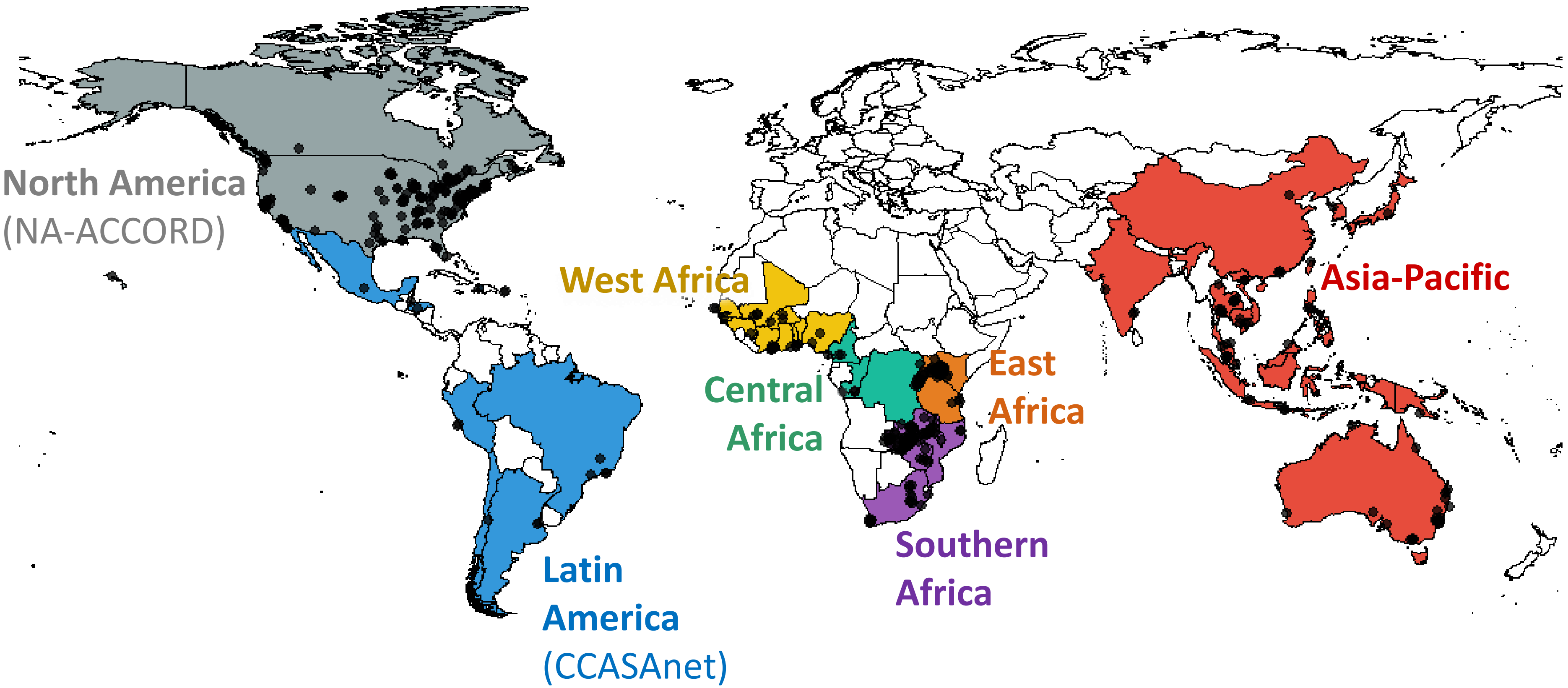
A Flexible Open Source Tool for Quality Checking and Sharing Observational HIV Data

Judith Lewis, Jeremy Stephens, Beverly Musick, Karen Malateste, Nicola Maxwell, Brenna Hogan, Hae-Young Kim, Karu Jayathilake, Azar Kariminia, Cam Ha Dao Ostinelli, Steven Brown, Stephany Duda

IeDEA
International epidemiology
Databases to Evaluate AIDS

The Challenge: Data Harmonization in IeDEA



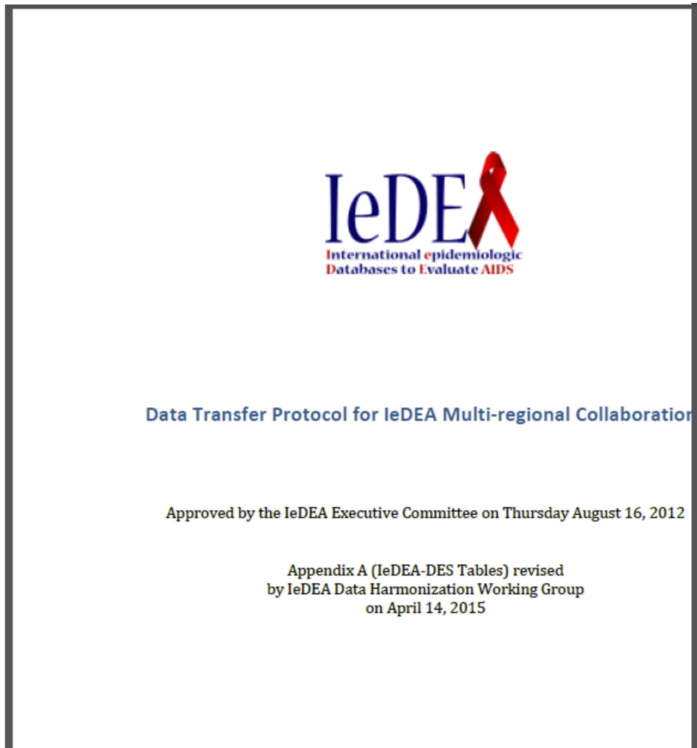
International epidemiology Databases to Evaluate AIDS

A global collaboration of seven regional HIV observational research networks with combined data on nearly two million persons living with HIV (PLWH)

IWHOD 2017

Establishing the IeDEA Data Exchange Standard

The **IeDEA Data Exchange Standard (DES)** is a common data model for sharing HIV data within IeDEA. The DES defines variable names, variable definitions, and code lists.

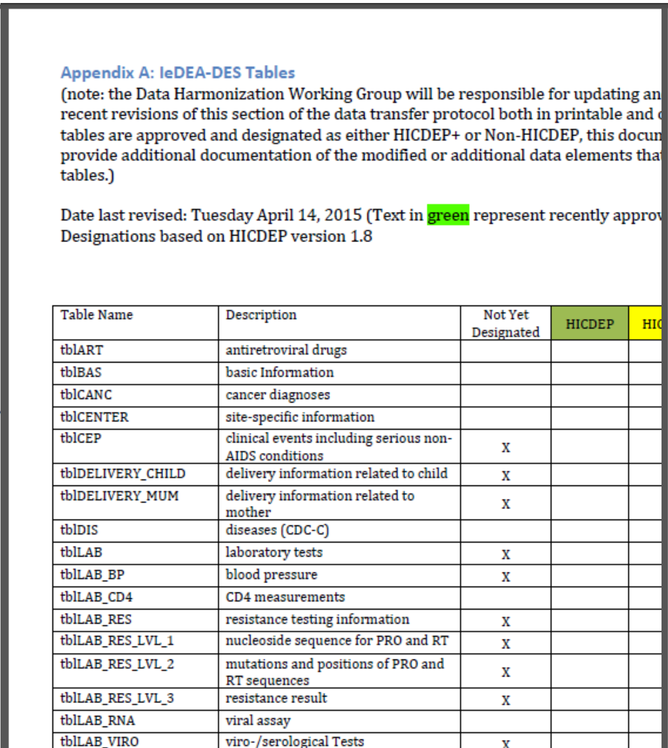


IeDEA
International epidemiologic Databases to Evaluate AIDS

Data Transfer Protocol for IeDEA Multi-regional Collaboratior

Approved by the IeDEA Executive Committee on Thursday August 16, 2012

Appendix A (IeDEA-DES Tables) revised
by IeDEA Data Harmonization Working Group
on April 14, 2015

