DI-cubed
Breast Cancer Data Sets and one Brain Cancer Data Set
Data Mapping and Harmonization

NIH NATIONAL CANCER INSTITUTE

December 4, 2017

Decisions

- Encode concepts and values where possible
- Use NCI Thesaurus whenever possible (https://ncit.nci.nih.gov/)
  - If we don’t find a concept in NCI Thesaurus we look for a different standard / value source
- Indicate source and code
  - e.g. NCIt:C38361
    (NCI Thesaurus (NCIt) = source; C38361=code for Estrogen Receptor)

What is i2b2?

- Informatics for Integrating Biology and the Bedside
- https://www.i2b2.org/index.html
- The goal of i2b2 is to develop the science and the engineering required to enable the clinical investigators of academic medical centers to conduct clinical research that is informed by state-of-the-art genomics and biomedical informatics.
- Open source research data warehousing system
- Ontology driven interface

Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Anatomic Site

- Anatomic Site (NCIt:C13717)
  - Brain (NCIt:C12439)
  - Breast (NCIt:C12971)
  - Lung (NCIt:C12468)

- Source of permissible values - NCI Thesaurus

Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Clinical Course of Disease

- Clinical Course of Disease (NCIt:C35461)
  - No Evidence of Disease (NCIt:C40413)
  - Recurrent Disease (NCIt:C38155)

- Source of permissible values - NCI Thesaurus
- Mapping Decision / Remark:
  - Two data element (Recurrence free survival & Progression or Recurrence) were considered similar and related to each other, although the semantics used in the data sets were not clear.
  - Based on looking at the valid values of the two data elements where the focus appeared to be deciding whether the disease recurred, we decide to select just one element that is focused on the concept of recurrence.

Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Data Set

- Data Set (NCIt:C47824)
  - Breast Diagnosis
  - Breast-MRI-NACT-Pilot
  - I-Spy1
  - Ivy-Gap
  - TCGA-BRCA

  - Source of permissible values: defaulted based on original data sources

Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
  - Age
  - Race
  - Sex
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Demographics

- Demographics (NCIt:C16495)

- Source of permissible values - NCI Thesaurus
- Mapping Decision / Remark: Extra level was introduced to group Age, Sex, and Race

Age

- Subject Age (NCIt:C69260)

- Source of permissible values: original values from source data
- Mapping Decision / Remark: Decisions we made on how to represent the data
  - Keep specific age for the records where we have the data
  - Preserve the original unit of age (years, decades)
  - Merged all flavors of age e.g. age, age at diagnosis, and age at specific event (MRI)
  - We realized that normalizing age related data elements was beyond the scope of this project.

Race

- Race (NCIt:C17049)
  - American Indian or Alaska native (NCIt:C41259)
  - Asian (NCIt:C41260)
  - Black or African American (NCIt:C16352)
  - Native Hawaiian or other Pacific Islander (NCIt:C41219)
  - Unknown (NCIt:C17998)
  - White (NCIt:C41261)

- Source of permissible values: OMB Race
- Mapping Decision / Remark:
  - Not given = unknown
  - Other = unknown
  - Not reported = unknown

Sex

- Sex (NCIt:C28421)
  - Female (NCIt:C16576)
  - Male (NCIt:C20197)

- Source of permissible values: NCI Thesaurus
- Mapping Decision / Remark:
  - The sex was assumed to be female for all but one of the breast cancer data sets. Only the TCGA BRCA data set specified gender.
  - The team recognizes different ways of describing sex (societal vs. phenotypic). This project used the phenotypic definition.

