Title

Observations from the Data Integration and Imaging Informatics (DI-Cubed) Project

Authors and Affiliations

David Clunie, PixelMed (ORCID 0000-0002-2406-1145)
Hubert Hickman, Essex Management (ORCID 0000-0002-7526-3193)
Wendy Ver Hoef, Samvit Solutions (ORCID 0000-0002-6033-0912)
Smita Hastak, Samvit Solutions (ORCID 0000-0001-7606-0374)
Julie Evans, Samvit Solutions (Retired)
Jon Neville, CDISC
Ulrike Wagner, Frederick National Laboratory for Cancer Research (ORCID 0000-0002-3230-5058)

Abstract

In this paper we explore extending the concept of common cross-study Common Data Element concepts beyond simple demographics to cover disease-specific concepts relevant to imaging. We test interactively linking the resulting database to the associated images in a federated manner. We examine the use of existing standards, not only for terminology, but for interchange of serialized data in forms familiar to imaging and clinical trials specialists and their dedicated systems. Our intent is to perform preliminary work to inform both the upcoming Imaging Data Commons specifically, as well as more general integration projects beyond imaging.

Introduction

Image-derived observations are a vital component of modern oncology clinical practice, therapeutic clinical trials as well as basic research, particularly those observations that are quantitative, reproducible and validated biomarkers, whether generated by traditional means or with the application of artificial intelligence [Savadjiev] [Alberich-Bayarri]. It is recognized that to be maximally useful, images and image-derived information need to be linked to clinical, treatment and response data as well as genetic and other information [Chennubhotla]. Heterogeneity and incongruity remain challenging [Toga] [Neu]. It is important to collect such real-world data over multiple time points [Lee] [Hu].

Secondary re-use of images and clinical data collected for other purposes (such as clinical trials) and unimpeded sharing of that data is now well accepted [Bertagnolli]. Proof of this is evident in the success of The Cancer Image Archive (TCIA) [Clark], which at the time of writing contains 122 collections, the vast majority of which (113) are publicly accessible and the remainder accessible on request [TCIA Collections].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

Common understanding and interpretation of data elements and values will be critical to successful implementation of the "findable" component of the FAIR principles for data reuse [Wilkinson]. Inconsistency of nomenclature remains a major barrier to re-use across different sources and linkage between domains [Green] [Tenenbaum] and for distributed machine learning [Deist]. Development of Common Data Elements (CDEs) as a partial solution has long been a National Cancer Institute (NCI) priority [Meadows] [NCI 2001a] [NCI 2001b] [Hubbard] [Winget]. Experience with the NCI CDEs has not been universally positive [Nadkarni], but they have proven useful in the NCI Genomic Data Commons (GDC) effort [Vesteghem]. Past experience has described the difficulties in mapping when data is not initially standardized, and highlighted the need for curation in addition to lexical matching [Pathak]. The importance of using common terms and enforcing uniform semantic interpretation of data elements is recognized in specialties other than oncology [Zhang]. The multitude of incomplete, inconsistent and overlapping standards for terminology and data elements [Sansone] make it hard to choose between them. The same harmonization issues arise whether data of a similar type is centralized, as in TCIA, or distributed and federated [Alberich-Bayarri] [Smedley], and are exacerbated when linking data from different domains [Denny].

The GDC [Grossman] project identified the data harmonization challenge posed by disparate data dictionaries and clinical study designs, the tension between reusing legacy collections and prospective good practice, and the need to map clinical data into GDC standards [Jensen]. The importance of reusing the GDC experience and components for other oncology applications has also been recognized [Vesteghem]. The recently defined GDC Common Cross-Study CDEs, though few in number, are particularly relevant [DMWG].

An ultimate goal is to achieve a global knowledge commons (resource shared by a group of people [Hess]) that spans all data sources of all types. Within NCI's scope, a national Cancer Research Data Commons is potentially achievable [Dearry].

The linkage of images to staging and treatment outcomes has been previously explored from a scientific and operational perspective and the role of interoperability standards considered, e.g., in radiotherapy [Fedorov] [Elhalawani] [Grossberg].

