National Cancer Institute

Data Integration and Imaging Informatics Project (DI-cubed) Status Report

December 04, 2017

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Project Objectives

- Demonstrate that standards such as BRIDG, CDISC and DICOM will support interoperability
- Develop a prototype that supports the ability to query across disparate data sets by leveraging the clinical data and linking them to images.
    - Could inform development of NCI Research Data Commons
- Provide mappings to CDISC SDTM for the scoped data sets

Standards referenced

- BRIDG (ISO 14199)
  - https://bridgmodel.nci.nih.gov/
  - Stakeholders: CDISC, FDA, HL7, ISO, NCI
  - Produce a shared view of the dynamic and static semantics for the domain of basic, pre-clinical, clinical, and translational research and its associated regulatory artifacts.
  - Balloted by CDISC, HL7 and ISO
- DICOM (ISO 12052)
  - International standard for medical images and related information
- CDISC SDTM (Study Data Tabulation Model)
  - Required for submission to FDA
  - https://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm248635.htm

BRIDG 5.0

- Connecting two ISO standards by including representative portions of DICOM into BRIDG which allows to link clinical data and images
- Contains new semantics from the harmonization of concepts from NCI's Surveillance, Epidemiology and End Results (SEER)
  - BRIDG can now support patient clinical care data in addition to the clinical trial data
- BRIDG 5.0 has be balloted by HL7, CDISC and ISO in a joint ballot cycle process.

Data Acquisition

- Scoped to publicly available data
- Focused on Breast Cancer

Collection/Data Set	Source	Cancer Type	Color Code
I-Spy1	TCIA caArray / caIntegrator	Breast Cancer	
I-SPY	TCIA	Breast Cancer	
Breast-MRI-NACT-Pilot	TCIA	Breast Cancer	
TCGA-BRCA	TCIA; TCGA Legacy Data GDC	Breast Cancer	
Breast Diagnosis	TCIA	Breast Cancer	
Ivy-Gap	TCIA	Glioblastoma	
Clinical Study	Duke / COH	Breast Cancer	
SEER	Cancer Registries	Breast Cancer	

High Level Approach

Concept Metadata:

ISPY1 caIntegrator

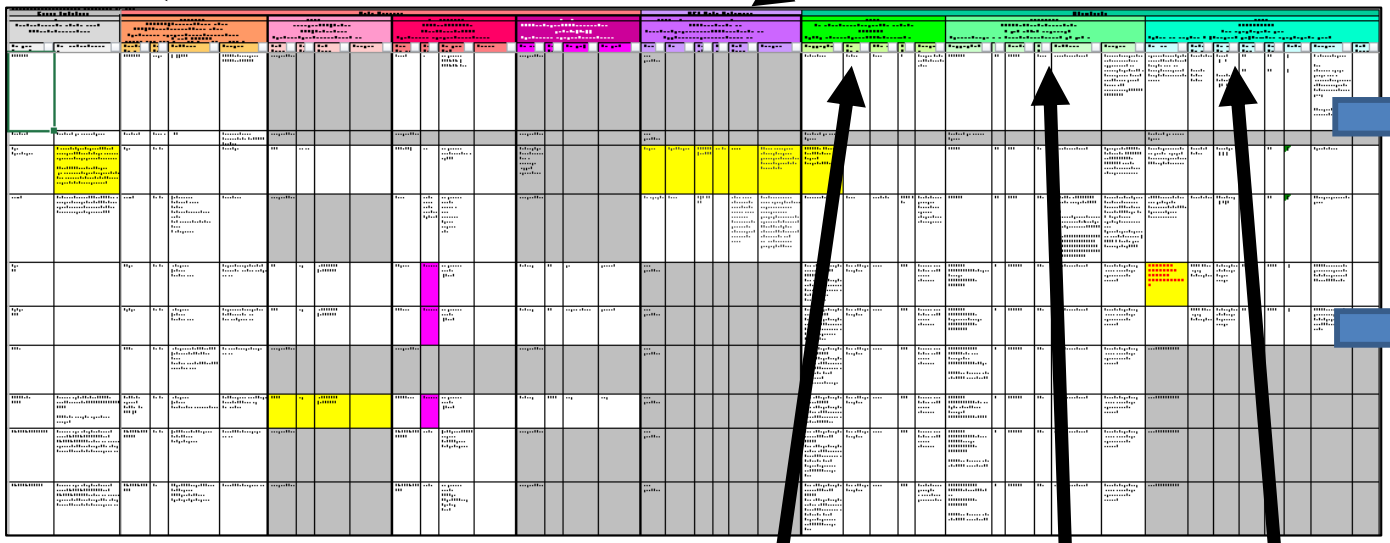
Breast-MRI-NACT-Pilot

Breast Diagnosis

ISPY1 TCIA

TCGA-BRCA

Concepts from all sources are aligned in the master mapping spreadsheet with the 3 international standards. Core metadata are identified for the prototype. DI-cubed prototype repository extended as needed to support cross-source querying



Map/Gap Analysis

Data design for

Repository

Standards Metadata:

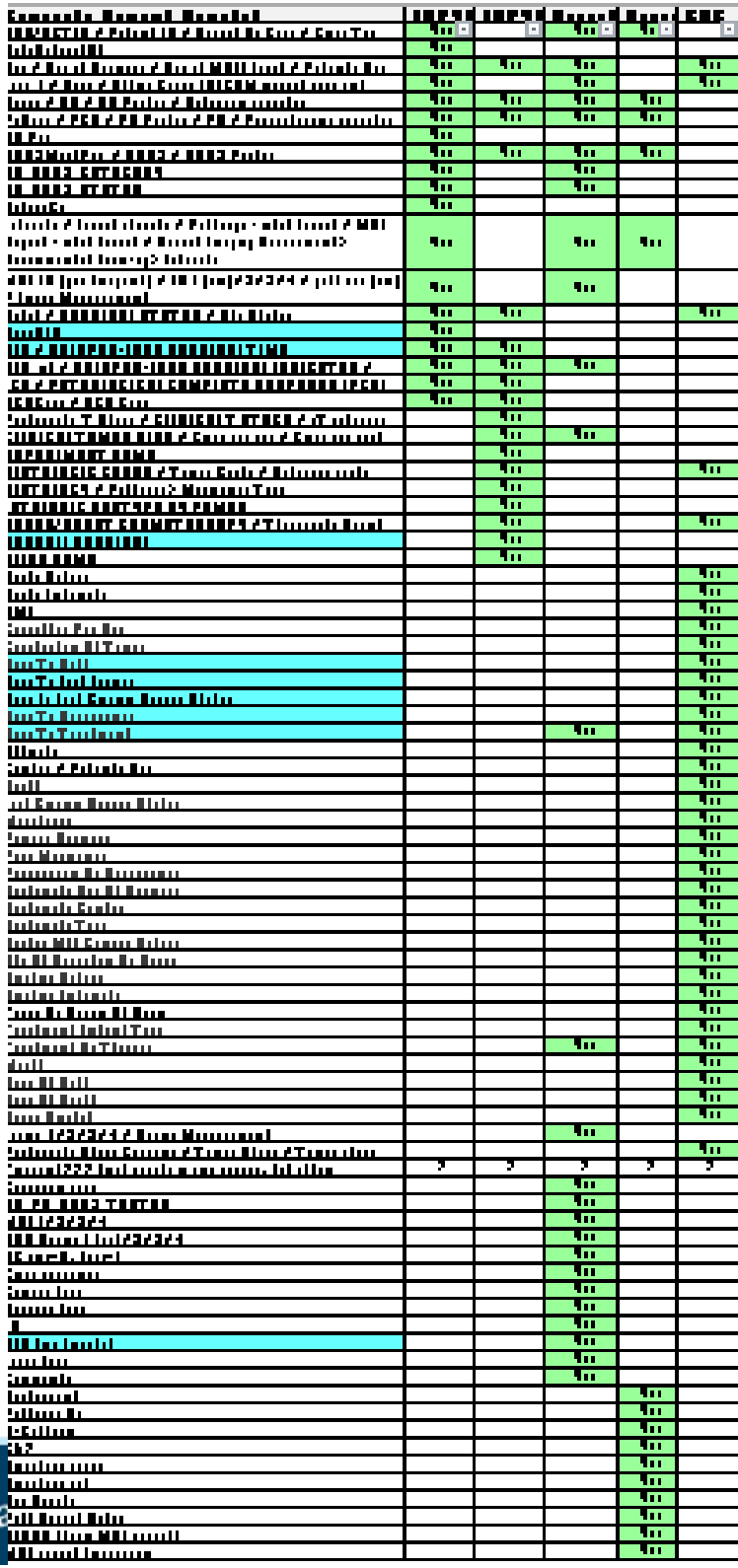
BRIDG

SDTM

DICOM

Findings – Little Overlap

- 84 distinct data elements from five sources (without TCGA-BRCA and IVY-Gap)
  - No data elements occur in all 5 sources
  - 4 occur in 4 sources (4.7%)
  - 5 occur in 3 sources (3.4%)
  - 11 occur in only 2 sources (13.1%)
  - 64 occur in only 1 source (76.2%)



