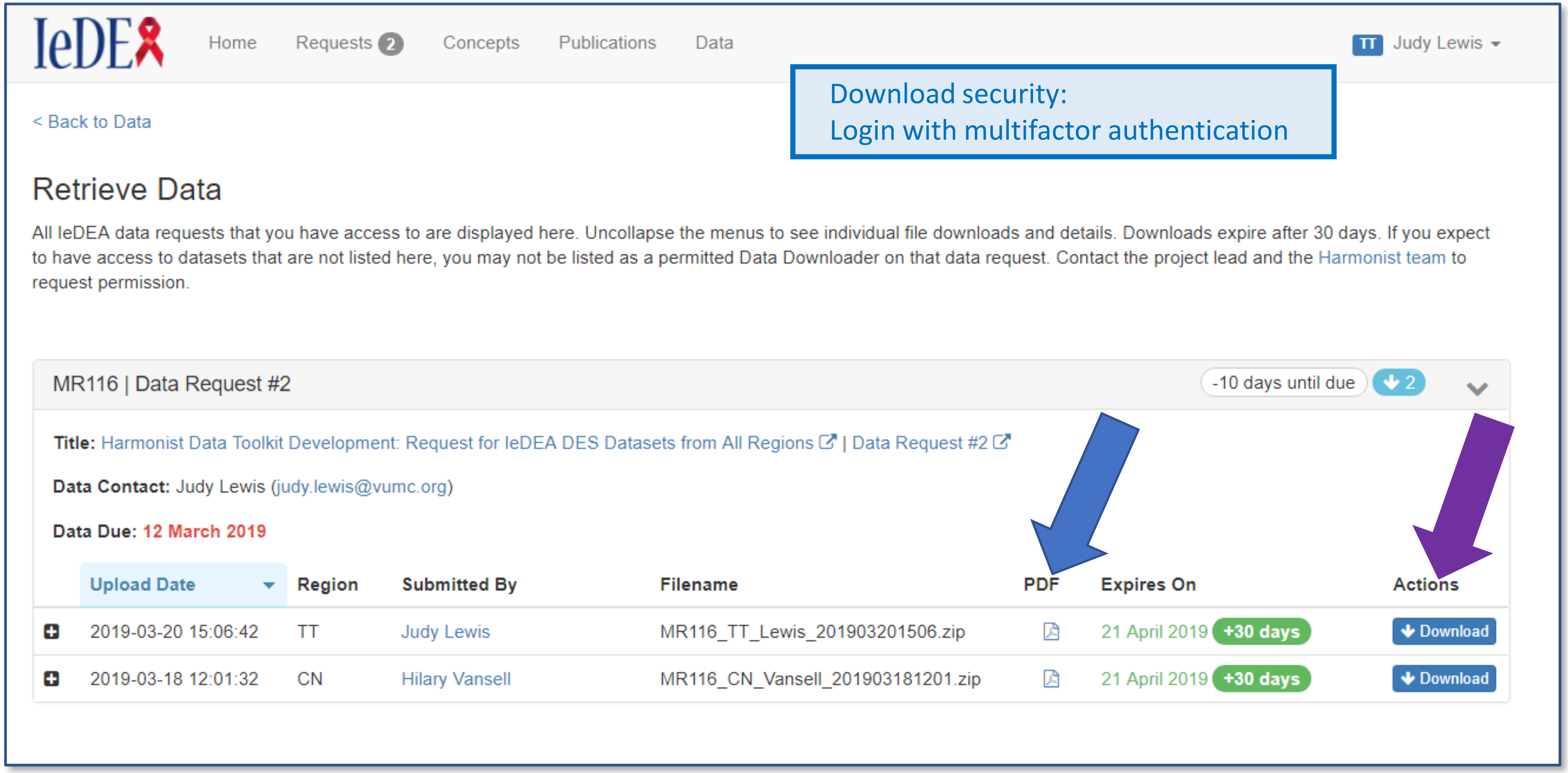


Review and Correct Errors

Please review your critical errors (Step 2) and upload a revised dataset (Step 1).

Submissions with Critical Errors require explanations.

7. Recipient Downloads Data



Download security:
Login with multifactor authentication

< Back to Data

Retrieve Data

All IeDEA data requests that you have access to are displayed here. Uncollapse the menus to see individual file downloads and details. Downloads expire after 30 days. If you expect to have access to datasets that are not listed here, you may not be listed as a permitted Data Downloader on that data request. Contact the project lead and the Harmonist team to request permission.

