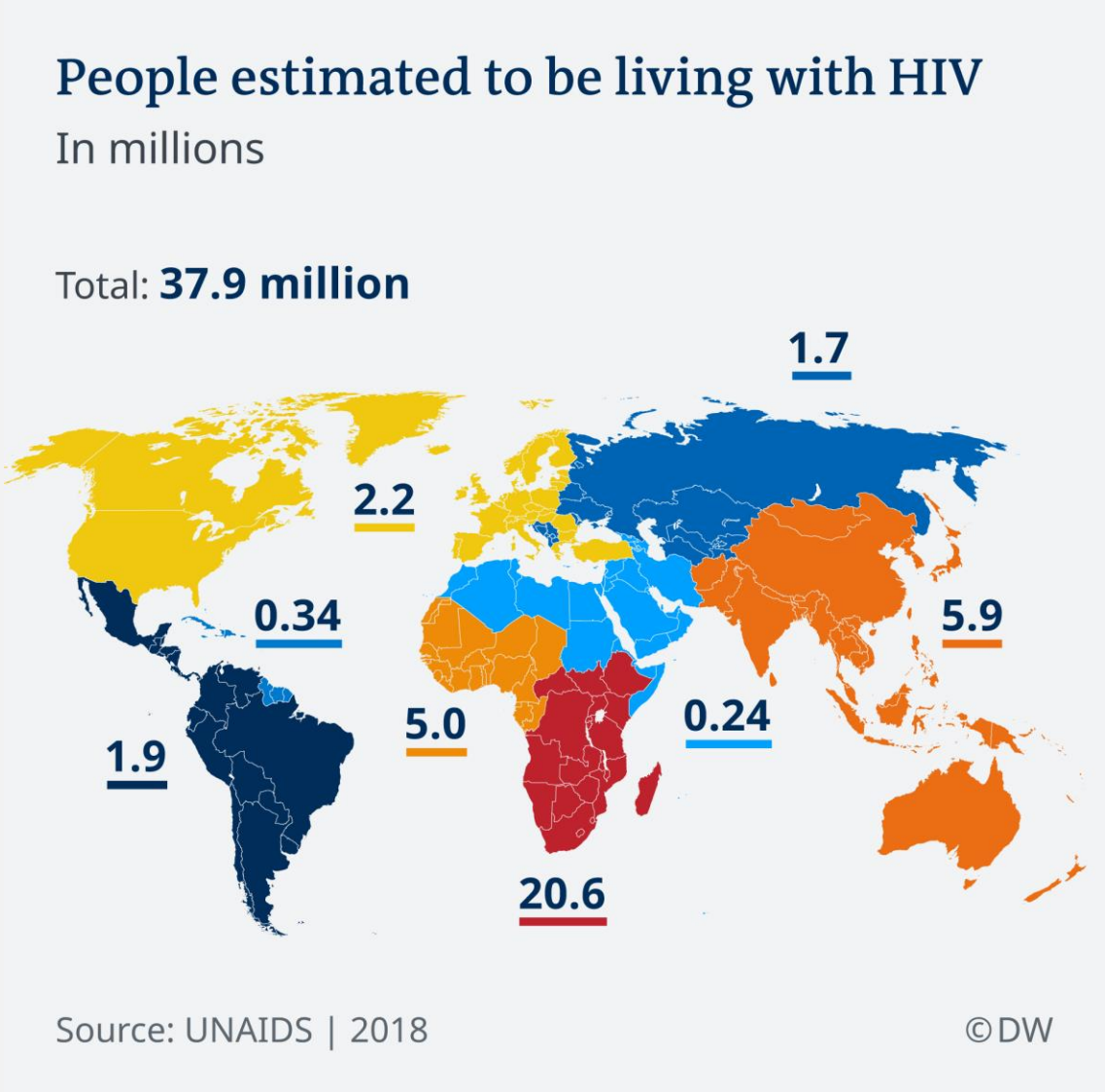
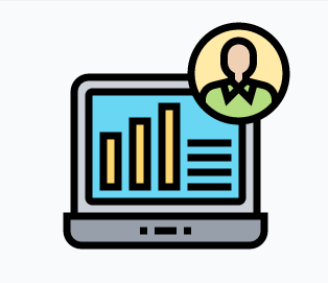
BIOMEDICAL INFORMATICS TOOLS FOR GLOBAL HIV/AIDS RESEARCH