Findings
Lack of Semantic Consistency across the Data Sources for a given Data Element

- Age – several variations:
  - **Age** (ISPY1 TCIA) – patient age (but not tied to time point)
  - **Age** (ISPY1 caIntegrator) – no definition
  - **Age at MRI1** (yrs) (Breast-MRI-NACT-Pilot) – timepoint-specific
  - Pathology - **Age Decade** (Breast Diagnosis) – not really the same concept but possibly attempt at de-identification or summary
  - **Age at Diagnosis** (GDC)
  - **Patient's Age** (DICOM)
  - **Relationship Age at Diagnosis** (GDC) – age or relative when diagnosed
  - **Year of Birth** and **Year of Death** (GDC) – only dates provided

Findings – Differences in Value Set Bindings

- Race – Core values map across all sources using them, but some differences exist and a scheme for handling them will be needed

ISPY1 TCIA Valid Values:
- 1=Caucasian
- 3=African American
- 4=Asian
- 5=Native Hawaiian/
- Pacific Islander
- 6=American Indian/
- Alaskan Native
- 50=Multiple race

Breast-MRI-NACT-Pilot example values:
- african-amer
- asian
- caucasian
- hispanic
- not given
- other

GDC Valid Values:
- White
- american indian or alaska native
- black or african American
- Asian
- native hawaiian or other pacific islander
- Other
- not reported
- not allowed to collect

SDTM Valid Values:
- WHITE
- AMERICAN INDIAN OR ALASKA NATIVE
- ASIAN
- NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER
- BLACK OR AFRICAN AMERICAN

Findings – List Implied Data Elements

- Breast cancer cases did not specifically list the gender of the patient or the anatomic site

Approach to Harmonization Minimum Clinical Meta-Data

1. Patient ID
2. Age
3. Sex
4. Race
5. Estrogen Receptor Status
6. Progesterone Receptor Status
7. HER2/Neu Status
8. Laterality
9. Vital Status
10. Clinical Course of Disease
11. Anatomic Site
12. Primary Diagnosis