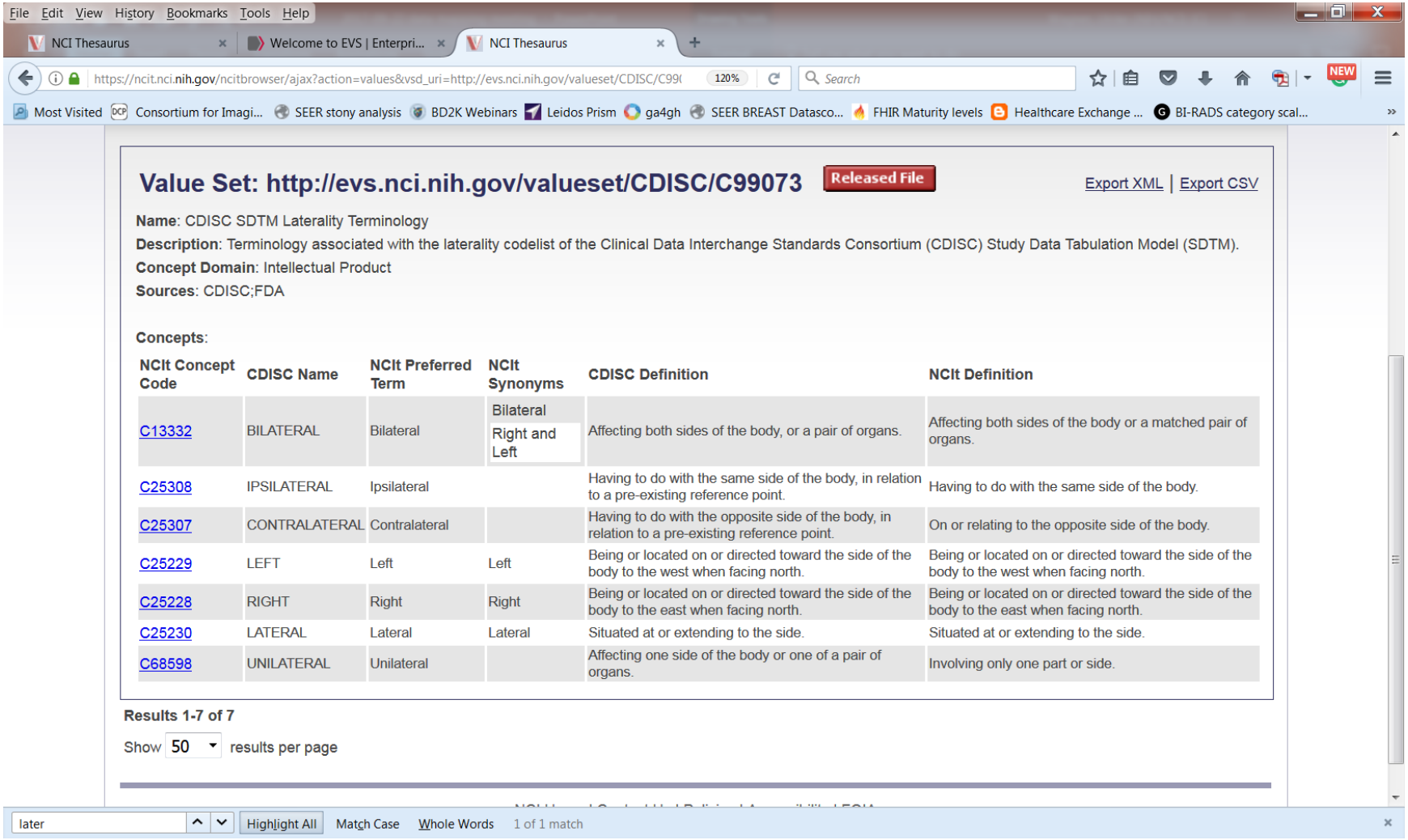
Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Laterality

- Laterality (NCIt:C25185)
  - Left (NCIt:C25229)
  - Right (NCIt:C25228)

- Source of permissible values: NCI Thesaurus / CDISC Study Data Tabulation Model (SDTM)
- Mapping Decision / Remark:
  We only include left and right from the available permissible value set (see next slide).



Value Set: http://evs.nci.nih.gov/valueset/CDISC/C99073

Released File

Export XML | Export CSV

Name: CDISC SDTM Laterality Terminology

Description: Terminology associated with the laterality codelist of the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM).

Concept Domain: Intellectual Product

Sources: CDISC;FDA

Concepts:

NCIt Concept Code	CDISC Name	NCIt Preferred Term	NCIt Synonyms	CDISC Definition	NCIt Definition
C13332	BILATERAL	Bilateral	Bilateral Right and Left	Affecting both sides of the body, or a pair of organs.	Affecting both sides of the body or a matched pair of organs.
C25308	IPSILATERAL	Ipsilateral		Having to do with the same side of the body, in relation to a pre-existing reference point.	Having to do with the same side of the body.
C25307	CONTRALATERAL	Contralateral		Having to do with the opposite side of the body, in relation to a pre-existing reference point.	On or relating to the opposite side of the body.
C25229	LEFT	Left	Left	Being or located on or directed toward the side of the body to the west when facing north.	Being or located on or directed toward the side of the body to the west when facing north.
C25228	RIGHT	Right	Right	Being or located on or directed toward the side of the body to the east when facing north.	Being or located on or directed toward the side of the body to the east when facing north.
C25230	LATERAL	Lateral	Lateral	Situated at or extending to the side.	Situated at or extending to the side.
C68598	UNILATERAL	Unilateral		Affecting one side of the body or one of a pair of organs.	Involving only one part or side.