Judy Lewis, PhD
Vanderbilt Institute for Clinical and Translational Research
Vanderbilt Department of Biomedical Engineering

THE CHALLENGE:

INVESTIGATORS NEED CLINICAL HIV/AIDS DATA FROM AROUND THE WORLD





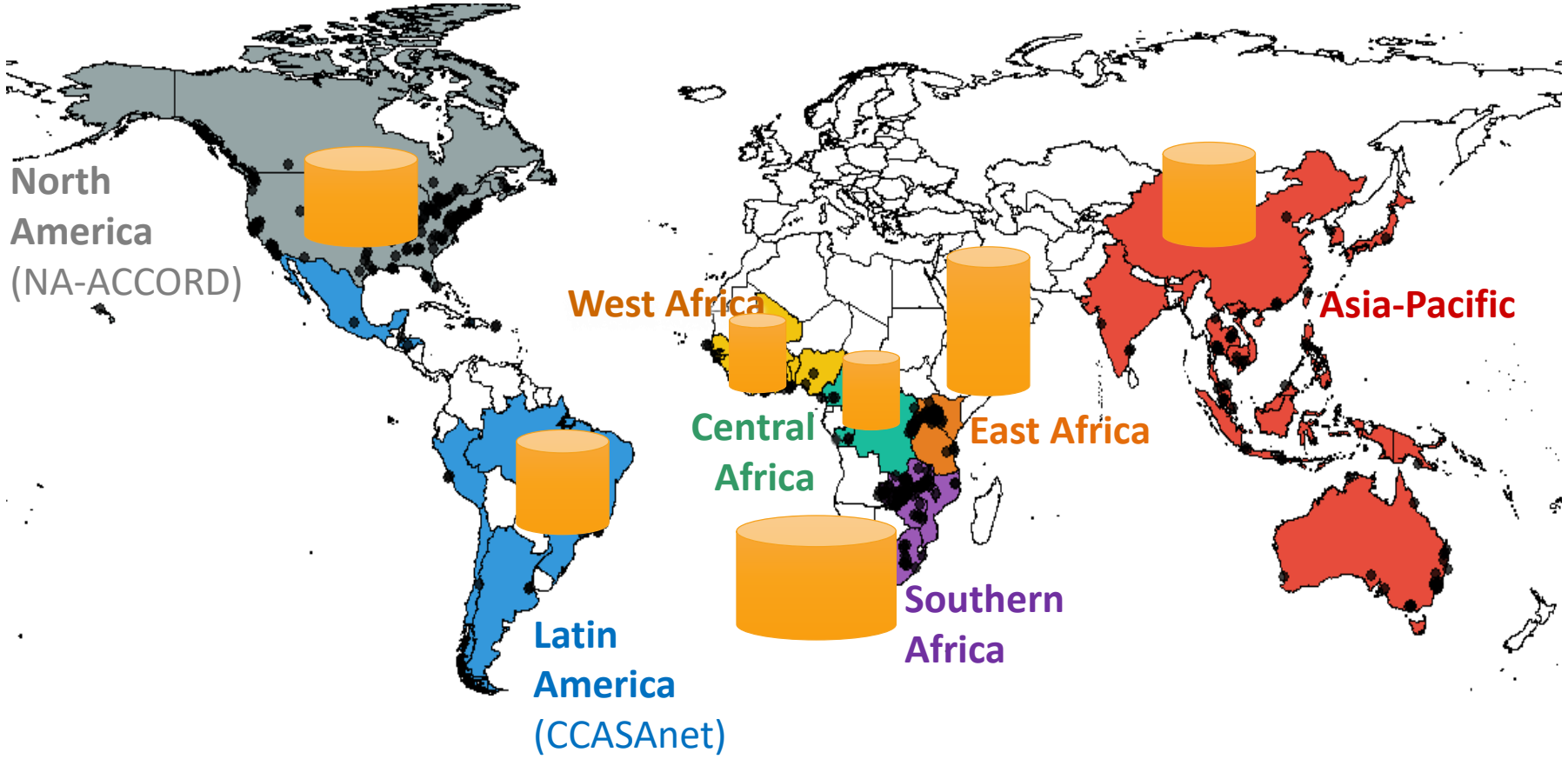
Does early antiretroviral treatment impact the risk of neurodevelopmental impairment among perinatally acquired HIV-infected preschool children?