There are numerous types of relevant interoperability standards, including:

- abstract information models, such as the Biomedical Research Integrated Domain Group (BRIDG) model [Fridsma],
- regulatory clinical trial data submission models, such as the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) [CDISC SDTM],
- information communication standards, such as Digital Imaging and Communications in Medicine (DICOM) [DICOM] for images and associated information, and
- lexicons, vocabularies, controlled terminologies, classifications and ontologies, such as NCI Thesaurus [Golbeck] [Sioutos] and SNOMED CT [Donnelly].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

Though the importance of linkage of imaging phenotypical data and clinical data as well as other specialized data source is recognized, and individual systems have been constructed for such purposes, the value of using existing interoperability standards has been less well explored. We hypothesized that:

- existing standard data models and concepts for CDEs and values could be used to harmonize clinical data and results across image collections from disparate sources,
- the resulting harmonized data could be queried and analyzed seamlessly regardless of source.
- the harmonized data could be easily linked to the related images stored in a different system, and
- the harmonized data could be exported to existing standard formats to support interoperability between systems.

It is hoped that the results of this project can be used to inform the development of NCI Research Data Commons in general, and in particular the Imaging Data Commons (IDC) [FNL IDC] [Jett IDC] [Dearry IDC].

Materials and Methods

The project involved the following conceptual steps:

1. Identification of TCIA collections to include in the project.
2. Identification of collection-specific data elements used in those collections.
3. Database schema definition of collection-specific data elements.
4. Extraction of collection-specific data elements and text and coded values from spreadsheets into database.
5. Identification of related standard data elements and coded values from GDC Common Cross-Study CDEs, BRIDG, CDISC SDTM and DICOM.
6. Database schema definition of standard data elements.
7. Definition of mapping of collection-specific data elements and values into standard data elements and coded values.
8. Translation of collection-specific data elements and values into standard data elements and coded values within database.
9. Implementation of a query interface.
10. Linking of imaging study records in database to TCIA imaging study identifiers.
11. Implementation of a database user interface extension to trigger display of selected studies in TCIA search window.
12. Report generation to export records in CDISC SDTM model in CSV format.
13. Report generation to export records in DICOM Breast Imaging Report in Structured Report (SR) format.
14. Loading of exported DICOM SR objects into TCIA collections.

Iteration over several of these steps was necessary to converge on a solution that achieved sufficient coverage and consistency of user experience within the interactive database.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

TCIA Collection Selection and Data Access

The collections selected needed to have clinical data associated with the images, be focused on a single disease and method of imaging, and have both their images and clinical data publicly available. Breast cancer MRI collections met these criteria. Data elements from the four following collections (five sources) were included:

- ISPY1 [Newitt multi] (TCIA Clinical and Outcome Data, caIntegrator sources)
- Breast-MRI-NACT-Pilot [Newitt single]
- TCGA-BRCA [Lingle]
- Breast Diagnosis [Bloch]

The GDC Clinical Data Elements were also used as an additional source of elements, though not data [DMWG].

The QIN-BREAST [Li] collection could not be used, since its data access was restricted and remains so, even though it contains outcomes data and individual access was requested and granted.

To perform a cross-check on the utility of the cross-study data elements selected, one collection with clinical data from an additional disease state was selected, MRI of glioblastoma:

- IvyGAP [Shah]

The images and their patient, study, series and image related metadata were accessed using the TCIA RESTful API [NBIA]. A subset of the metadata in the DICOM images stored in TCIA is extracted during the collection ingestion process, stored in the TCIA database and made available through the query API. Additionally, DICOM image metadata that is not so indexed can be obtained by downloading the images and extraction from the DICOM “header” (non-pixel data attributes) [Källman].

Every TCIA collection is described in a manually created Wiki page, which contains information about the collection as well as reference to data acquisition protocol description documents and ancillary files containing clinical data and analysis results. For each TCIA collection extracted, the data dictionary and clinical data spreadsheet files were downloaded. For extraction, data in Microsoft Excel (XLS) spreadsheets were first exported into non-proprietary comma separated value (CSV) files amenable to text processing tools.