Results 1-7 of 7

Show 50 results per page

Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Primary Diagnosis

- Diagnosis (NCIt:C15220)

- Source of permissible values: NCI Thesaurus
- Mapping Decision / Remark: mapped values from original data sets to NCI Thesaurus codes, see next slide including mapping spreadsheet

Primary Diagnosis – Valid Values

(Breast and Primary Brain tumors, non-lymphoma)

1. Anaplastic Astrocytoma (NCIt:C9477)
2. Astrocytoma (NCIt:C60781)
3. Benign (NCIt:C14172)
4. Breast Carcinoma (NCIt:C4872)
5. Breast Fibroadenoma (NCIt:C3744)
6. Breast Fibrocystic Change (NCIt:C3039)
7. Ductal Breast Carcinoma In Situ (NCIt:C2924)
8. Glioblastoma (NCIt:C3058)
9. Invasive Breast Carcinoma (NCIt:C9245)
10. Invasive Ductal Carcinoma, Not Otherwise Specified (NCIt:C4194)
11. Invasive Lobular Breast Carcinoma (NCIt:C7950)
12. Mixed Neoplasm (NCIt:C6930)
13. Stromal Hyperplasia (NCIt:C35857)
14. Unknown (NCIt:C17998)

Mapping of valid values

Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Receptor Status

- Receptor Status (NCIt:C94299)
  - Estrogen Receptor Status (NCIt:C16150)
  - HER2/Neu Status (NCIt:C16152)
  - Progesterone Receptor Status (NCIt:C16149)

Estrogen Receptor Status

- Estrogen Receptor Status (NCIt:C16150)
  - Estrogen Receptor Negative (NCIt:C15493)
  - Estrogen Receptor Positive (NCIt:C15492)
  - Estrogen Receptor Status Unknown (NCIt:C15495)

- Source of permissible value sets: NCI Thesaurus
- Mapping decisions / Remarks:
  - Indeterminate mapped to “unknown”
  - Weak positive / strong positive / positive mapped to “positive”
  - Blank mapped to “unknown”
  - We assumed that neg means negative, pos means positive
  - We assumed that the label of the field Erpos actually means “ER Status”

HER2/Neu Status

- HER2/Neu Status (NCIt:C16152)
  - HER2/Neu Negative (NCIt:C68749)
  - HER2/Neu Positive (NCIt:C68748)
  - HER2/Neu Status Unknown (NCIt:C68750)

- Source of permissible value sets: NCI Thesaurus
- Mapping decisions / Remarks:
  - Indeterminate mapped to “unknown”
  - Weak positive / strong positive / positive mapped to “positive”
  - Any flavor of negative mapped to “negative”
  - Blank mapped to “unknown”

Progesterone Receptor Status

- Progesterone Receptor Status (NCIt:C16149)
  - Progesterone Receptor Negative (NCIt:C15497)
  - Progesterone Receptor Positive (NCIt:C15496)
  - Progesterone Receptor Status Unknown (NCIt:C15498)

- Source of permissible values: NCI Thesaurus
- Mapping Decision
  - Indeterminate mapped to “unknown”
  - Weak positive / strong positive / positive mapped to “positive”
  - Any flavor of negative mapped to “negative”
  - Blank mapped to “unknown”

Data Organization

- Anatomic Site
- Clinical Course of Disease
- Data Set
- Demographics
- Laterality
- Primary Diagnosis
- Receptor Status
- Vital Status

Vital Status

- Vital Status (NCIt:C25717)
  - Alive (NCIt:C37987)
  - Dead (NCIt:C28554)
  - Lost to follow up (NCIt:C48227)
  - Unknown (NCIt:C17998)

- Source of permissible values: NCI Thesaurus