Are women living with HIV at higher risk for cervical cancer?

Stunting and growth velocity of adolescents with perinatally acquired HIV: is it different between genders?

IeDEA

International epidemiologic Databases to Evaluate AIDS



- Established in 2005
- Funded by NIH
- 46 countries
- ~2 million patients
- 100's of publications

COLLABORATION WITH IEDEA REGIONAL DATA MANAGERS: DEFINED NEEDS

- Data must be in consistent **format**--can be merged by investigator
  - **Need** *common data model*: Capacity to evolve, 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 investigator
  - **Need** secure method for *submitting and receiving datasets*
- Regions must **communicate** to track requests, submit votes on concepts, etc
  - **Need** *project management* hub

1. COMMON DATA MODEL

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

SEX
Male
Female
Other
Unknown

Requires a Common Data Model

?

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

IeDEA Data Exchange Standard (DES)

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

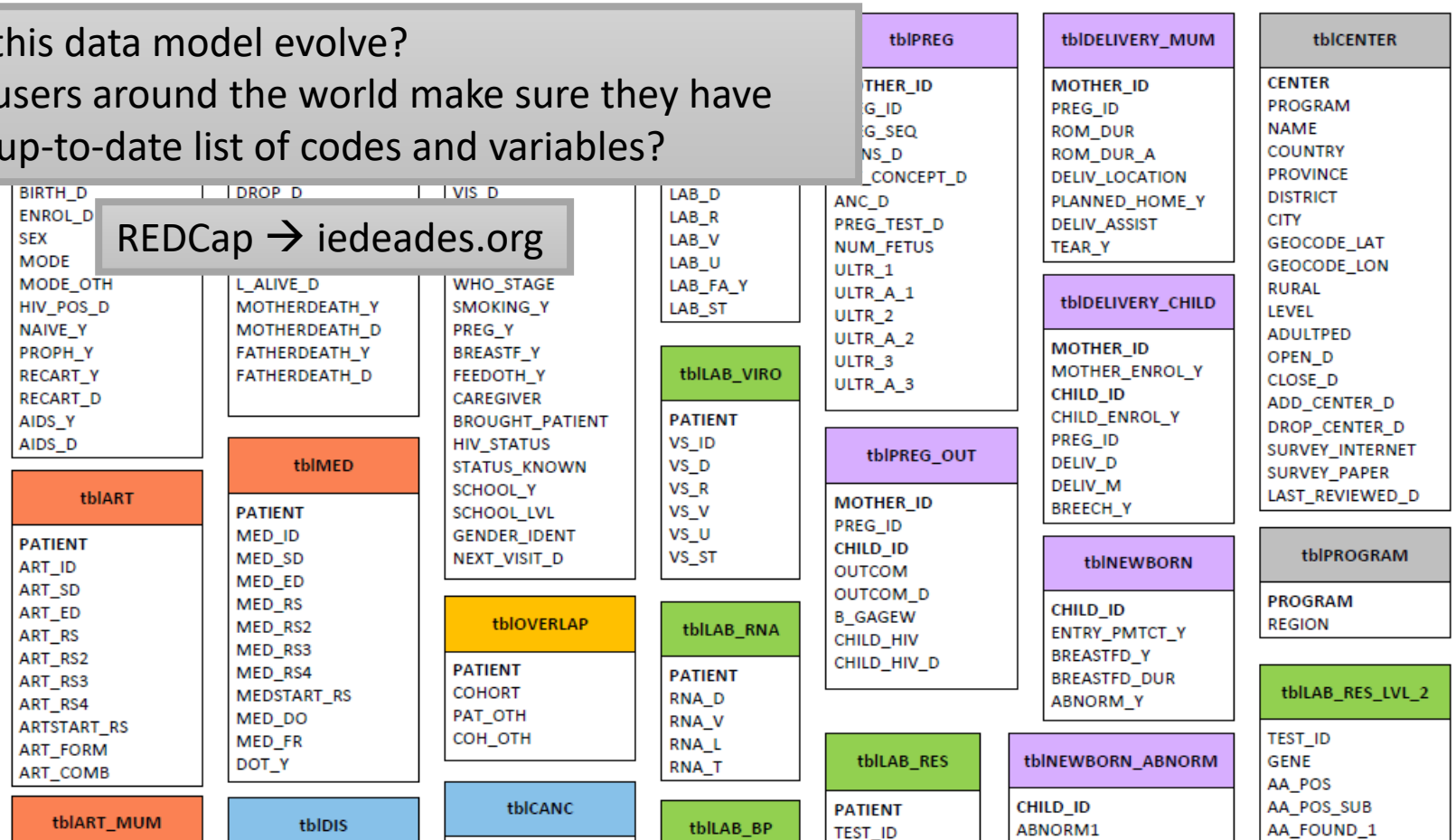
tblCENTER

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 the site was dropped from the cohort

How can this data model evolve?
How can users around the world make sure they have the most up-to-date list of codes and variables?

REDCap → iedeades.org



2. DATA QUALITY CHECKS

Snippet of previous code

```
## CHECK FOR UNEXPECTED CODING
badcodes(gender,c(1,2,9),basic)
# Mode of Infection
#   1 = homo/bisexual
#   2 = injecting drug user
#   3 = (1+2)
#   4 = haemophiliac
#   5 = transfusion, non-haemophilia related
#   6 = heterosexual contact
#   7 = (6+2)
#   8 = Perinatal
#   9 = Sexual contact (homo/hetero not specified)
#   10 = Sexual abuse
#   90 = other
#   99 = unknown
badcodes(mode,c(1:8,90,99),basic)