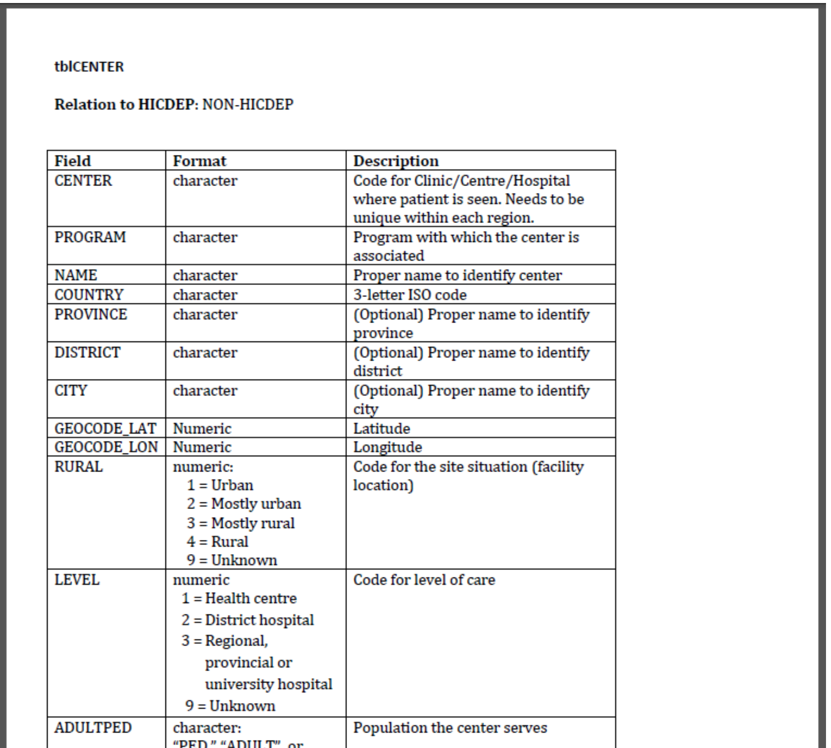


Appendix A: IeDEA-DES Tables

(note: the Data Harmonization Working Group will be responsible for updating an recent revisions of this section of the data transfer protocol both in printable and tables are approved and designated as either HICDEP+ or Non-HICDEP, this docun provide additional documentation of the modified or additional data elements tha tables.)

Date last revised: Tuesday April 14, 2015 (Text in green represent recently approv Designations based on HICDEP version 1.8

Table Name	Description	Not Yet Designated	HICDEP	HIC
tblART	antiretroviral drugs			
tblBAS	basic Information			
tblCANC	cancer diagnoses			
tblCENTER	site-specific information			
tblCEP	clinical events including serious non-AIDS conditions	X		
tblDELIVERY_CHILD	delivery information related to child	X		
tblDELIVERY_MUM	delivery information related to mother	X		
tblDIS	diseases (CDC-C)			
tblLAB	laboratory tests	X		
tblLAB_BP	blood pressure	X		
tblLAB_CD4	CD4 measurements			
tblLAB_RES	resistance testing information	X		
tblLAB_RES_LVL_1	nucleoside sequence for PRO and RT	X		
tblLAB_RES_LVL_2	mutations and positions of PRO and RT sequences	X		
tblLAB_RES_LVL_3	resistance result	X		
tblLAB_RNA	viral assay			
tblLAB_VIRO	viro-/serological Tests	X		



tblCENTER

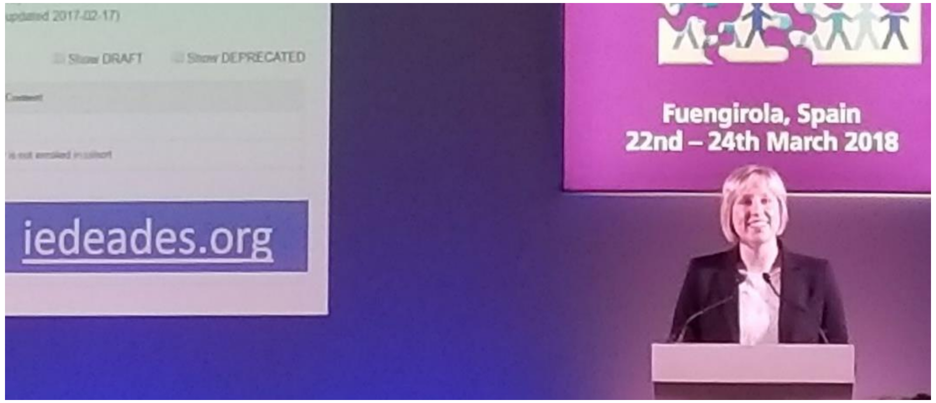
Relation to HICDEP: NON-HICDEP

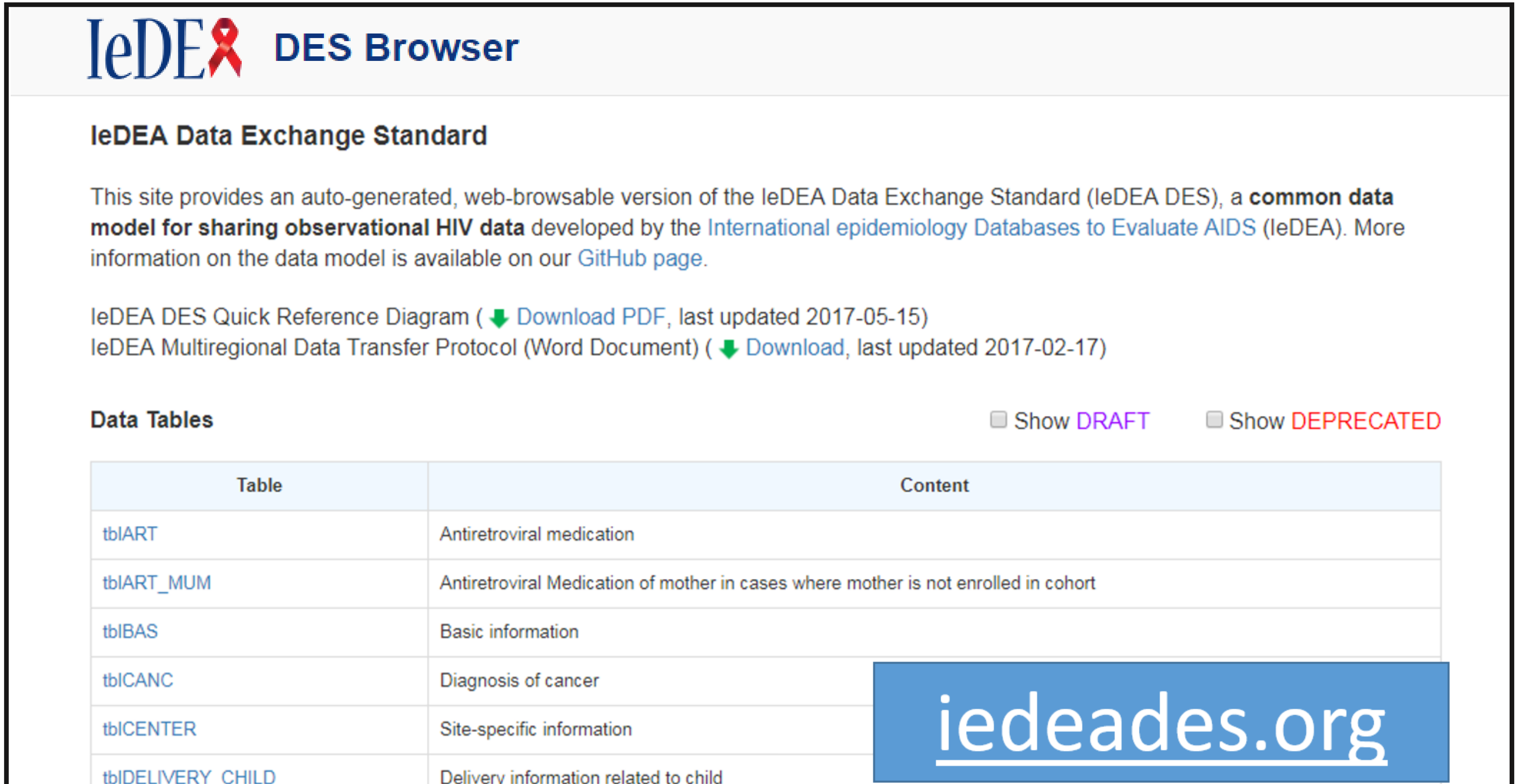
Field	Format	Description
CENTER	character	Code for Clinic/Centre/Hospital where patient is seen. Needs to be unique within each region.
PROGRAM	character	Program with which the center is associated
NAME	character	Proper name to identify center
COUNTRY	character	3-letter ISO code
PROVINCE	character	(Optional) 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	Population the center serves

