Streamlining Data Sharing in a Global HIV Research Consortium

Stephany Duda PhD and Judy Lewis PhD
Vanderbilt University Medical Center
October 1, 2021
IeDEA – A global HIV research consortium

International epidemiology Databases to Evaluate AIDS
A global, NIAID-funded collaboration of seven regional HIV observational research networks with combined data on nearly two million persons living with HIV (PLWH)
IeDEA is a collaboration, not a single network

• Each region operates independently, has its own coordinating center, and conducts regional research.
• Regions come together for multiregional research projects.
• There is no overall coordinating center for IeDEA (as of July 2016).
• There has never been an overall data coordinating center for IeDEA.

(This turned out to be great)
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.

• 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.
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)
Researchers wanted to collaborate efficiently, but they needed the right tools to support them. We focused on three areas:

(1) Support for Data Model
   *IeDEA Data Exchange Standard* → Clear data standards

(2) Consortium Portal
   *Harmonist Hub* → Managed research workflow

(3) Data Quality and Sharing Tool
   *Harmonist Data Toolkit* → High quality data

Goal: address IeDEA’s needs but always look for generalizable solutions that can benefit others.
The Harmonist Project

1. Support for Data Model
   *IeDEA Data Exchange Standard*
   → Clear data standards

2. Consortium Portal
   *Harmonist Hub*
   → Managed research workflow

3. Data Quality and Sharing Tool
   *Harmonist Data Toolkit*
   → High quality data
The **IeDEA Data Exchange Standard (DES)** is a common data model for HIV data sharing within IeDEA. The document defines variable names, variable definitions, and code lists.
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

• 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
What is REDCap?

~5440 sites
142 countries
1.3M users

projectredcap.org
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

**Table Metadata**
- **Record ID**: Editing existing Record ID
- **Table Metadata**

**Table Definition**
- **Table name**: tblBAS
- **Table type**: One row per patient
- **Table definition (brief text)**: Basic Information

**Display Settings**
- **Display this table in human-readable documents and forms?**: Yes
- **Table display name**
- **Display order for this table**: 1
- **Text (HTML-formatted) to display before the table definition**: Every Patient ID must have one and only one entry in tblBAS.
Initial REDCap Mapping

Word version of the IeDEA DES
Part of Machine-Readable DES

(JSON formatted export via REDCap API)
Data Model Browser

website is auto-generated based on the content of the three REDCap projects.
Why this matters

• One “source of truth” for the data model
• One place for users to visit
• Easy to add new variables and codes
• Machine-readable data model can inform additional software tools

• REDCap templates and module code are shared on GitHub
  • [https://github.com/vanderbilt-redcap/data-model-browser](https://github.com/vanderbilt-redcap/data-model-browser)
  • Currently developing an “External Module” version (one-click installation) to share via the REDCap Repository of External Modules
# The Harmonist Project

<table>
<thead>
<tr>
<th>(1) Support for Data Model</th>
<th>Clear data standards</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>IeDEA Data Exchange Standard</em></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(2) Consortium Portal</th>
<th>Managed research workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Harmonist Hub</em></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(3) Data Quality and Sharing Tool</th>
<th>High quality data</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Harmonist Data Toolkit</em></td>
<td></td>
</tr>
</tbody>
</table>
The Harmonist Project

(1) Support for Data Model
   *IeDEA Data Exchange Standard*
   → Clear data standards

(2) Consortium Portal
   *Harmonist Hub*
   → Managed research workflow

How can IeDEA investigators request data for projects?

(3) Data Quality and Sharing Tool
   *Harmonist Data Toolkit*
Before...

- IeDEA investigators with an approved concept sheet developed a “Data Request” (Word document).
- Data request documents and due dates were circulated via email.
- A review of the first 4 data requests of 2016 (by different people) found that none aligned with the IeDEA DES.
We did have a private consortium website, but

- It only indexed concept sheets, not data requests or due dates
- We couldn’t tell which projects had received data and which ones were pending
- Nobody used it because we could never remember our passwords...