Several collections (ISPY1 and Breast-MRI-NACT-Pilot) described result data and related metadata buried within the DICOM image headers. However, since the same data was already included in the ancillary spreadsheets, and tumor location information was not required for this project, it was not necessary to use a DICOM attribute extraction tool to extract this information, though tests were performed to demonstrate the feasibility of such an approach using the dicom3tools utilities [dicom3tools].

 doi:10.20944/preprints202008.0474.v1

Database Selection and Implementation

The i2b2 database was selected because it is patient (human subject) centric, was designed for extraction of medical records and clinical data, makes use of a “fact table” of observations consisting of name-value pairs, supports a web client that facilitates query creation by non-expert users, and is open source, mature and designed to be extensible [Murphy]. Its capabilities were expected to be sufficient for the proof of concept and potentially scalable to an operational deployment.

The source data for each collection was obtained from the TCIA collection Wiki pages. This data was then imported into a staging area in i2b2 in a minimally processed representation. The data was then transformed and mapped, such that it conformed to the structure of the i2b2 fact table Each transformed prototype fact table was then loaded into the main i2b2 fact table, which enabled its use by the i2b2 query user interface.

Once the target data elements had been identified and a mapping defined (as described below), an i2b2 implementation of the mapping rules was written to create additional entries in the fact table for the harmonized target data elements. Both coded and preferred name entries were created to encode each concept value.

The resulting user experience in the default query interface was then reviewed by technical and subject matter experts to confirm the consistency and usability of the mapped data for various plausible use cases for imaging study cohort selection. The organization of the fact table entries was iteratively refined until the users were satisfied.

Data Element Identification and Mapping

The data element identification and mapping process overlapped with the loading of the collection-specific data elements and values into i2b2. For each collection included in the project, the columns of the ancillary clinical data spreadsheet files were matched to any accompanying documentation (such as data dictionary files). The names of the columns were entered into a mapping spreadsheet as well as an i2b2 schema. The values used in each collection were identified from data dictionaries as well as actual values encountered in the records once loaded into i2b2.

The mapping spreadsheet used BRIDG, CDISC SDTM, and DICOM standards to map the CDEs. It also included demographic and CDEs used in the Genomic Data Commons (GDC) Common Cross-Study CDEs [DMWG]. Metadata from the sources were mapped to each other and also to the standards. Mapping was done at a granular level that leveraged the element name, definition (when available), data type/format, and the valid value list, as applicable.

All collection-specific source data elements related to subject demographics, diagnosis and outcome were considered as candidates for mapping. Those that were present with values in three or more collections were chosen for harmonization. The text or coded values specific to each collection or standard were extracted from the data dictionaries and using an initial load

 doi:10.20944/preprints202008.0474.v1

into i2b2, compared with the actual values encountered in the data sets. Typographic errors and inconsistencies in text values were resolved.

Standard coded concepts and values were then selected for each candidate source data element and value. Priority was given to the NCI Thesaurus as a source of coded concepts for the name-value pairs, using the GDC Common Cross-Study CDE [DMWG] choices when appropriate. The NCI Thesaurus Code of each concept was encoded in the mapping and database for each data element and value together with the NCI Thesaurus Preferred Name. The caDSR Public IDs were not used.

Additional mapping of coded concepts and values to the various database export formats was also performed. The DICOM standard uses either SNOMED CT, LOINC or DICOM's own codes in the SR templates and context groups (value sets) and so a mapping of NCI Thesaurus codes to SNOMED CT and other codes was created. The NCI Thesaurus concepts for data elements were also mapped to CDISC SDTM table columns and the concepts for values were mapped to CDISC SDTM string codes if found, otherwise the NCI Thesaurus Preferred Name was used as the value.

The mapping of data elements and values was iteratively refined using offline spreadsheets and summary documents until both technical and subject matter experts were satisfied with its scope of coverage and consistency and reviewed again once implemented in i2b2.

TCIA Linkage

The primary key used to establish correspondence of the de-identified imaging studies in the TCIA to the mapped clinical data in i2b2 was the Subject ID (DICOM Patient ID). Where more than one imaging study was present in TCIA for a single subject, information related to the time point (DICOM Clinical Trial Time Point ID) and date on which the study was performed (DICOM Study Date) was used to select the appropriate study. Though the study dates in TCIA have been de-identified, their relative temporal offset is maintained. For those collections for which an explicit time point identifier or imaging study date was not present in the clinical source data, the relative position of a study (first MRI, second MRI, etc.) was used to establish correspondence.