# ART naive upon enrollment
#   0 = No
#   1 = Yes
#   9 = Unknown
badcodes(naive_y,c(0,1,9),basic)
# Prior to enrollment, has the patient been exposed to antiretroviral therapy for p
#   0 = No
#   1 = Yes
#   9 = Unknown
badcodes(proph_y,c(0,1,9),basic)
#Has the patient ever received antiretroviral treatment? (excludes antiretroviral d
#   0 = No
#   1 = Yes
#   9 = Unknown
badcodes(recart_y,c(0,1,9),basic)
# Has patient ever been given an AIDS diagnosis? (clinical)
#   0 = No
#   1 = Yes
#   9 = Unknown
badcodes(aids_y,c(0,1,9),basic)
badcodes(birth_d_a,c("<",">","D","M","Y","U"),basic)
badcodes(enrol_d_a,c("<",">","D","M","Y","U"),basic)
badcodes(recart_d_a,c("<",">","D","M","Y","U"),basic)
badcodes(aids_d_a,c("<",">","D","M","Y","U"),basic)
```

Now, thanks to

REDCap™
Research Electronic Data Capture

```
errorFrame <- checkCodedVariables(errorFrame)
```

tblART_checks.R
tblART_MUM_checks.R
tblBAS_checks.R
tblCANC_checks.R
tblCENTER_checks.R
tblCEP_checks.R
tblDELIVERY_CHILD_check
tblDELIVERY_MUM_checks
tblDIS_checks.R
tblDIS_TB_checks.R
tblLAB_BP_checks.R
tblLAB_CD4_checks.R
tblLAB_checks.R
tblLAB_RES_checks.R
tblLAB_RES_LVL_2_checks.
tblLAB_RES_LVL_3_checks.
tblLAB_RNA_checks.R
tblLAB_VIRO_checks.R
tblLTFU_checks.R
tblMED_checks.R
tblNEWBORN_ABNORM_c
tblNEWBORN_checks.R
tblNEWBORN_checks_old.
tblPREG_checks.R
tblPREG_OUT_checks.R
tblPROGRAM_checks.R
tblVIS_checks.R

```
## NAME OF TABLE FOR WRITING QUERIES
tablename <- "tblBAS"
## NAMES EXPECTED FROM HICDEP+/IeDEAS DES
expectednames <- c("patient","birth_d","e
```

```
## CHECK FOR UNEXPECTED CODING
badcodes(gender,c(1,2,9),basic)
# Mode of Infection
#    1 = homo/bisexual
#    2 = injecting drug user
#    3 = (1+2)
#    4 = haemophiliac
```