<table>
<thead>
<tr>
<th>Concept Number</th>
<th>Title</th>
<th>Lead Investigator</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.042</td>
<td>Outcomes of children and adolescents treated with raltegravir in the iDeA consortium</td>
<td>Mary Davi</td>
</tr>
<tr>
<td>2.041</td>
<td>Descriptions and outcomes of HIV infected patients treated for tuberculosis without microbiologic confirmation in HIV care programs within the iDeA consortium</td>
<td>Kara Woo</td>
</tr>
<tr>
<td>2.040</td>
<td>Diagnostic, treatment and outcome of extra-pulmonary tuberculosis in HIV-co-infected adults and children</td>
<td>Maria Balli</td>
</tr>
<tr>
<td>2.039</td>
<td>iDeA-WHO collaboration 2016: Adolescent treatment outcomes in the iDeA global consortium</td>
<td>Anna Soh</td>
</tr>
<tr>
<td>2.038</td>
<td>iDeA-WHO Collaboration: Global analysis of delays from ART eligibility to antiretroviral treatment initiation among adults</td>
<td>Deni Nas</td>
</tr>
<tr>
<td>2.037</td>
<td>iDeA-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</td>
<td>Valerie Ler</td>
</tr>
<tr>
<td>2.036</td>
<td>iDeA-WHO Collaboration: Global analysis of retention in care in initial HIV care and treatment program</td>
<td>Eliza Zanoli</td>
</tr>
<tr>
<td>2.035</td>
<td>Update of mortality estimates of HIV-infected patients for Spectrum</td>
<td>Congyi Yan</td>
</tr>
<tr>
<td>2.034</td>
<td>Association between clinic-level factors and individual retention, engagement, and loss to follow-up following ART initiation in the iDeA collaboration from 2009 to 2014</td>
<td>Bill L...</td>
</tr>
</tbody>
</table>
Designing the Harmonist Hub

We worked closely with IeDEA administrators, PIs, and investigators to design a web-based consortium management platform that could

• Index concept sheets and publications
• Coordinate Executive Committee review and voting for concept sheets, abstracts, manuscripts, etc.
• Implement a workflow with dashboards and automatic emails
• Generate metrics based on IeDEA’s scientific portfolio

• Not a separate website, just another module installed on REDCap
IeDEA Hub: Password-Free Login

If email address is in a list of active users, send an email with a URL that contains a one-time use token.
IeDEA Hub: Home Page

The IeDEA Hub has been updated to v2.10 on 2020-03-10. Contact us at harmonist@yume.org if you encounter any unexpected website behavior.

### Open Requests
- Concept Sheets
- Manuscripts
- Posters
- Fast Track
- Other Items

### Deadlines and Events

<table>
<thead>
<tr>
<th>Date</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020-07-06</td>
<td>+7 days AIDS 2020: Virtual begins</td>
</tr>
<tr>
<td>2020-07-06</td>
<td>+7 days TB RePORT-IeDEA Collaboration Workshop (register)</td>
</tr>
<tr>
<td>2020-08-04</td>
<td>+36 days IeDEA 4 FOA submission due date</td>
</tr>
<tr>
<td>2020-11-16</td>
<td>+140 days HIV Pediatrics Workshop begins (postponed)</td>
</tr>
</tbody>
</table>

### Active Data Calls
- Check the list of active data calls for IeDEA multi-regional projects.
- View Data Calls

### Hub Metrics
- View more
- Requests
- All IeDEA Hub requests by category.

### Recent Activity
- Per submitted a vote for Cervical Cancer Prevention and Care Cascade for wo... 2 days ago
The ieDEA Hub provides a set of tools that allow you to **request, submit, and retrieve** ieDEA data. The purpose of these tools is to make it easier to use the ieDEA Data Exchange Standard (DES) and share standardized and quality-checked data in a secure way.