The DICOM Study Instance UID was used as the actual key for linkage of i2b2 records to TCIA imaging studies once the conceptual mapping was established.

The TCIA RESTful API was used both to download imaging study related metadata and to create an interactive link between the i2b2 user interface and the TCIA Search web page. Two new tables were added to the i2b2 database, one for the imaging study, and one for the series within the study; these contain relevant metadata for mapping and linking. A Python script was written to retrieve the data to populate these tables via the RESTful API. The DICOM images did not need to be downloaded to perform the mapping since the relevant metadata was indexed in the TCIA database and was available via the API. Nor were the images stored locally in the i2b2 system. An i2b2 extension was written to allow the user to select a cohort of subjects, studies or individual studies and trigger display of the TCIA page for that subset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

Data Export

Several data formats were selected as targets for export of the harmomized TCIA-linked clinical data. Export as a DICOM Structured Report (SR) was chosen in order to demonstrate that the clinical data could be archived and distributed in a standardized structured coded format intended for clinical use, and encoded in the same manner as the images so that they could be made available in image databases and to imaging workstations. Export in CDISC SDTM was performed to show that the same data could be distributed in the standard used for clinical trial research for regulatory agency submissions.

The GUI screens and output files for the export processes were created using the R language and R's Shiny web interface toolkit, and could be launched via i2b2's plugin architecture or run as a standalone web application.

DICOM Breast Imaging Report SR Export

The DICOM Structured Report (SR) was designed to be a "self-describing information structure" that could be "tailored to diverse clinical observation reporting applications by utilization of templates and context-dependent terminology" [Bidgood]. A DICOM SR document consists of an ordinary DICOM "header" containing demographic and identification information, accompanied by a "content tree" that consists of a recursive structure of name-value pairs. Extensive use is made of codes rather than plain text, numeric measurements, and references to images and coordinates of regions on those images [Clunie SRClinTrials] [Clunie SR]. DICOM defines specific templates to be used to constrain the flexibility of the SR content tree and standardized the codes and structures used [Noumeir]. DICOM SR instances are intended to be displayed by imaging workstations together with their associated images [Hussein].

To create DICOM SRs, a report was generated in i2b2 to produce an intermediate CSV file containing a single table of the information to be included in a DICOM Breast Imaging Report SR, with one table record per row containing sufficient information for higher level entities replicated in that row. In addition to the CDEs common across the source data collections, when image-derived information was present, such as tumor size measurements, these were also exported to be included in the DICOM SR.

One DICOM SR file was created per imaging study by processing the i2b2 exported CSV file with a set of XSLT stylesheets [W3C XSLT 2.0] to create an intermediate XML representation that was then converted into the standard binary DICOM format using an open source DICOM toolkit [PixelMed]. The DICOM SR template used was TID 4200 Breast Imaging Report [TID4200], extended with additional concept as necessary. Concept and value code mapping from NCI Thesaurus to SNOMED CT was performed in one of the XSLT stylesheets. Each DICOM SR file was added to the same imaging study as the images to which it is related by reusing the Study Instance UID of the images. In a small number of cases, not all the clinical data had matching imaging studies, and these SR files were excluded.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

The DICOM SR files were then validated for conformance with basic DICOM encoding rules using dicom3tools dciodvfy [dciodvfy] and for conformance with TID 4200 using DicomSRValidator [DicomSRValidator]. The i2b2 imported data, data export report and XSLT stylesheets were iteratively revised to correct mapping errors until a satisfactory result was achieved.

The DICOM SR files were also imported into a publicly available off the shelf open source image viewer [Horos], which can display the content of the report alongside the images, in human readable form, by using the dcmtk toolkit utility dsr2html [dsr2html].

The DICOM SR files were then submitted to TCIA for inclusion in the original collections for public release.