Using R and the REDCap API, all of that is replaced by streamlined code:

```
for (tableName in names(uploadedTables())){
    errorFrame <- checkDatesInTable(errorFrame, tableName)
    errorFrame <- findMissingValues(errorFrame, tableName)
    errorFrame <- checkCodesInTable(errorFrame, tableName)
    errorFrame <- checkNumericValues(errorFrame, tableName)
    errorFrame <- PatientIDChecks(errorFrame, tableName)
    errorFrame <- globalDateChecks(errorFrame, tableName)
    errorFrame <- duplicateRecordChecks(errorFrame, tableName)
    errorFrame <- withinTableDateOrder(errorFrame, tableName)
}
```

```
antiretroviral therapy for proph
t? (excludes antiretroviral drugs
ical)
```

```
outoforder(birth_d,aids_d,basic)

## CHECK FOR DATES OCCURRING TOO FAR
futuredate(birth_d,basic)
futuredate(enrol_d,basic)
futuredate(recart_d,basic)
futuredate(aids_d,basic)

## CHECK FOR DUPLICATE PATIENT IDs
queryduplicates(patient,basic)
```

```
#    1 = Yes
#    9 = Unknown
badcodes(aids_y,c(0,1,9),basic)
badcodes(birth_d_a,c("<",">","D","M","Y","U"),basic)
badcodes(enrol_d_a,c("<",">","D","M","Y","U"),basic)
badcodes(recart_d_a,c("<",">","D","M","Y","U"),basic)
badcodes(aids_d_a,c("<",">","D","M","Y","U"),basic)

## QUERY PATIENTS WITH NO RECORD IN tblPROGRAM
if(exists("program") & exists("program",basic)){badrecord(program,basic,program)}
```

IeDEA DATA QUALITY CHECKING AND REPORTING

DESIGN GOALS	SOLUTION: HARMONIST DATA TOOLKIT
Minimal resources required by user →	Web interface (Shiny)
Easy for data managers to use →	Accepts variety of file types (SAS, Stata, SPSS, CSV)
Summarize datasets and errors →	Rmarkdown = reproducible reports
Should adapt to evolving data model →	REDCap (API → R)
Secure mechanism for file exchange →	AWS (API → R)

3. PROJECT MANAGEMENT HUB

IEDEAHUB.ORG

IeDEA Home Requests 4 Concepts Data Resources CA Stephany Duda

Upload Data

The following IeDEA Concepts have active Data Requests. Please select the request for which you wish to upload data. Data files should be in the **IeDEA Data Exchange Standard (IeDEA DES) format**. If you are transfering non-DES, non-data files, please use the File Transfer Tool instead.

Due Date	Concept	Title	Contact Person	Data Request	CN	Actions
2017-10-27 -2 days	MR014	Duration of first-line antiretroviral regimens in children: a global perspective (CIPHER)	Harmonist TestPerson (CN)		*No uploads*	Upload Data
2017-11-20 +22 days	MR077	Outcomes of children and adolescents treated with raltegravir in the IeDEA consortium	Gem Patten (SA)		Uploaded 2017-10-26	Upload Data View Upload
2017-11-27 +29 days	MR108	IeDEA-WHO collaboration: global analysis of the pre-ART cascade and delay from diagnosis to start of antiretroviral therapy in HIV-infected children aged 0-19 years	Cam Ha Ostinelli (SA)		Uploaded 2017-10-26	Upload Data View Upload

IeDEA HARMONIST PROJECT

- **iedeades.org:** Common data model
- **iedeahub.org**
  - Data requests
  - Research project management
- **iedeadata.org**
  - Data quality checking
  - Report generation
  - Secure file transfer

TOOL DEVELOPMENT: GLOBAL COLLABORATION

WORKFLOW

Workflow Begins in IeDEA Project Portal

IeDEA | Home | Requests 2 | Concepts | Publications | Data | TT Judy Lewis

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.

1

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

IeDEA | Home | Requests 2 | Conc

< 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

2

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

IeDEA Harmonist Data Toolkit

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

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

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

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

- [x] Summary statistics of tables
- [x] Summary of data quality checks
- [x] 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