IWHOD 2018

Maintaining the IeDEA DES in REDCap

Research Electronic Data Capture





IeDEA **DES Browser**

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 2017-05-15)
IeDEA Multiregional Data Transfer Protocol (Word Document) (Download, last updated 2017-02-17)

Data Tables □ Show DRAFT □ Show DEPRECATED

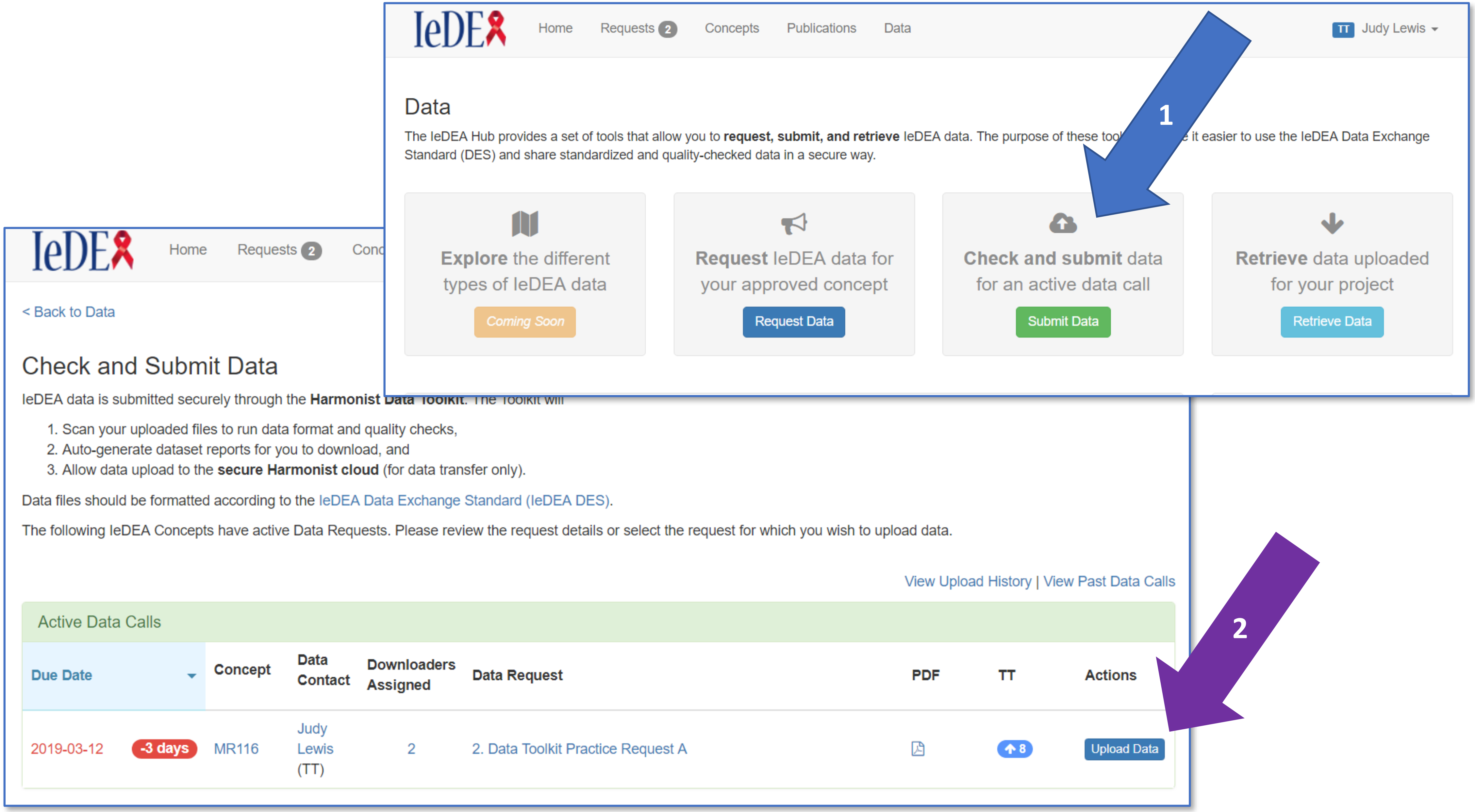
Table	Content
tblART	Antiretroviral medication
tblART_MUM	Antiretroviral Medication of mother in cases where mother is not enrolled in cohort
tblBAS	Basic information
tblCANC	Diagnosis of cancer
tblCENTER	Site-specific information
tblDELIVERY_CHILD	Delivery information related to child

iedeades.org

Today: IeDEA Harmonist Data Toolkit

- Collaborative project with all seven IeDEA regions
- Web application
- Developed with open source tools
- Features:
  - ▦ Ensures datasets conform to common data model
  - ✓ Performs data quality checks
  - 📊 Generates reproducible reports
  - ☁ Submits approved datasets to secure cloud storage

Workflow Begins in IeDEA Project Portal



Data

The IeDEA Hub provides a set of tools that allow you to **request, submit, and retrieve** IeDEA data. The purpose of these tool 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).