CDISC SDTM Export

A report was generated in i2b2 to produce CDISC SDTM CSV and SAS Transport files. The alternative XML representation of CDISC SDTM was not produced. Tables for the Demographic (DM), Disposition (DS), Microscopic Findings (MI), Procedures (PR), Subject Status (SS), Tumor Identification (TU) and Tumor Results (TR) domains were created [CDISC SDTM v1.3] [CDISC SDTM IG HT], and each could be displayed as a separate tab in the report generator user interface. Some required information for the various domains could not be populated. For example, the Planned Arm Code (ARMCD) variable in the Demographics (DM) domain was not populated due to lack of data.

The SDTM datasets were then validated using the Pinnacle 21 CDISC validation software [Pinnacle 21], which reviews datasets according to their degree of conformance to rules developed for the purposes of FDA submissions of electronic data. The resulting reports were annotated with explanations of the flagged issues and notes regarding findings from the exploration of their sources within the mapped data. Corrected datasets were revalidated in an iterative fashion until group consensus was achieved. In some cases fixed data values were created to minimize errors (e.g., SITEID in the DM domain).

Results

Mapping

Since the cross-mappings across the sources and standards were done in a single master mapping spreadsheet, the results of the mapping were quantifiable. Additionally, color coding of cells and columns allowed for a quick visual assessment of which data element concepts were common across the various data sources and highlighted the areas where equivalent mappings did not exist. The approach is illustrated in Figure R-MP-1. Supplementary Data Set 1 contains the final spreadsheet.

Without regard to their importance, very little overlap for the patient clinical data associated with these collections was found. A total of 149 unique clinical data elements were analyzed. An

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

additional 8 data elements were added subsequently by direct query to TCIA to supplement the original 149. Aside from Subject ID, no data elements were common to all collections. Below is a frequency breakdown for the 157 clinical data elements:

- 4 data elements occur in all 5 sources
- 5 data elements occur in 4 sources
- 4 data elements occur in 3 sources
- 8 data elements occur in 2 sources
- 136 occur in only 1 source

On the basis of these findings, it was decided to proceed with only data elements that were present in at least three of the sources.

For the selected data elements, we found that the existing standard NCI Thesaurus concepts for CDEs and values could be used to harmonize the clinical data and results across image collections from the collections selected. The list of data elements and values used are shown in Table R-1.

The majority of these presented little challenge in establishing a mapping between the source concept and values and the standard target. This was despite considerable variation in the naming of the data elements and values in the source data dictionaries and data sets. Formal definitions of elements and values were often lacking. It was necessary to review the actual data to gain insight into what was being collected and educated guesses were required in some cases. For many data elements the meaning is context sensitive. This is particularly the case with observations done at a particular time point relative to a significant milestone, such as tumor size measurements relative to a treatment event.

Age and date concepts were challenging and of limited use, given the wide variation found in epoch relative to which age was specified (e.g., age at diagnosis versus other events), and inconsistent precision with which nominally de-identified dates were provided (e.g., year of birth or death only).

For some source data elements, we found that more than one concept was conflated in a single data element, e.g., survival and follow up status. The opposite was also true in the case of race and ethnicity, which was challenging to map to DICOM because ethnicity is a US-centric (Hispanic or not) rather than international concept.

Values for data elements also provided mapping challenges. In general, the various data sources contained numerous coded values, however all were human-readable short text or local data element specific numeric codes rather than any standard code from an external lexicon. The degree of specificity varied, especially in the case of histopathological diagnosis. Harmonizing values for data elements from sources that allowed for choices of “multiple”, “other” or different flavors of null was awkward. Race, for example, was particularly awkward in this respect.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

A number of data elements that were not part of the source data sets were actually available either via other systems, APIs or via derivation from the protocol or other metadata. For instance, gender and disease diagnosis were often missing but derivable from the study protocol as prerequisites for participation in the study. For example, all of the breast cancer patients were known to be female. In hindsight, these data elements could easily have been included in the source data sets to make meaningful comparisons across studies and to make cross-study searches easier to do.

There are more complicated examples of logic that were required to properly map some of the data items. The tumor measurements that were available in two of the source collections required uniting data only available from the TCIA API (image study dates) together with data from the data spreadsheets. These two collections had measurement data with a possibility of four measurement time points. Some study participants did not have imaging studies performed for all of the four possible time points, and hence our processing logic had to accurately account for those gaps.