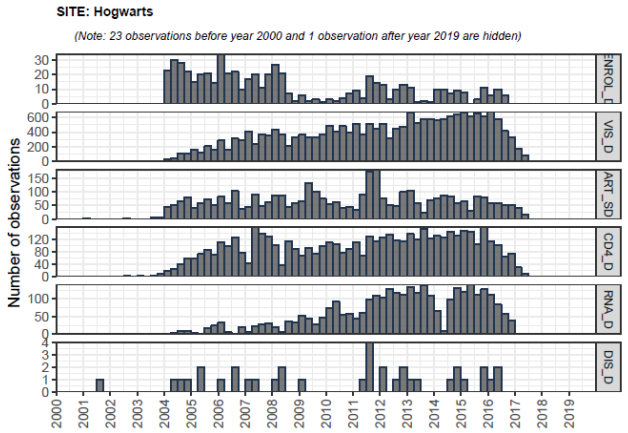
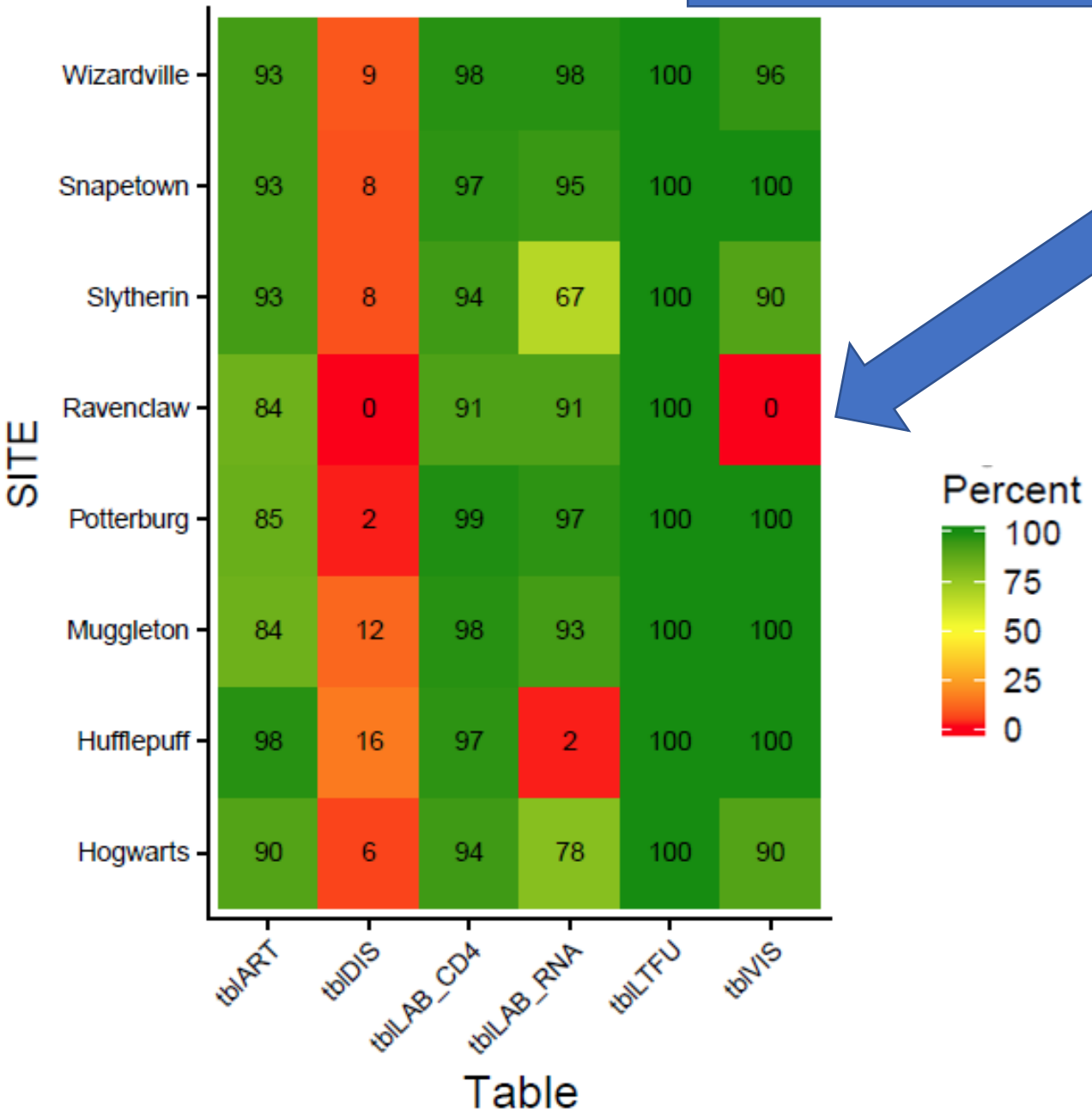