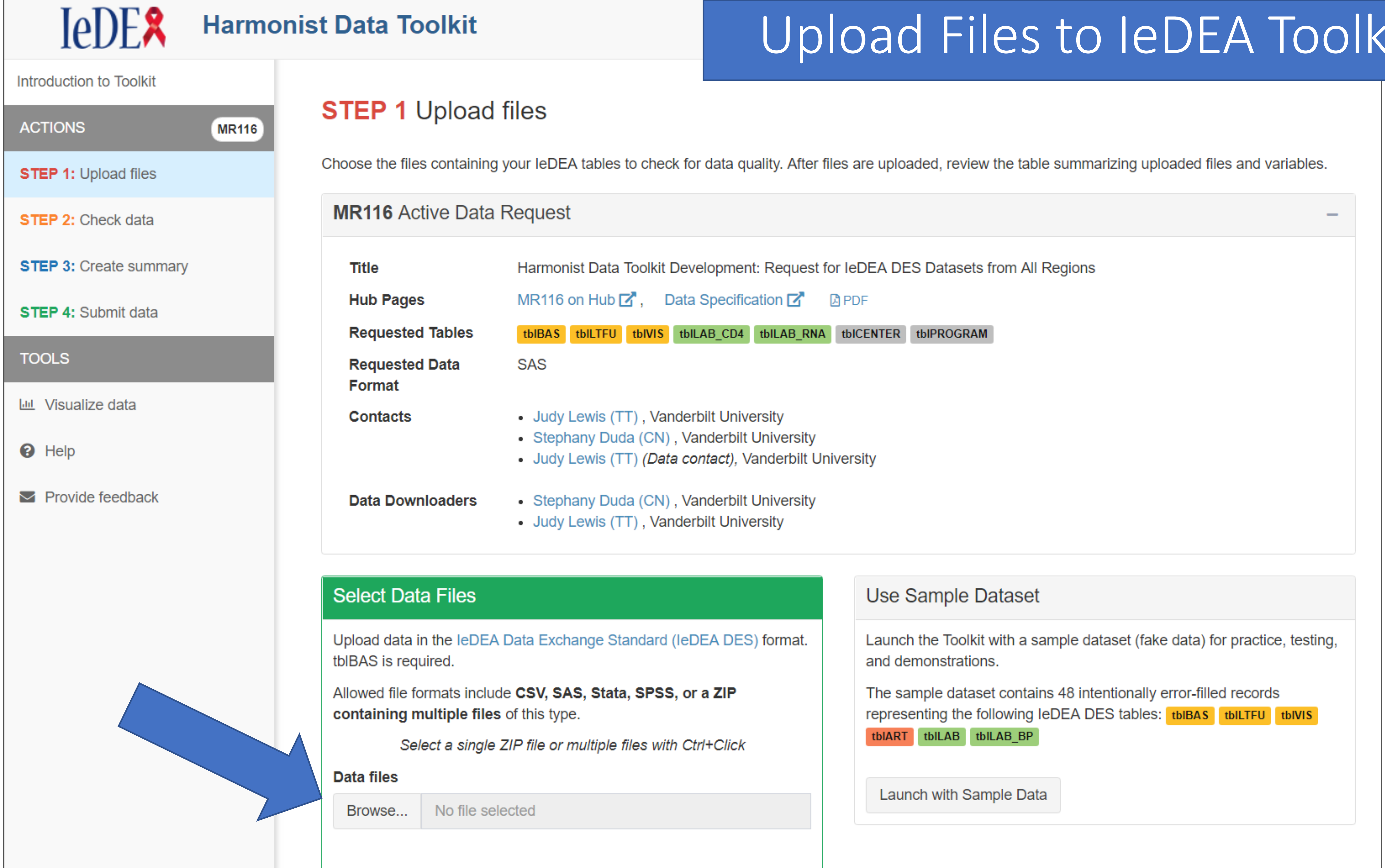
The following IeDEA Concepts have active Data Requests. Please review the request details or select the request for which you wish to upload data.