**Request ieDEA data for your approved concept**

- **Explore** the different types of ieDEA data
- **Check and submit** data for an active data call
- **Retrieve** data uploaded for your project

**Recent Data Activity**

- **Beverly** uploaded EA data for MR180.
- **Karu** uploaded CN data for MR180.
- **Belinda** downloaded CN data for MR186.
Creating a Data Request: Step 1

**STEP 1: Setup**

The **leDEA Data Request Builder** is going to walk you through four steps:

1. Identify the concept for which you want patient-level data,
2. Pick the variables you need from the current version of the **leDEA Data Exchange Standard**,
3. Define your inclusion/exclusion criteria, suggested data submission date, and who will have permission to download the data, and
4. Preview your draft Data Request.

At the end, we'll guide you through **making your draft public** so you can share it with the leDEA Data Harmonization Working Group for feedback.

**Select Your Concept:**

```
MR116 - Harmonist Data Toolkit Development: Request for leDEA DES Datasets
For test requests, select MR000.
```

**Setup Type:**

- Create new data request
- Load draft

**Save and continue**
**Creating a Data Request: Step 2**

**STEP 2: Choose Variables**

*MRI16: Data Request #63*

Select the lD IDEA DES tables and variables you need for your analysis.

Keep in mind that the lDEA DES only specifies how data are transferred among regions; it does not define what data each region collects. Therefore **not all these variables will be available from lDEA regions**. Known variable availability (based on previous requests) is marked in the "Availability" column. The Data Harmonization Working Group will help you refine these selections.

<table>
<thead>
<tr>
<th>Select</th>
<th>Field</th>
<th>Availability</th>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PATIENT</td>
<td>Character or Numeric</td>
<td></td>
<td>Code to identify patient (Cohort Patient ID)</td>
</tr>
<tr>
<td></td>
<td>PROGRAM</td>
<td>Character</td>
<td></td>
<td>Direct one-to-one link from the patient to their program</td>
</tr>
<tr>
<td></td>
<td>BIRTH_D</td>
<td>YYYY-MM-DD</td>
<td></td>
<td>Birth date</td>
</tr>
<tr>
<td></td>
<td>BIRTH_D_A</td>
<td>Character</td>
<td>&lt;= Before this date</td>
<td>Date approximation for BIRTH_D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D = Exact to the date</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>M = Exact to the month</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y = Exact to the year</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&gt; = After the date</td>
<td></td>
</tr>
</tbody>
</table>
Creating a Data Request: Step 2.5

STEP 2: Choose Variables
Select the IeDEA DES tables and variables you need for your analysis.

Keep in mind that the IeDEA DES only specifies how data are transferred among regions; it does not define what data each region collects. Therefore **not all these variables will be available from IeDEA regions**. Known variable availability (based on previous requests) is marked in the “Availability” column. The Data Harmonization Working Group will help you refine these selections.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Select All</th>
</tr>
</thead>
<tbody>
<tr>
<td>tblBAS</td>
<td></td>
</tr>
<tr>
<td>tblLTU</td>
<td></td>
</tr>
<tr>
<td>tblVIS</td>
<td></td>
</tr>
<tr>
<td>tblLAB_CD4</td>
<td></td>
</tr>
<tr>
<td>tblLAB_RNA</td>
<td></td>
</tr>
<tr>
<td>tblART</td>
<td></td>
</tr>
<tr>
<td>tblDIS</td>
<td></td>
</tr>
<tr>
<td>tblCENTER</td>
<td></td>
</tr>
<tr>
<td>tblPROGRAM</td>
<td></td>
</tr>
<tr>
<td>tblART_MUM</td>
<td></td>
</tr>
<tr>
<td>tblGANG</td>
<td></td>
</tr>
</tbody>
</table>
**STEP 3: Add Details**

The form below collects some extra information needed to generate the Data Request.