MR116 | Data Request #2 -10 days until due 2

Title: Harmonist Data Toolkit Development: Request for IeDEA DES Datasets from All Regions | Data Request #2

Data Contact: Judy Lewis (judy.lewis@vumc.org)

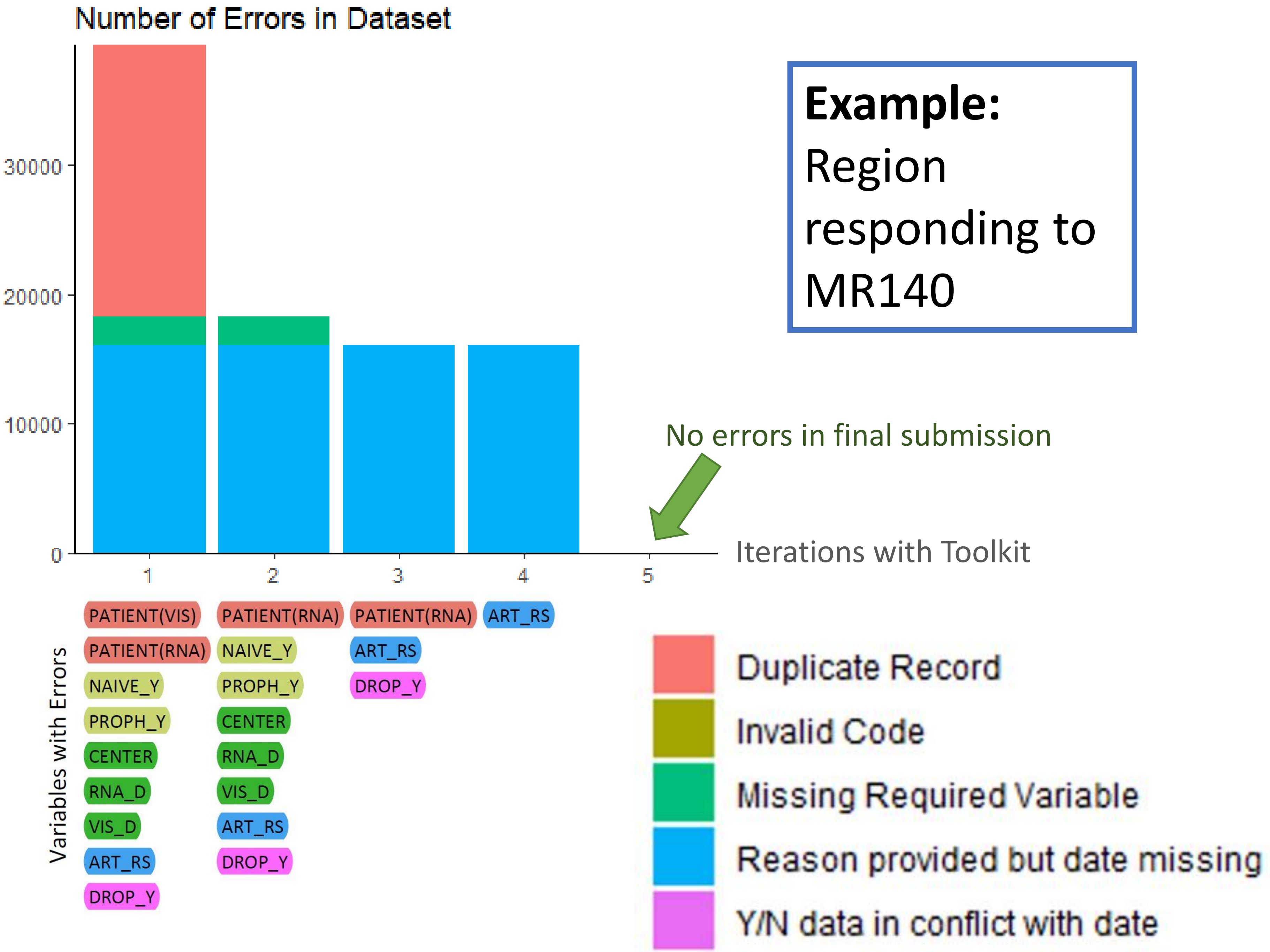
Data Due: 12 March 2019

	Upload Date	Region	Submitted By	Filename	PDF	Expires On	Actions
+	2019-03-20 15:06:42	TT	Judy Lewis	MR116_TT_Lewis_201903201506.zip		21 April 2019 +30 days	Download
+	2019-03-18 12:01:32	CN	Hilary Vansell	MR116_CN_Vansell_201903181201.zip		21 April 2019 +30 days	Download