Database

The i2b2 database proved useful not only for delivering the harmonized data sets, but also for performing the work of loading and translation (mapping) of the various source data sets as well. It was useful for making lists of values encountered for source data elements, to use as input to the mapping decision making process.

When the mapping had been defined, design decisions were required for the use of coded terminology rather than free text values for standard data elements. Standard coded concepts such as those in NCI Thesaurus have codes as well as human readable preferred names. The i2b2 support for dictionaries of coded concepts was used to assure that both the code and the preferred name were available for query and report generation.

Once harmonized, the data in i2b2 could be queried and analyzed seamlessly regardless of source. Figure R-DB-1 illustrates a simple query to identify a patient set that spans several source data sets and selects patients with a specific disease and specific receptor status.

The TCIA REST API was found to be sufficient to obtain additional imaging study specific metadata, which could be obtained via the API without downloading any of the images. The harmonized data in i2b2 was then easily linked to the related images stored in TCIA by using the TCIA REST API to trigger the display of the appropriate TCIA search pages, as shown in Figure R-DB-2.

Export

The harmonized data was successfully exported to existing standard formats to support interoperability between systems. Automated validation tools were used as a surrogate for actual transport of the data into a different system to demonstrate interoperability, since we did

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

not have access to such a system. An alternative would have been to perform a round-trip importation of the exported data, but the resources were not available to pursue this option.

DICOM SR Export

We found that there was a relatively consistent mapping between the clinical data and measurements identified for this project and recorded in i2b2, and the TID 4200 supplementary data, as well as the patient characteristics described in the typical DICOM “header” such as age and sex. The initial mapping that was established when DICOM TID 4200 was considered as a source of input for this project and as a source of common data elements could be reversed to support the export of the i2b2 data linked with the TCIA image information to DICOM SR TID 4200.

Overall, the exercise of exporting the i2b2 content demonstrated that the information gathered from the collection sources, linked to TCIA and stored in i2b2, was largely sufficient to generate near-valid standard clinical breast imaging reports with only minor deviations. Relevant clinical content can be extracted and stored in a conventional DICOM archive and displayed in a conventional DICOM viewer together with the images. The conversion process was relatively straightforward to implement, given familiarity with existing off-the-shelf tools, XML and XSLT. The XSLT consisted mostly of code mapping rules, and could easily be replicated by other means, e.g., in a procedural rather than declarative language, and could perhaps be largely automatically generated in future given a tabular input of code mappings for concepts and values. The DICOM Breast Imaging Report template would benefit from some updates and extensions to better support reporting of MR imaging studies as well as some more recently developed clinical content. The normal DICOM standard maintenance process can be used to implement such improvements.

The limitations of TID 4200 observed, and rectified by a change proposal submitted to the DICOM organization [CP 1838], which has now been approved, included:

- the addition of patient characteristics (clinical background) to include clinical course and extension of subject context to include racial group,
- the addition of HER2/Neu receptor status,
- the addition of laterality at finding level, since the procedure reported may be bilateral even though the finding is unilateral,
- the need to relax the requirement to include the pathology sampling date time, since it may not be available even though the related data is still useful, and
- the need to expand the list of breast procedures reported to include procedures not covered, more specific procedures, and LOINC codes for various procedures, especially breast MRI.

CDISC Export

Group discussion determined which of the errors reported by the CDISC validation software could and should be addressed, and how to do so. Many of the underlying rules were deemed not relevant to this project. Therefore, some of the issues were ignored. In other words, though

the CDISC SDTM exports were in a sense incomplete, the SDTM standard itself was found to be sufficient to encode the data generated.

Discussion

The results clearly indicate that although patient clinical data is being gathered for the collections considered, there is a lack of consistency on what is collected or at least what is shared. For NCI to meet the informatics goals defined by the Cancer Ecosystem and Cancer Moonshot effort [NCI Ecosystem], developing a core or minimum set of patient clinical data elements that will be collected and shared is critical. Our findings are consistent with those from a similar study using lung and brain collections from TCIA, which recapitulates the need to reduce the burden of retrospective data harmonization through use of standards [Basu].