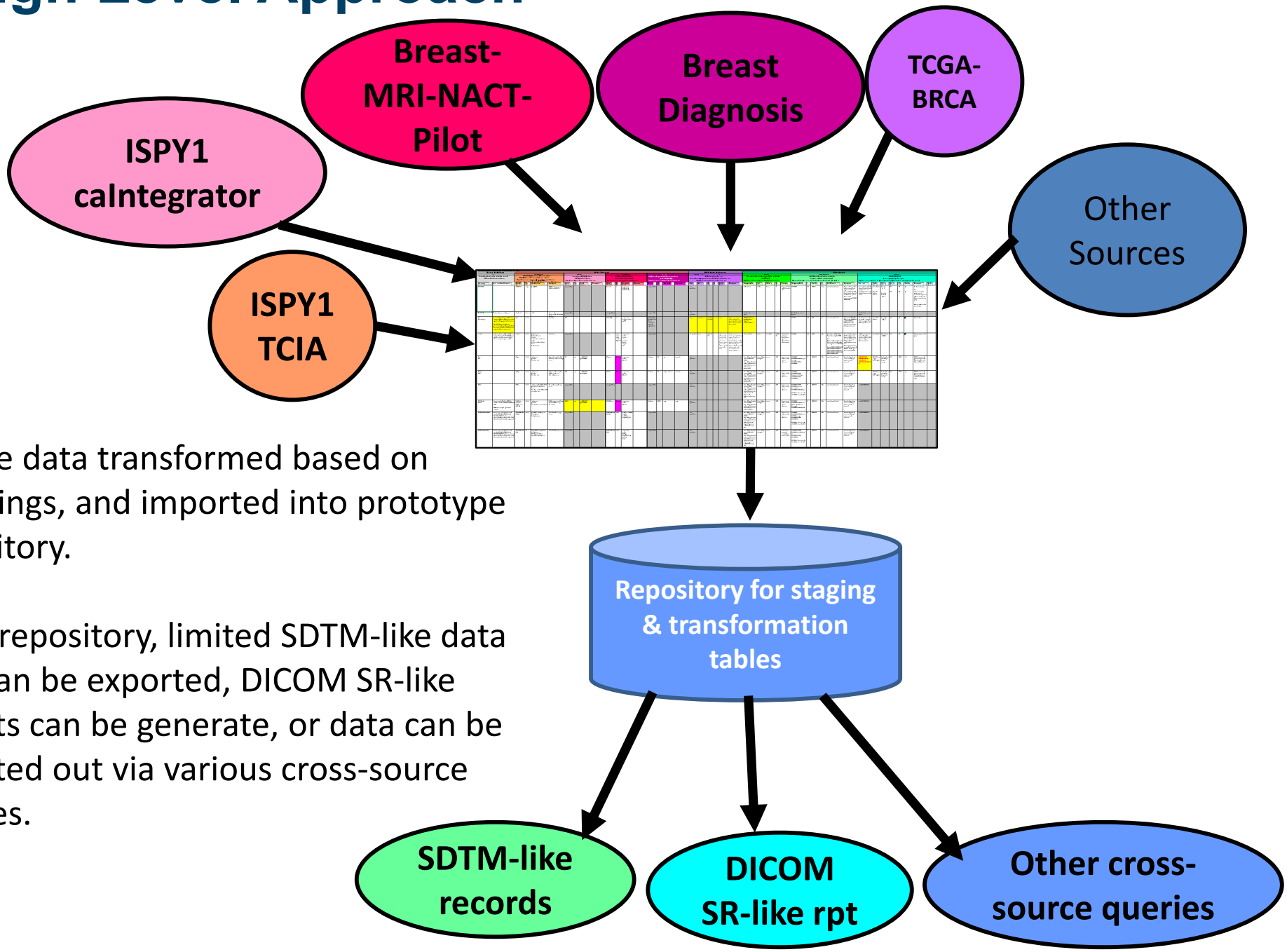
Used "rule of three" to tag a element as "minimum/core" and add to this list

Developed/identified the following for each element:

- Definition
- Data Type/Format
- Valid values (as applicable)

Bound every concept to a NCI EVS concept code

High Level Approach



Source data transformed based on mappings, and imported into prototype repository.

From repository, limited SDTM-like data sets can be exported, DICOM SR-like reports can be generate, or data can be reported out via various cross-source queries.

Demo

Conclusions

- Looking at the state of clinical data collections presently accessible to us shows that they are disparate

Recommendations

- Identification of a minimum standardized clinical data set will enable semantic interoperability
- Prospective data collection efforts based on standards are more efficient than retrospective transformation
- Building standards based solutions opens the door to integrating with other large-scale projects
- Consider leveraging BRIDG as the common information model for clinical trial and clinical care data

Documents available on NCI wiki

- Master mapping spreadsheet
- Mapping decisions
- https://wiki.nci.nih.gov/x/xllnFQ

The Team (sorted alphabetically)

Lauren Becnel (CDISC)
David Clunie (Pixelmed)
Julie Evans (Samvit)
Smita Hastak (Samvit)
Ed Helton (NCI)
Hubert Hickman (Essex Management)
Sam Isa (Essex Management)
Wendy VerHoef (Samvit)

The Bigger Picture

- Information on FDA's Common Data Model Harmonization Project (CDMH)
- Led by Mitra Rocca, FDA
- Leveraging BRIDG and CDISC SDTM

Goal and Objectives of CDMH

Goal:

Build a data infrastructure for conducting research using Real World Data* derived from the delivery of health care in routine clinical settings.

Objective:

Develop the method to harmonize the Common Data Models of various networks (Sentinel, i2b2/ACT, PCORNET, OMOP) allowing researchers to simply ask research questions on much larger amounts of Real World Data* than currently possible, leveraging open standards and controlled terminologies to advance Patient-Centered Outcomes Research.

* Examples of Real World Data are EHR data, claims data, data from registries, patient-generated data, and data from mobile devices.