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

Upload Files to IeDEA Toolkit



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

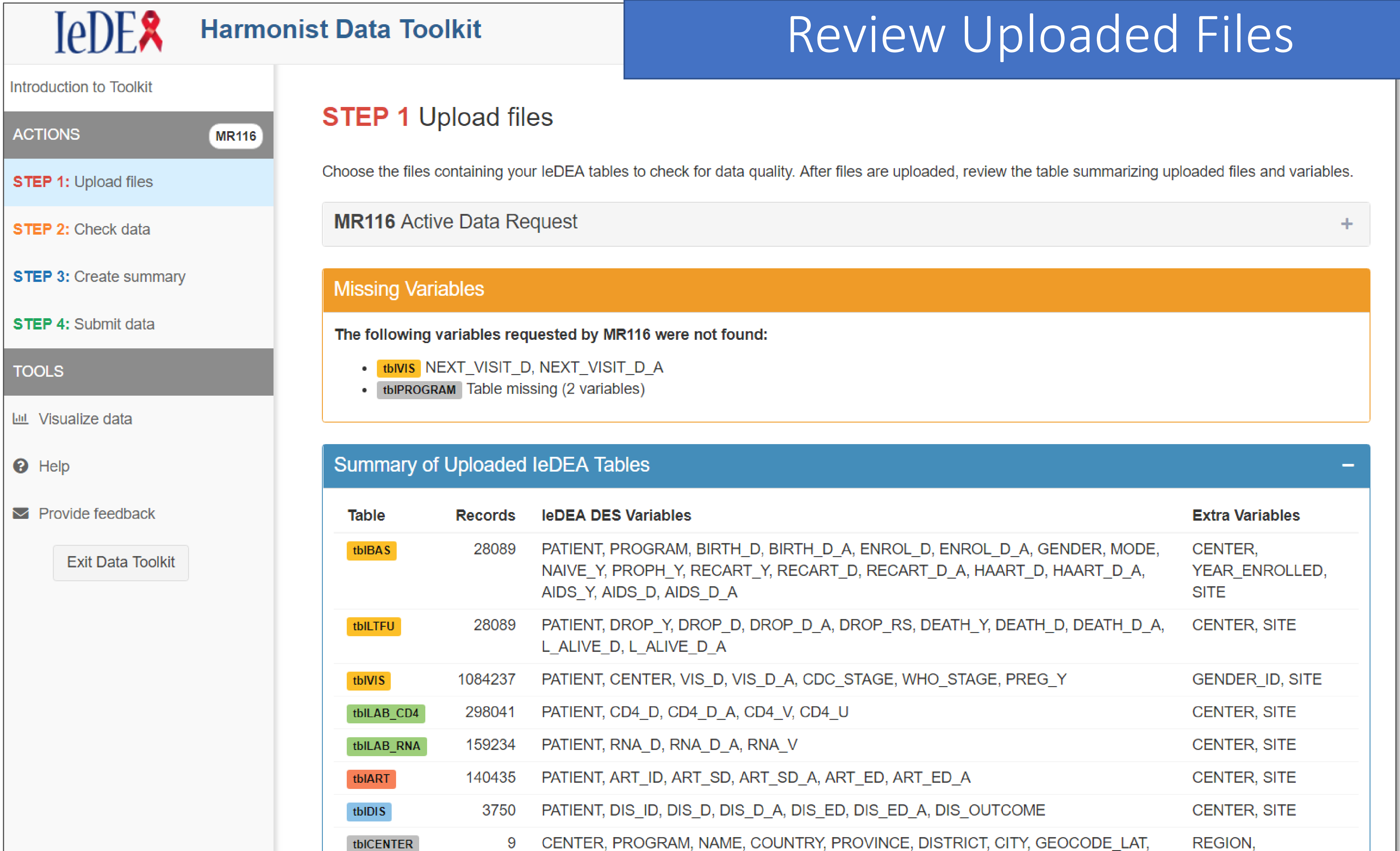
Data files

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

Review Uploaded Files



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 +

Missing Variables

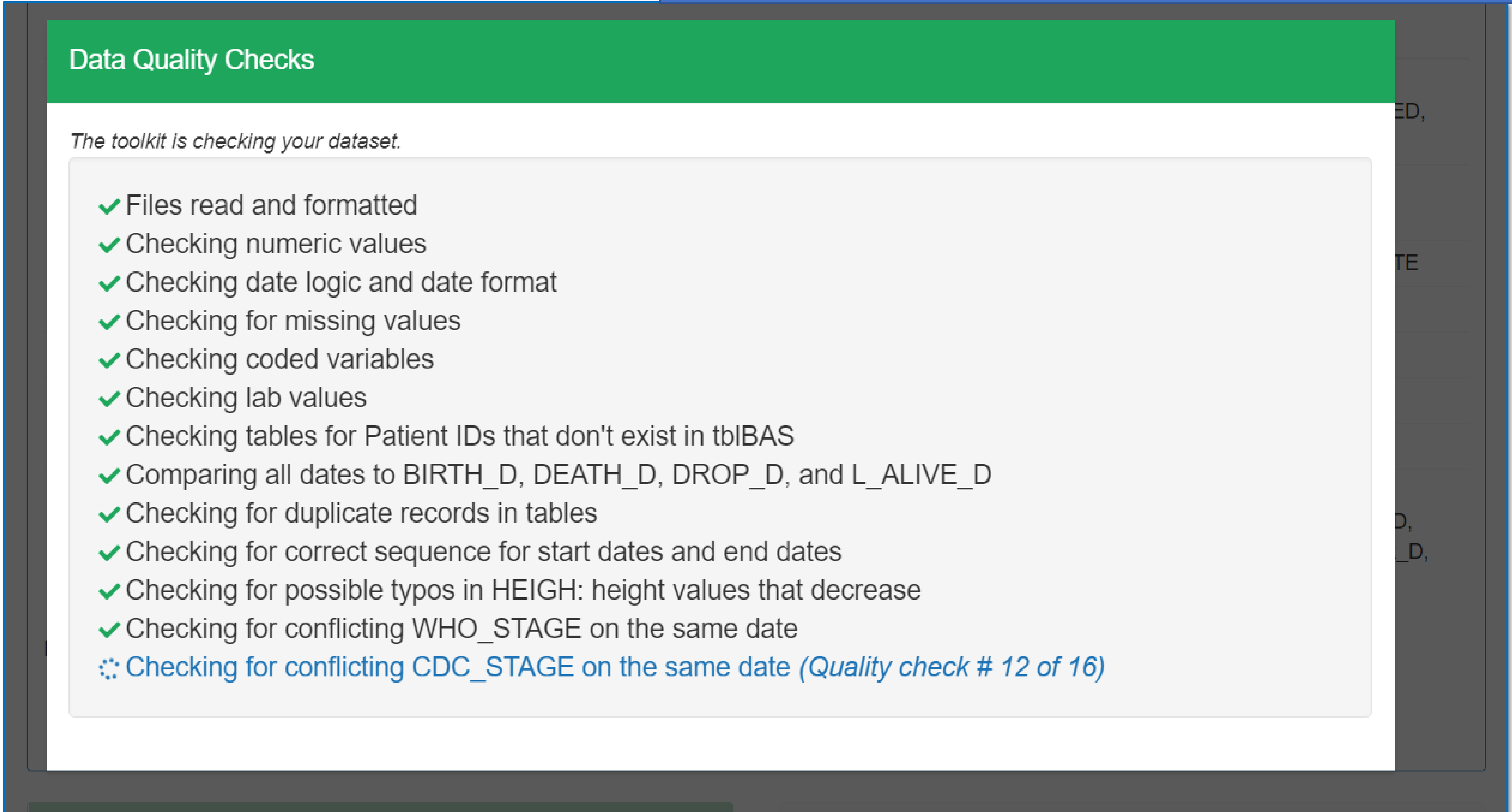
The following variables requested by MR116 were not found:

- tblVIS NEXT_VISIT_D, NEXT_VISIT_D_A
- tblPROGRAM Table missing (2 variables)

Summary of Uploaded IeDEA Tables –

Table	Records	IeDEA DES Variables	Extra Variables
tblBAS	28089	PATIENT, PROGRAM, BIRTH_D, BIRTH_D_A, ENROL_D, ENROL_D_A, GENDER, MODE, NAIVE_Y, PROPH_Y, RECART_Y, RECART_D, RECART_D_A, HAART_D, HAART_D_A, AIDS_Y, AIDS_D, AIDS_D_A	CENTER, YEAR_ENROLLED, SITE
tblLTFU	28089	PATIENT, DROP_Y, DROP_D, DROP_D_A, DROP_RS, DEATH_Y, DEATH_D, DEATH_D_A, L_ALIVE_D, L_ALIVE_D_A	CENTER, SITE
tblVIS	1084237	PATIENT, CENTER, VIS_D, VIS_D_A, CDC_STAGE, WHO_STAGE, PREG_Y	GENDER_ID, SITE
tblLAB_CD4	298041	PATIENT, CD4_D, CD4_D_A, CD4_V, CD4_U	CENTER, SITE
tblLAB_RNA	159234	PATIENT, RNA_D, RNA_D_A, RNA_V	CENTER, SITE
tblART	140435	PATIENT, ART_ID, ART_SD, ART_SD_A, ART_ED, ART_ED_A	CENTER, SITE
tblDIS	3750	PATIENT, DIS_ID, DIS_D, DIS_D_A, DIS_ED, DIS_ED_A, DIS_OUTCOME	CENTER, SITE
tblCENTER	9	CENTER, PROGRAM, NAME, COUNTRY, PROVINCE, DISTRICT, CITY, GEOCODE_LAT,	REGION,