Toolkit Impact on Data Quality

- As of May 2020:
  - **>700 datasets** processed
  - 1,800 to 986,089 patients per dataset
  - Used for 7 official multiregional IeDEA data calls
- Regional data managers uploaded datasets and reviewed data quality results multiple times before final submission
- Results suggest that data managers used Toolkit data quality reports to improve datasets before submission
- The number and types of errors decreased with each iteration of Toolkit use.

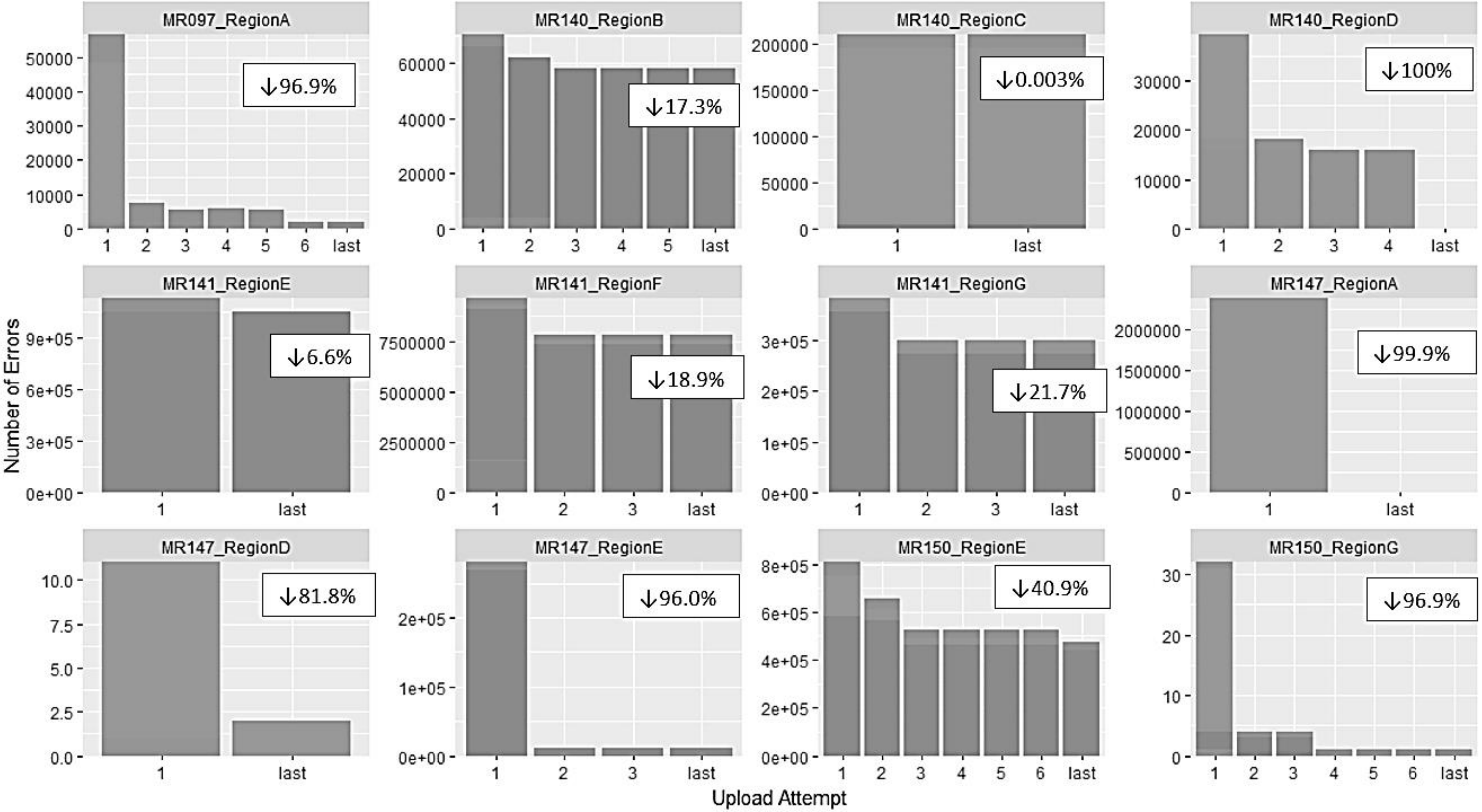
Example:
Region responding to MR140



No errors in final submission

Iterations with Toolkit

Toolkit Use Impact on Data Quality:
Median Percent Decrease in Number of Errors = 61.3%