<table>
<thead>
<tr>
<th>Inclusion criteria (list variable names if possible)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Include all HIV-positive older adolescents and adults (( \geq 16 ) years of age at enrollment) in the IeDEA cohorts from 1995 onward.</td>
</tr>
<tr>
<td>- Both ART-naive and non-naive patients at enrollment are eligible.</td>
</tr>
<tr>
<td>- Include all years of data from CENTERS that diagnose cryptococcal meningitis (CM) and provide data on CM/extra pulmonary cryptococcosis, even if the CENTER only started diagnosing CM recently. (e.g., If a clinic started in 2013, but CM diagnoses only began in 2015, include all data from 2013 onward.)</td>
</tr>
<tr>
<td>- Include both patients with CM and patients without CM from these CENTERS.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exclusion criteria (list variable names if possible)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Exclude CENTERS that do not systematically collect any opportunistic infection (OI) data.</td>
</tr>
<tr>
<td>- Exclude CENTERS that have never diagnosed CM.</td>
</tr>
<tr>
<td>- Exclude CENTERS that do not collect or provide data on CM for IeDEA (This means exclude CENTERS that do not ever collect or provide DIS_ID = ‘CRCO’, see definition of CM below.)</td>
</tr>
</tbody>
</table>
Step 3.5

New content guided by the IeDEA Data Harmonization Working Group (not part of Word document Data Requests)
STEP 4: Preview Data Request

Preview the document below. You can return to any previous step to edit the draft. The next step will save the draft Data Request to your library and create a PDF. You can still return to edit the document afterwards and regenerate a new PDF.

International epidemiology Databases to Evaluate AIDS (IeDEA)
DATA TRANSFER REQUEST – MR131

Clinical Outcomes in HIV-positive patients with Cryptococcal Meningitis in the IeDEA global cohort consortium

Data Due: 17 October 2018

Research Contact(s)

Annie
user.email@example.com
Brenda
user.email@example.com

Data Contact
Steps Complete ✅

Your Data Request has been generated successfully. You can review and download the PDF below or download a ZIP file with an HTML and a PDF version. If you need to make changes, you can go back to edit your data request.

The next step is to make your draft public and request a review by the leDEA Data Harmonization Working Group. The working group chairs may ask to meet with you to revise this document together, or may ask if you can discuss this data request on the next WG call.

< Back to Edit Data Request

International epidemiology Databases to Evaluate AIDS (leDEA)

DATA TRANSFER REQUEST – MR131

Clinical Outcomes in HIV-positive patients with Cryptococcal Meningitis in the leDEA global cohort consortium
International epidemiology Databases to Evaluate AIDS (IeDEA)

DATA TRANSFER REQUEST – MR131

Clinical Outcomes in HIV-positive patients with Cryptococcal Meningitis in the IeDEA global cohort consortium

Data Due: 17 October 2018

Research Contact(s)

Anna
user.email@example.com

Brenda
user.email@example.com

Data Contact

Stephany Duda
user.email@example.com

Automatically Generated Data Request
Why this matters

• No more manually assembled Word documents! Point-and-click data request builder saves time.

• The list of requestable variables is pulled from the machine-readable IeDEA DES. The tool automatically adds required variables if not selected
  • Therefore all data requests will be compliant with the standard.

• One “source of truth” for data requests (on the Hub)

• All resources in GitHub and the REDCapRepo

And...

• Data requests are stored in a machine-readable format now too.
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.

<table>
<thead>
<tr>
<th>Due Date</th>
<th>Data Request Details</th>
<th>Data Contact</th>
<th>AP</th>
<th>CA</th>
<th>CN</th>
<th>EA</th>
<th>NA</th>
<th>SA</th>
<th>WA</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2021-09-08</td>
<td>MR190 Infectious and non-infectious causes of liver disease among adults living with HIV from low and middle income countries, the Sentinel Research Network of IeDEA</td>
<td>Kare</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>Change Status</td>
</tr>
<tr>
<td>2021-09-15</td>
<td>MR181 Diagnostic value of serological markers for liver steatosis and fibrosis in people living with HIV: a multicenter study from the IeDEA Network</td>
<td>H</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>Change Status</td>
</tr>
<tr>
<td>2021-09-20</td>
<td>MR185 Associations of mental health and substance use disorders with virological failure by sex among aging adults living with HIV from low/middle-income countries: the Sentinel Research Network of IeDEA</td>
<td>Keith</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>Change Status</td>
</tr>
</tbody>
</table>
The Harmonist Project