Data Quality Checks

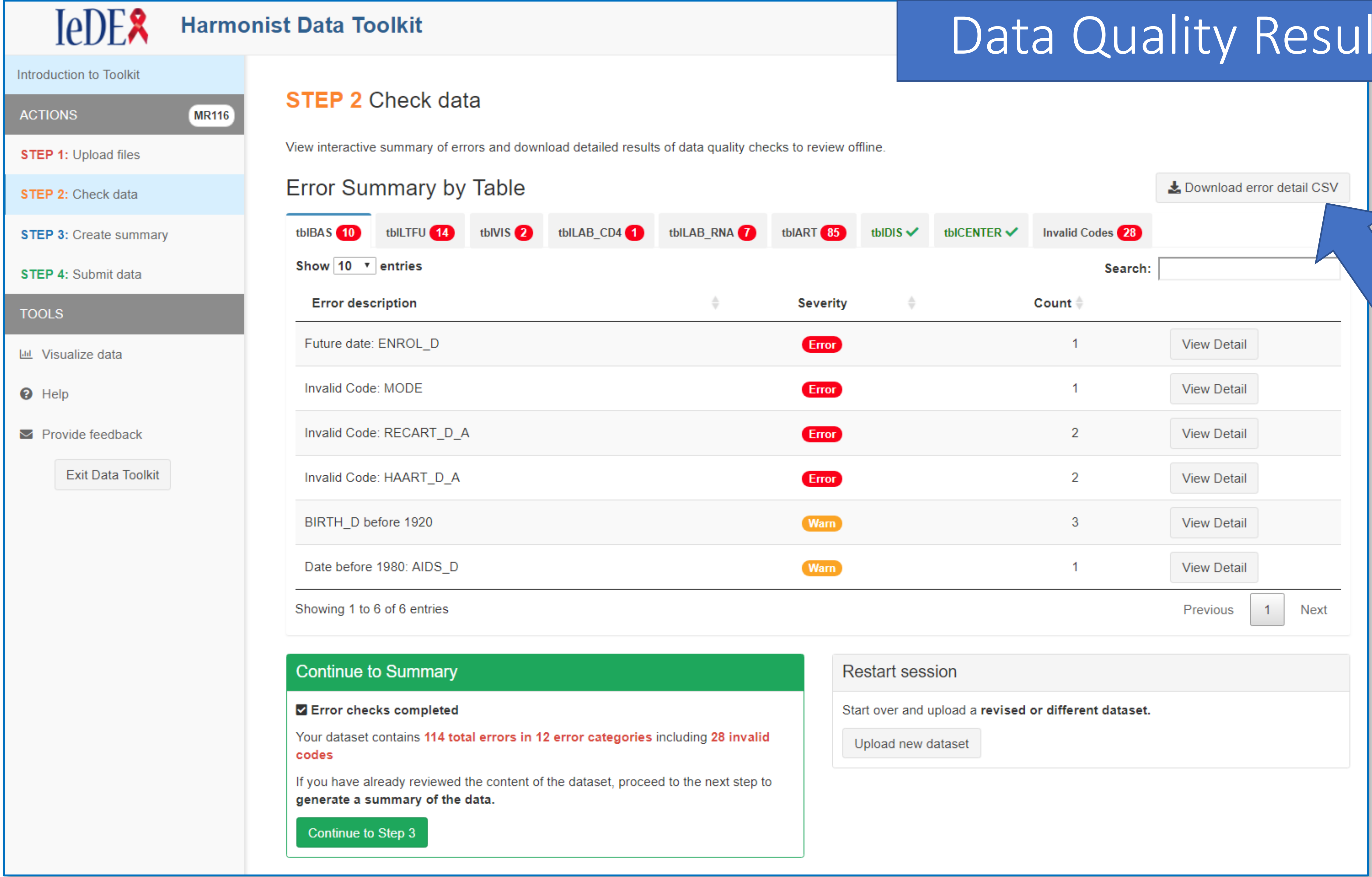


Data Quality Checks

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)*

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

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

Continue to Summary

☑ **Error checks completed**

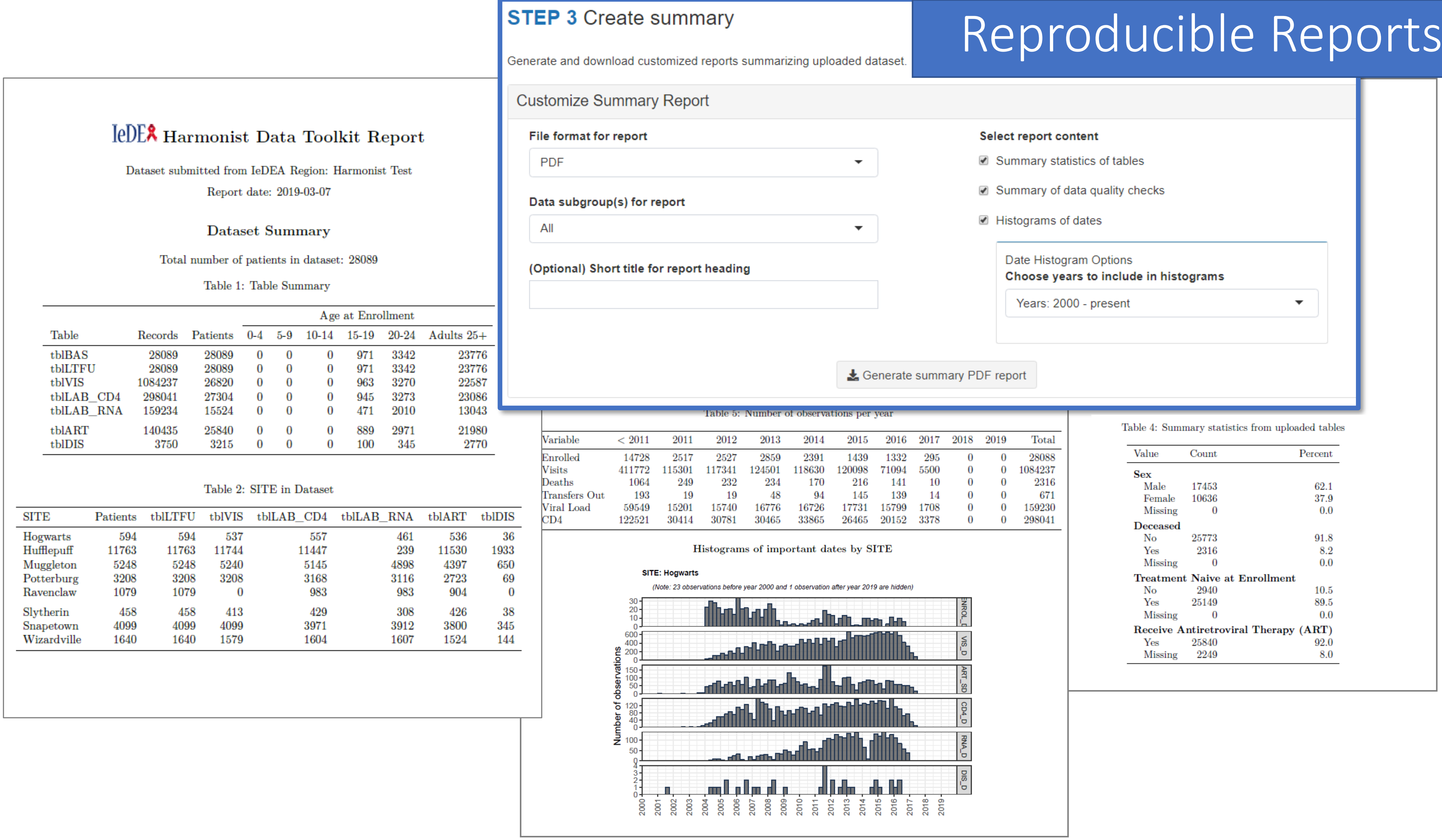
Your dataset contains **114 total errors in 12 error categories** including **28 invalid codes**

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

Restart session

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

Reproducible Reports



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

IeDEA 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

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

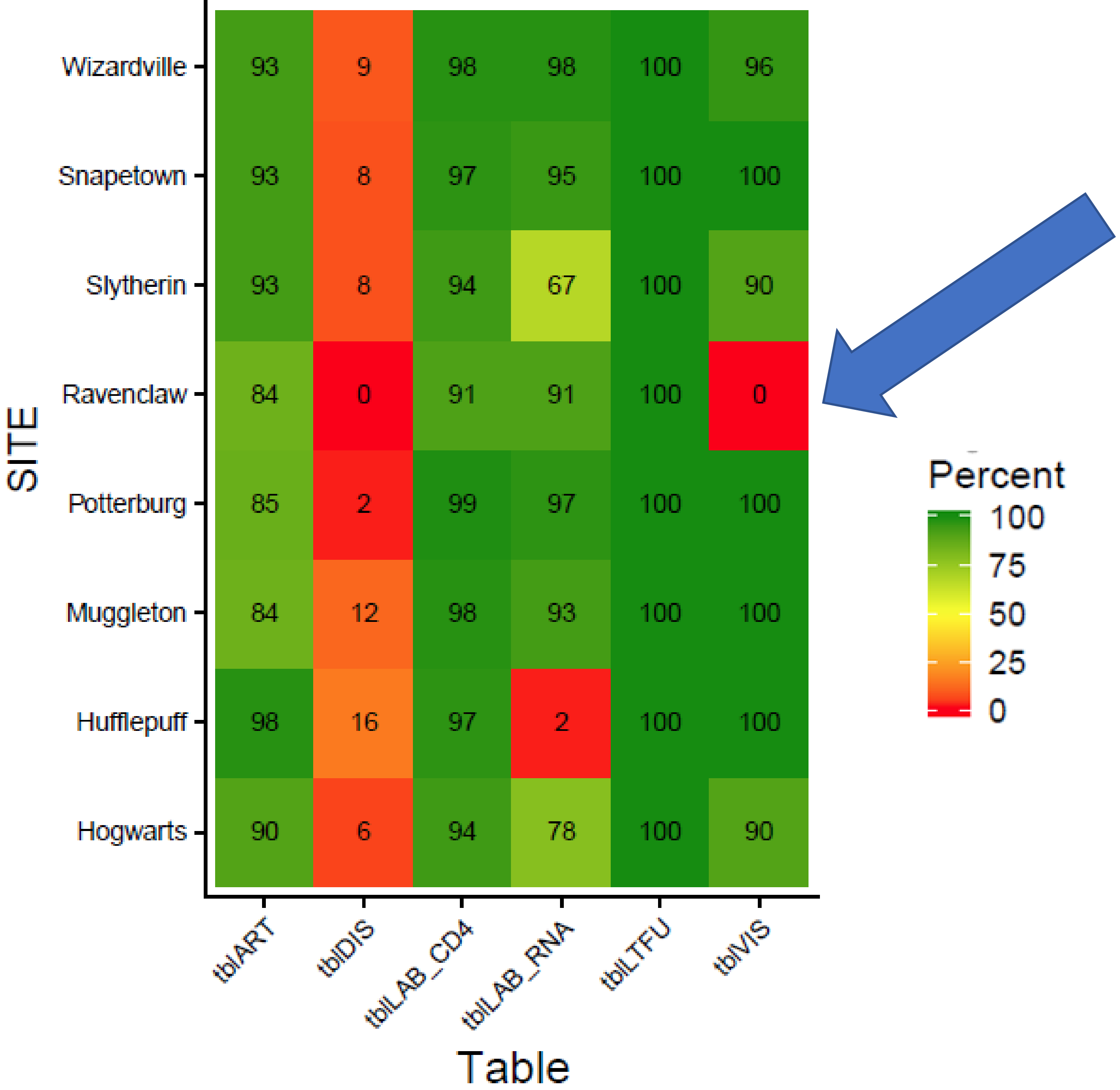
(Note: 23 observations before year 2000 and 1 observation after year 2019 are hidden)