Table 4: Summary statistics from uploaded tables

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



IeDEA

5. 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

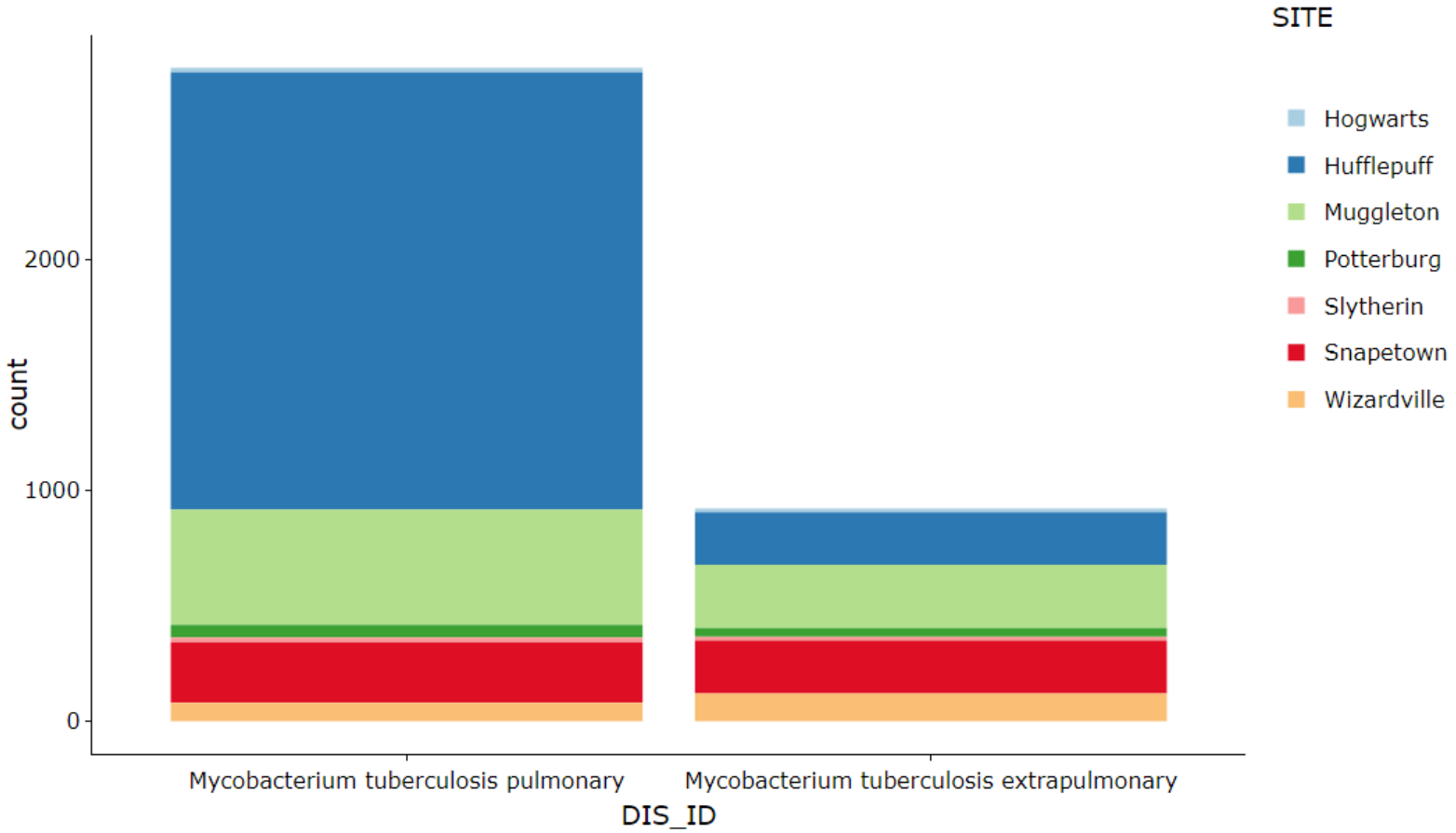
After selecting the desired table and variable(s) to include in your graph, click Generate graph

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



6. Transfer Data

IeDEA Harmonist Data Toolkit

IeDEA Home Requests 2 Concepts Publications Data TT Judy Lewis

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

Submit Data