(1) Support for Data Model
   *IeDEA Data Exchange Standard*
   → Clear data standards

(2) Consortium Portal
   *Harmonist Hub*
   → Managed research workflow

(3) Data Quality and Sharing Tool
   *Harmonist Data Toolkit*
   → High quality data
IeDEA Harmonist Data Toolkit

• Collaborative project with all seven IeDEA regions
• Web application
• Developed with freely available/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
Flow of IeDEA Data for Global Projects

IeDEA Sites

Region #1 Data Center

Region #2 Data Center

IeDEA Regions

Global IeDEA Projects
“Multi-regional”

Harmonist Data Toolkit sits here
Workflow Begins in IeDEA Project Portal (IeDEA Hub)
1. 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

<table>
<thead>
<tr>
<th>Title</th>
<th>Harmonist Data Toolkit: Request for IeDEA DES Datasets from AI Regions</th>
</tr>
</thead>
</table>

Hub Pages

- MR116 on Hub
- Data Specification
- PDF

Requested Tables

- RHIN
- BAHfit
- BAH
- BAH_com
- BAH_DNA
- DHCENTER
- DHP

Requested Data Format

- SAS

Contacts

- Judy Lewis (TT), Vanderbilt University
- Stephanie Duca (CN), Vanderbilt University
- Judy Lewis (TT) (Data contact), Vanderbilt University

Data Downloaders

- Stephanie Duca (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, Stats, SPSS, or a ZIP containing multiple files of this type.

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

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:

- BAHfit
- BAH
- BAH_DNA
- BAH_com
- DHCENTER
- DHP

Launch with Sample Data
### 2. File Review/Data Quality Checks

#### STEP 1: Upload files

Choose the files containing your iDEA 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:

- `NEXT_VISIT_D, NEXT_VISIT_D_A`
- `DROP_CICAM` Table missing (2 variables)

**Summary of Uploaded iDEA Tables**

<table>
<thead>
<tr>
<th>Table</th>
<th>Records</th>
<th>iDEA DES Variables</th>
<th>Extra Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR116S</td>
<td>1084237</td>
<td>PATIENT, CENTER, VIS_D, VIS_D_A, CDC_STAGE, WHO_STAGE, PREG_Y</td>
<td>GENDER_ID, SITE</td>
</tr>
<tr>
<td>MR116U</td>
<td>298041</td>
<td>PATIENT, CD4_D, CD4_D_A, CD4_V, CD4_U</td>
<td>CENTER, SITE</td>
</tr>
<tr>
<td>MR116A_D</td>
<td>150234</td>
<td>PATIENT, RNA_D, RNA_D_A, RNA_V</td>
<td>CENTER, SITE</td>
</tr>
<tr>
<td>MR116D</td>
<td>140435</td>
<td>PATIENT, ART_ID, ART_SD, ART_SD_A, ART_ED, ART_ED_A</td>
<td>CENTER, SITE</td>
</tr>
<tr>
<td>MR116E</td>
<td>3750</td>
<td>PATIENT, DIS_ID, DIS_D, DIS_D_A, DIS_ED, DIS_ED_A, DIS_OUTCOME</td>
<td>CENTER, SITE</td>
</tr>
<tr>
<td>MR116T</td>
<td>9</td>
<td>CENTER, PROGRAM, NAME, COUNTRY, PROVINCE, DISTRICT, CITY, GEOCODE_LAT, GEOCODE_LON, RURAL, LEVEL, ADULTPED, OPEN_D, CLOSE_D</td>
<td>REGION, ADD_CENTER_D, DROP_CENTER_D, SITE</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Error description</th>
<th>Severity</th>
<th>Count</th>
<th>View Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future date: ENROL_D</td>
<td>Error</td>
<td>1</td>
<td>View Detail</td>
</tr>
<tr>
<td>Invalid Code: MODE</td>
<td>Error</td>
<td>1</td>
<td>View Detail</td>
</tr>
<tr>
<td>Invalid Code: RECART_D_A</td>
<td>Error</td>
<td>2</td>
<td>View Detail</td>
</tr>
<tr>
<td>Invalid Code: HAART_D_A</td>
<td>Error</td>
<td>2</td>
<td>View Detail</td>
</tr>
<tr>
<td>BIRTH_D before 1920</td>
<td>Warn</td>
<td>3</td>
<td>View Detail</td>
</tr>
<tr>
<td>Date before 1900 AIDS_D</td>
<td>Warn</td>
<td>1</td>
<td>View Detail</td>
</tr>
</tbody>
</table>