Error types most often…

Corrected by final submission

- Invalid IDs (patient not in tblBAS)
- Invalid codes (tblLTFU)
- Duplicate records
- Out-of-range values

Remaining in final submission

- Invalid codes (ART, labs)
- Date logic errors

Why This Matters

- High quality data is essential to meaningful research.
- Tools like this can help:
  - Improve adherence to data model and standards
  - Reduce time for data preparation and checking
  - Highlight data completeness and coding problems
  - Increase security and uniform workflow for data exchange
- Generalized design using REDCap allows software to be adapted to other domains.

Lessons Learned

- Close collaboration with stakeholders and users is key
  - Monthly Data Harmonization Working Group Calls
  - Structured testing and training exercises with users
  - International meetings to collaborate in person on design
- Defining details in REDCap and using the REDCap API make it possible to design tools that adapt with the changing data model
- Web-based tools are easy to use and require no user maintenance or equipment

Future Development

- Expand data quality checks, report content
- Enhance code portability
- Dataset quality metrics
- New application domains

Code available github.com/IeDEA/Harmonist

New Quality Metrics Report

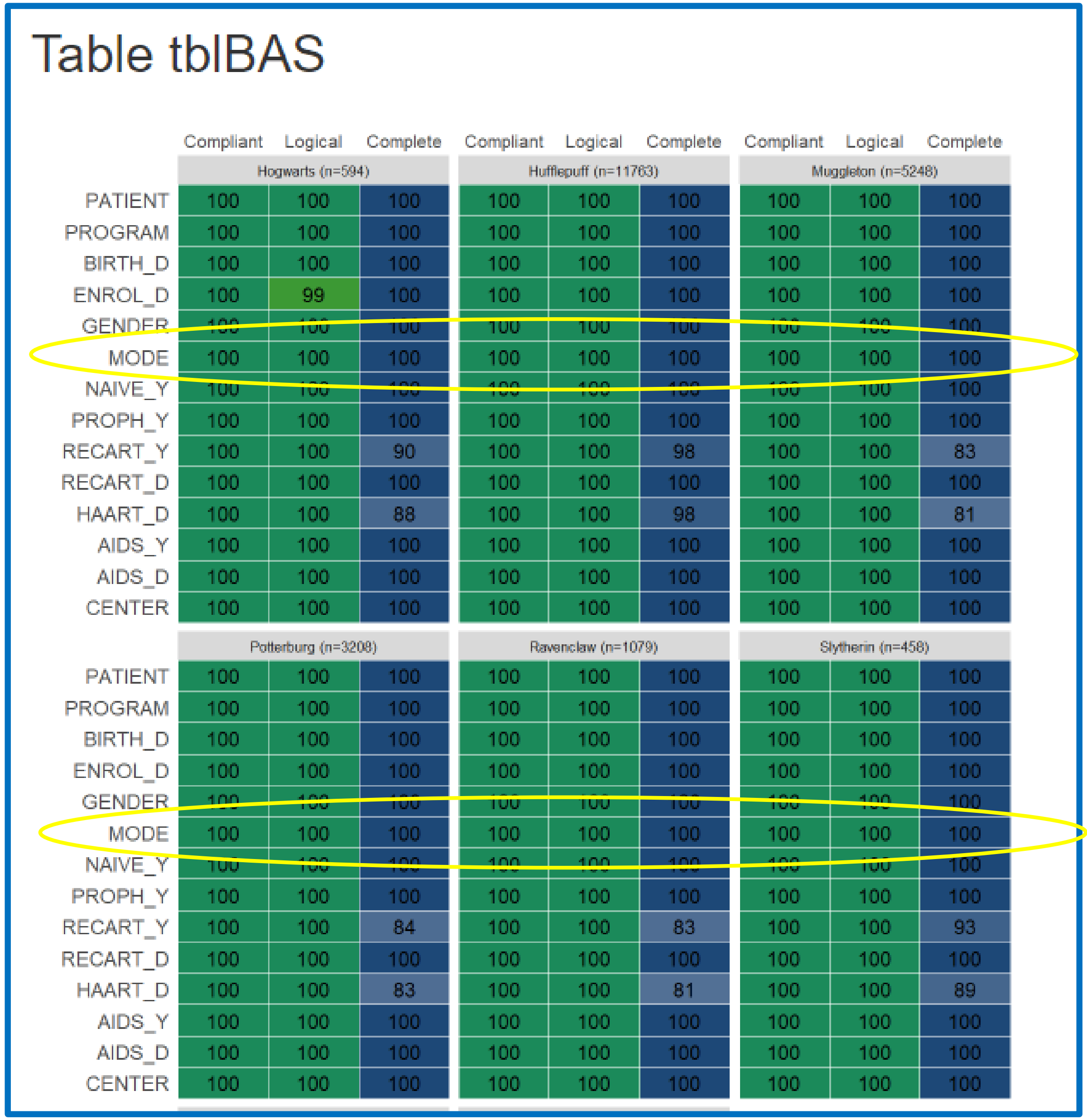
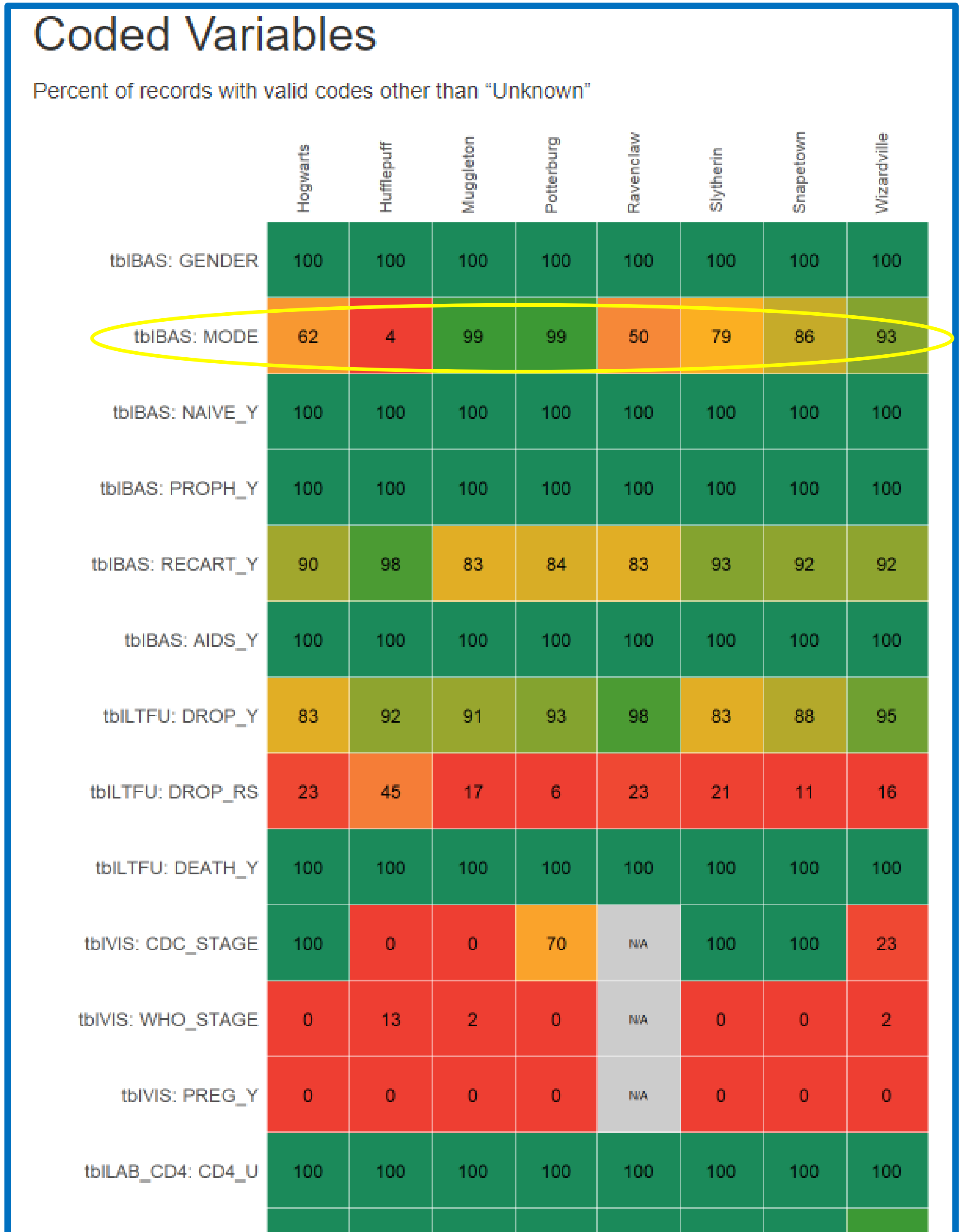