Our task would have been considerably easier if the projects that created the source data sets had reused existing CDEs and coded values rather than inventing their own. If new CDEs are actually required, they would be far easier to map if element names reflected greater semantic precision, even if it increases verbosity, and better yet, if they were accompanied by a formal, clear and unambiguous definition that followed some formal guidelines. Establishment of a well-defined set of common events and milestones common across similar types of research studies would also be useful, as would decisions about whether to represent temporal information as instants or intervals [Zhou] [Allen] and with what granularity [Goralwalla]. Adoption of a common standard of methodology for de-identification of ages and dates would greatly improve cross-study harmonization of temporally-related concepts. The use of standard concepts from well accepted sources, not only for CDEs, but also for their values is also crucial. A standard, minimum set of core data elements for comparison sake could be developed and required, even if all values from a given research study have a fixed value. Though universal consistency of terminology may not be achievable [Wiederhold], the current state could certainly be improved.

The NCI Thesaurus proved to be sufficient for the vocabulary needs of this project, and all selected data elements and their values were able to be mapped. This observation may not have general applicability, since only two diseases (breast cancer and glioblastoma) were tested, but is promising. Ultimately a more rigorous ontological approach that supports reasoning will be required [Haendel] [Mate], but for the time being, incremental progress towards that goal can be achieved using the NCI Thesaurus. Ultimately, it is expected that the National Cancer Institute's (NCI's) new Center for Cancer Data Harmonization (CCDH) component of the Cancer Research Data Commons (CRDC) will deliver suitable models [CCDH].

Mapping NCI Thesaurus concepts to the SNOMED CT codes used in the DICOM Breast Imaging Report was also largely successful, with the exception of one of the hormone receptor types (HER2/Neu), some more arcane histopathological diagnoses, and a protocol-specific quantitative imaging technique (signal enhancement ratio). Though use of SNOMED CT is eschewed in some jurisdictions due to lack of a national free license for use [Cangioli], or a pan-European license [ASSESS CT], the existence of a DICOM-SNOMED agreement assures

royalty and license fee free use globally for any SNOMED CT concept codes defined for the subset of codes used in the DICOM Standard.

The mapping exercise confirmed the utility of the GDC Common Cross-Study CDEs [DMWG]. Indeed, our work produced only a very modest increment over the GDC set, most of which were disease specific, and for some concepts, different choices. Though GDC lists the CDEs, it appears at first sight to only provide text values for categorical concepts. However, since the CDEs are defined by a Cancer Data Standards Registry and Repository (caDSR) [CBIIT Metadata] Public ID, the Value Domain defined for each CDE can be accessed via the NCI CDE Browser [NCI CDE Browser], and those Value Domains do indeed contain appropriate NCI Thesaurus concepts. We observed that similar choices of coded concepts and values for non-disease-specific data elements were made in a vascular disease mapping project [Pathak].

The DICOM Breast Imaging Report was initially designed with X-Ray mammography in mind, and extended for ultrasound, but has not been specifically updated for MRI results. Yet it proved largely sufficient to encode the data gathered in this project. Some minor improvements to the template that were identified, proposed and have since been accepted into the standard. DICOM SR templates have been used before for breast reporting in a clinical (non-research) setting, albeit with a non-standard locally developed template, but the value of such structured data for secondary re-use is acknowledged [Segrelles].

The CDISC SDTM was sufficient to encode the range of demographic, diagnostic procedural, histopathological, radiological result and outcome data encountered, with the caveat that only selected elements of various domains could be encoded. SDTM categorical elements are not well suited to encoding concepts defined by external lexicons, having no standard means to encode both a code value and the source (coding scheme) that defines that value. SDTM text value sets were not sufficient to encode some of our selected data elements, so the convention of using the NCI Thesaurus Preferred Name was used, in preference to encoding a tuple formed from the scheme and code (e.g., “Estrogen Receptor Negative” rather than “NCIt:C15493”). Some difficulty was observed in trying to encode linkage information to the TCIA imaging studies, and the decision was made to encode the TCIA URL in a SDTM XFN (external file name) variable.