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.

Continue to Step 3

Restart session
Start over and upload a revised or different dataset.
4. Reproducible Reports
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

Submit 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.

<table>
<thead>
<tr>
<th>MR116</th>
<th>Data Request #2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> Harmonist Data Toolkit Development: Request for IeDEA DES Datasets from All Regions</td>
<td>Data Request #2</td>
</tr>
<tr>
<td><strong>Data Contact:</strong> Judy Lewis (<a href="mailto:judy@harmonist.com">judy@harmonist.com</a>)</td>
<td></td>
</tr>
<tr>
<td><strong>Data Due:</strong> 12 March 2019</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Upload Date</th>
<th>Region</th>
<th>Submitted By</th>
<th>Filename</th>
<th>PDF</th>
<th>Expires On</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-03-20 15:06:42</td>
<td>TT</td>
<td>Judy Lewis</td>
<td>MR116_TT_Lewis_201903201506.zip</td>
<td>Download</td>
<td>21 April 2019</td>
<td>+30 days</td>
</tr>
<tr>
<td>2019-03-18 12:01:32</td>
<td>CN</td>
<td>Hilary Vandell</td>
<td>MR116_CN_Vandell_201903181201.zip</td>
<td>Download</td>
<td>21 April 2019</td>
<td>+30 days</td>
</tr>
</tbody>
</table>

**Download security:**
Login with multifactor authentication
Toolkit Impact on Data Quality

• The Toolkit was **released in March 2019**
  • > 800 datasets have been checked and summarized
  • 1,800 to 986,089 patients per dataset
  • 81 secure file transfers for 17 multiregional IeDEA projects
• Regional data managers **uploaded datasets and reviewed data quality results multiple times** before final transfer
• Data managers used Toolkit data quality reports to **improve datasets** before transfer
• The number and types of **errors decreased** with each iteration of Toolkit use, average of 61% decrease in errors
Real De-identified Example:
Region responding to one data request

No errors in final submission
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.
Lessons Learned

• Close collaboration with stakeholders and users is key
  • Frequent contact with stakeholders (e.g., admins, data managers, working groups)
  • Structured testing and training exercises with users
  • In-person (now virtual) meetings to collaborate on design
• Using REDCap as a platform facilitates development and sharing
• Defining details in REDCap and using the REDCap API make it possible to design tools that adapt with the changing data model

Tools have been well received by IeDEA; multiregional concept proposals and data requests have doubled.
Development & Opportunities

• Future tasks
  • Expand features
  • Gather insights from other consortia
  • Improve generalizability of tools

• We welcome all questions and input!
  • Contact us at harmonist@vumc.org

• Code available
  • github.com/IeDEA/Harmonist
  • github.com/vanderbilt-redcap
  • redcap.vanderbilt.edu/consortium/modules/index.php
Thank You

• Harmonist and REDCap technical teams
• IeDEA leadership, Data Harmonization Working Group, and collaborators
• Other consortia helping to make our processes generalizable!

Research reported in this publication was supported by the National Institute Of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R24AI124872. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.