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

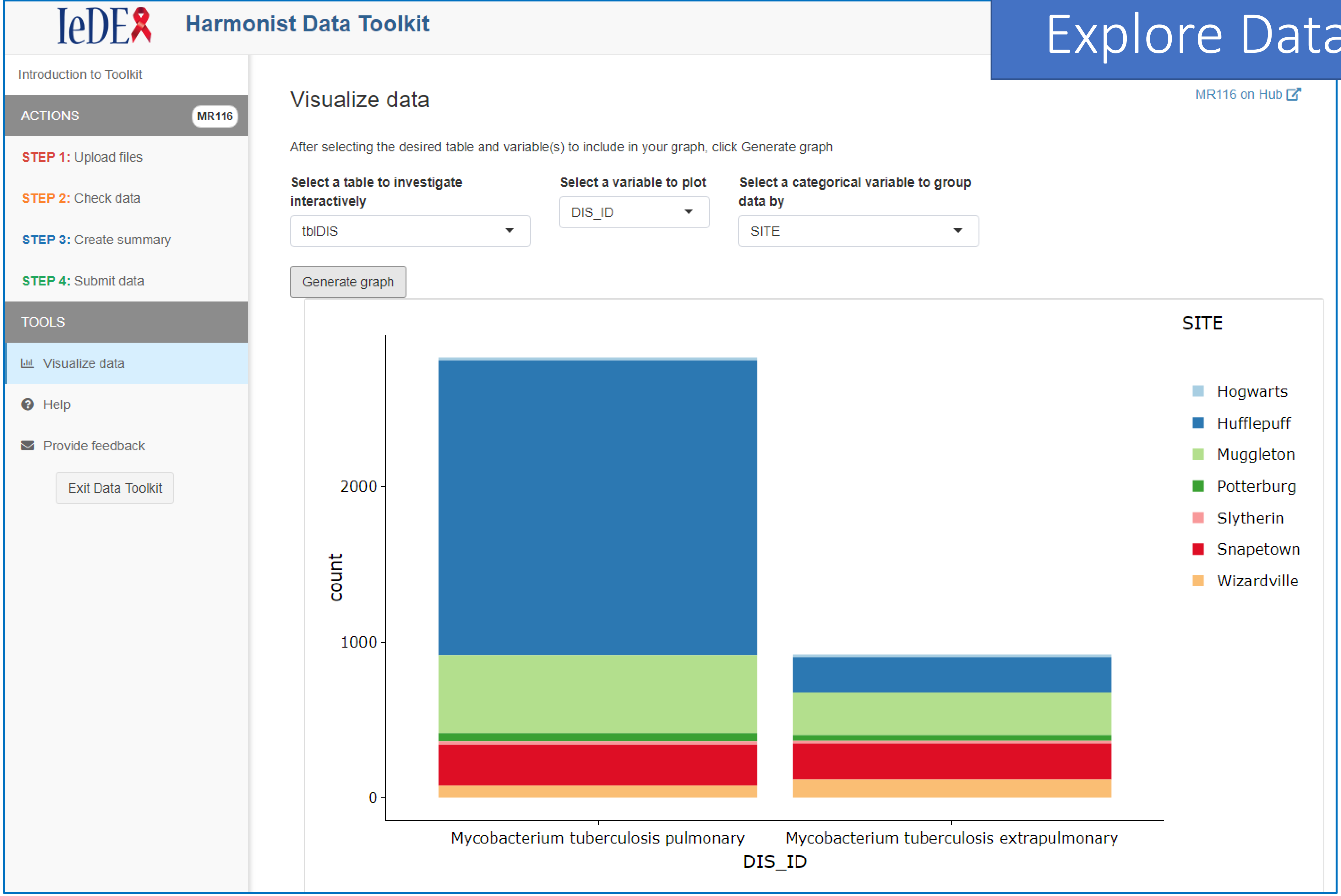
Example Report Content

Percent of Patients Included in Tables



SITE	tblART	tblDIS	tblLAB_CD4	tblLAB_RNA	tblLTFU	tblVIS
Wizardville	93	9	98	98	100	96
Snapetown	93	8	97	95	100	100
Slytherin	93	8	94	67	100	90
Ravenclaw	84	0	91	91	100	0
Potterburg	85	2	99	97	100	100
Muggleton	84	12	98	93	100	100
Hufflepuff	98	16	97	2	100	100
Hogwarts	90	6	94	78	100	90