Table tblBAS

	Compliant	Logical	Complete	Compliant	Logical	Complete	Compliant	Logical	Complete
	Hogwarts (n=594)			Hufflepuff (n=11763)			Muggleton (n=5248)		
PATIENT	100	100	100	100	100	100	100	100	100
PROGRAM	100	100	100	100	100	100	100	100	100
BIRTH_D	100	100	100	100	100	100	100	100	100
ENROL_D	100	99	100	100	100	100	100	100	100
GENDER	100	100	100	100	100	100	100	100	100
MODE	100	100	100	100	100	100	100	100	100
NAIVE_Y	100	100	100	100	100	100	100	100	100
PROPH_Y	100	100	100	100	100	100	100	100	100
RECART_Y	100	100	90	100	100	98	100	100	83
RECART_D	100	100	100	100	100	100	100	100	100
HAART_D	100	100	88	100	100	98	100	100	81
AIDS_Y	100	100	100	100	100	100	100	100	100
AIDS_D	100	100	100	100	100	100	100	100	100
CENTER	100	100	100	100	100	100	100	100	100

	Potterburg (n=3208)			Ravenclaw (n=1079)			Slytherin (n=458)		
PATIENT	100	100	100	100	100	100	100	100	100
PROGRAM	100	100	100	100	100	100	100	100	100
BIRTH_D	100	100	100	100	100	100	100	100	100
ENROL_D	100	100	100	100	100	100	100	100	100
GENDER	100	100	100	100	100	100	100	100	100
MODE	100	100	100	100	100	100	100	100	100
NAIVE_Y	100	100	100	100	100	100	100	100	100
PROPH_Y	100	100	100	100	100	100	100	100	100
RECART_Y	100	100	84	100	100	83	100	100	93
RECART_D	100	100	100	100	100	100	100	100	100
HAART_D	100	100	83	100	100	81	100	100	89
AIDS_Y	100	100	100	100	100	100	100	100	100
AIDS_D	100	100	100	100	100	100	100	100	100
CENTER	100	100	100	100	100	100	100	100	100



Coded Variables

Percent of records with valid codes other than "Unknown"

	Hogwarts	Hufflepuff	Muggleton	Potterburg	Ravenclaw	Slytherin	Snapetown	Wizardville
tblBAS: GENDER	100	100	100	100	100	100	100	100
tblBAS: MODE	62	4	99	99	50	79	86	93
tblBAS: NAIVE_Y	100	100	100	100	100	100	100	100
tblBAS: PROPH_Y	100	100	100	100	100	100	100	100
tblBAS: RECART_Y	90	98	83	84	83	93	92	92
tblBAS: AIDS_Y	100	100	100	100	100	100	100	100
tblLTFU: DROP_Y	83	92	91	93	98	83	88	95
tblLTFU: DROP_RS	23	45	17	6	23	21	11	16
tblLTFU: DEATH_Y	100	100	100	100	100	100	100	100
tblVIS: CDC_STAGE	100	0	0	70	N/A	100	100	23
tblVIS: WHO_STAGE	0	13	2	0	N/A	0	0	2
tblVIS: PREG_Y	0	0	0	0	N/A	0	0	0
tblLAB_CD4: CD4_U	100	100	100	100	100	100	100	100

Thank You

- Harmonist and REDCap technical teams
- IeDEA Data Harmonization Working Group and collaborators
- HICDEP colleagues
- IWHOD
- This work was funded by US NIAID under grant R24 AI124872 (“Harmonist”)