Explore Dataset



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

Exit Data Toolkit

Visualize data

MR116 on Hub

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

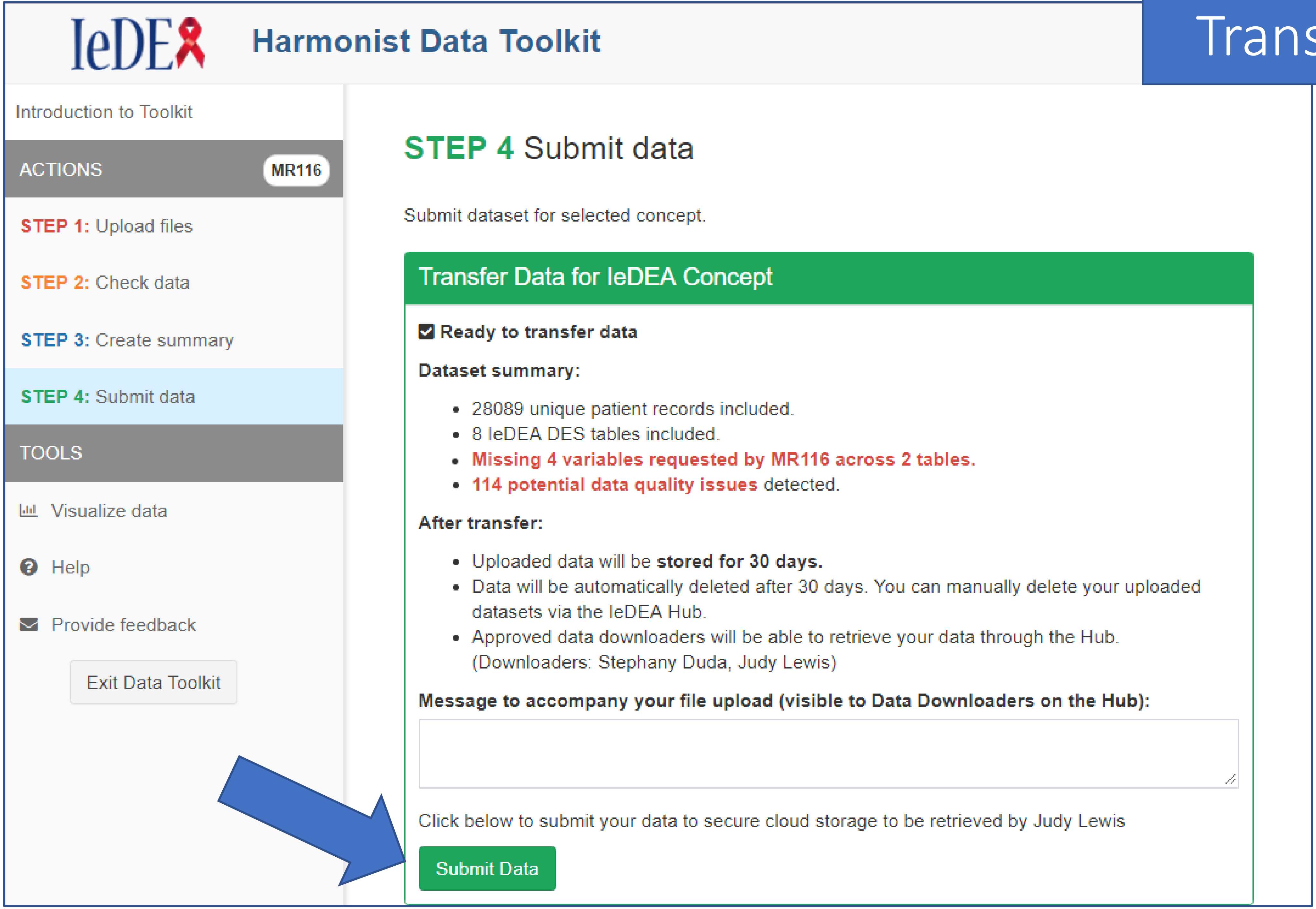
Select a table to investigate interactively: tblDIS

Select a variable to plot: DIS_ID

Select a categorical variable to group data by: SITE

Generate graph

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.

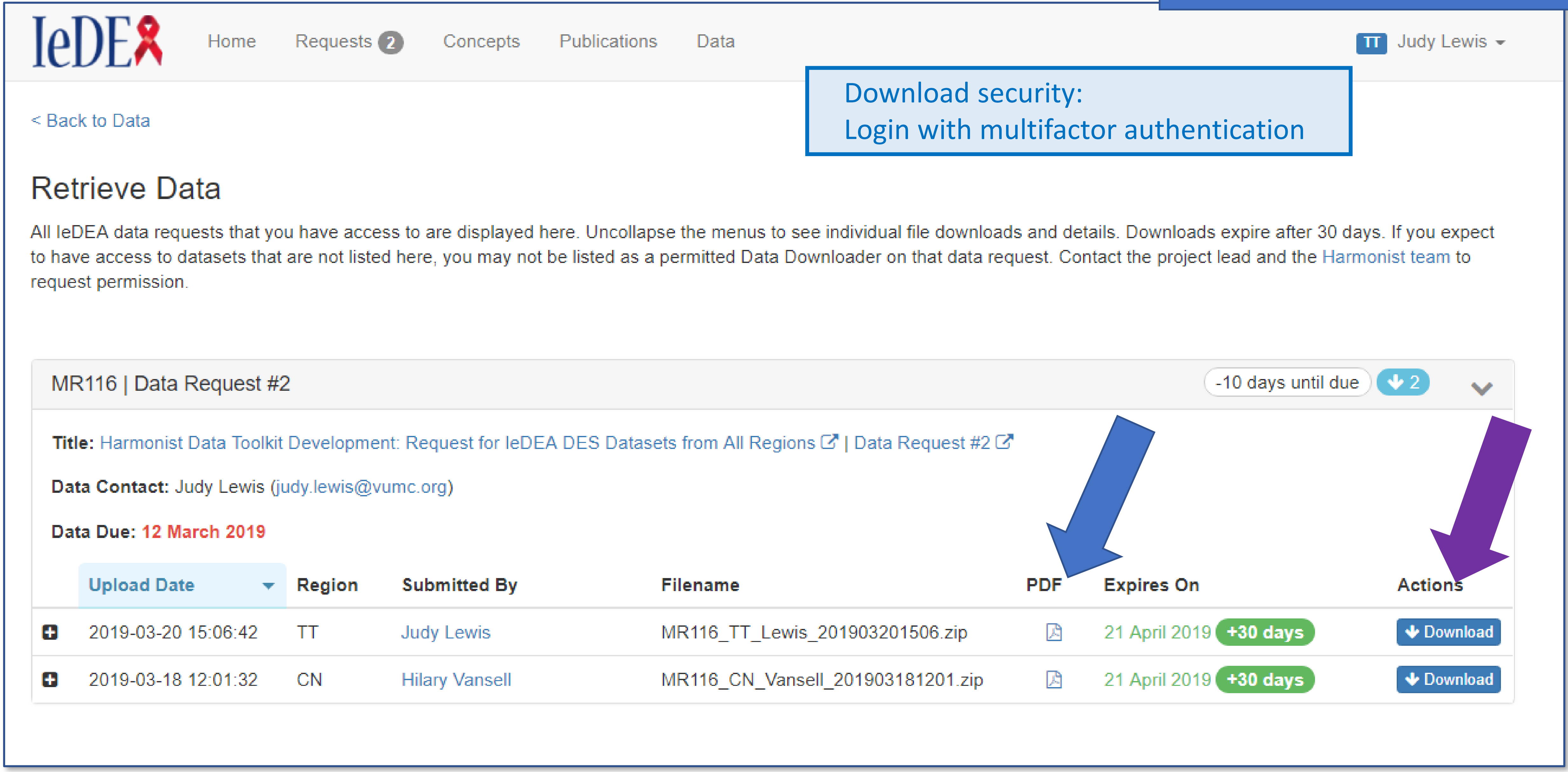
After transfer:

- 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

Download 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

Initial User Feedback

- “This is a fabulous tool and I will certainly utilize it to help manage the data in every aspect”
- “I feel that this toolkit will be a great asset to the regions in ensuring the efficient collation of data for concept analysis. It will help streamline and ensure data harmonization is achieved.”
- “Will save lots of time”
- “It will be very helpful when I receive data submission from other sites, so I know every table is submitted with IeDEA standards”
- “I like the heat map tables, so quick to understand what is happening in the data across programs”
- “immensely helpful in providing the cleanest data set possible”

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 allows software to be adapted to other domains.

Development & Opportunities

Future tasks:

- Expand data quality checks, report content
- Enhance code portability

We welcome all ideas and feedback!

- Dataset quality improvement metrics
- New application domains
- Advice/experience with similar tools

Code available github.com/IeDEA/Harmonist

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”)