The choice of i2b2 proved to be satisfactory, and all requirements for the selected database were satisfied. In particular, querying for a cohort of studies or subjects based on the mapped clinical data and imaging metadata was straightforward. Linkage to TCIA was also simple, given the extensibility of i2b2 and access to a usable and well-documented, even if non-standard, TCIA RESTful API. It would be preferable if TCIA implemented a standard API for image access, both for downloading selected studies, such as DICOMweb [DICOMweb] [Genereaux], or for invoking their display such as the Integrating the Healthcare Enterprise (IHE) Invoke Image Display (IID) profile [IHE IID]. Scalability issues were not explored, nor cross-application security considerations, such as user authentication, access control or single-sign on. Ours is not the first use of i2b2 in association with medical images; I2b2 has been used to link researchers to image systems in local hospital environments [Murphy] [Gollub] [mi2b2]

[MGH/HST] though as far as we know, not to external repositories of already deidentified research image data. Other data warehouse projects with model harmonization have also used i2b2, and in future it may be desirable to consider harmonization of our data with the OMOP and PCORNet data models [Klann].

In some jurisdictions, emphasis is placed on the use of only ISO standards. To this end, we note that both DICOM [ISO 12052] and BRIDG [ISO 14199] are ISO standards. The caDSR is based on an ISO standard model for metadata registration [ISO 11179] [Ngouongo].

Like other studies [Basu], our metadata mapping approach was manual, performed by humans. Future automation of such mapping shows promise, e.g., as evident from the results of the NCI CRDC's Metadata Automation DREAM Challenge [MAD] [MADResults]. The need for a standard mapping target remains though, in the absence of proactive implementation of standard CDEs and values.

In conclusion, we have explored extending the concept of common cross-study concepts beyond simple demographics to cover disease-specific concepts relevant to imaging and interactively linking the resulting database to the associated images in a federated manner. We have emphasized the use of standards, not only for terminology, but for interchange of serialized data in forms familiar to imaging and clinical trials specialists and their dedicated systems.

We expect that this preliminary work will serve to inform both the upcoming Imaging Data Commons specifically, as well as more general integration projects beyond imaging, such as those in radiation oncology [Mayo].

Dedication

We dedicate this paper to the memory of the late Dr. Edward Helton, who at the time of his death was government sponsor for the NCI CBIIT's Clinical Imaging Informatics Program and instigator of the Data Integration and Imaging Informatics (DI-Cubed) project.

Acknowledgements

We wish to thank Carolyn Klinger for maintenance and management of the supporting wiki and web sites.

Support

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

External links

- Data Integration and Imaging Informatics (DI-Cubed) Files to Share page: http://wiki.nci.nih.gov/display/DIcubed/DI-Cubed+Files+to+Share
- Mapping decisions: http://wiki.nci.nih.gov/download/attachments/359107012/DI-cubed%20data%20mapping%20decisions.pdf
- Mapping primary diagnosis: http://wiki.nci.nih.gov/download/attachments/359107012/Primary%20Diagnosis%20Mapping.pdf
- Mapping spreadsheet: http://wiki.nci.nih.gov/download/attachments/359107012/DI-Cubed%20Master%20Metadata%20Mapping%20_%20comm%20call.xlsx
- Videos: http://www.youtube.com/playlist?list=PL2uforSa-XbPSLzaK4GVq-SMSGQ6ZdPOj
- Source code: http://github.com/CBIIT/DiCubed
- Clinical and study data in CDISC SDTM form - on TCIA collections Wiki - Hickman H, Ver Hoef W, Hastak S, Neville J, Clunie D, Wagner U, et al. SDTM datasets of clinical data and measurements for selected cancer collections to TCIA [Dataset] - TCIA DOIs - Cancer Imaging Archive Wiki. The Cancer Imaging Archive (TCIA). doi: 10.7937/TCIA.2019.zfv154m9 Available from: http://wiki.cancerimagingarchive.net/display/DOI/SDTM+datasets+of+clinical+data+and+measurements+for+selected+cancer+collections+to+TCIA
- Clinical and study data in DICOM form - on TCIA collections Wiki - Clunie D, Hickman H, Ver Hoef W, Hastak S, Wagner U, Helton E. DICOM SR of clinical data and measurement for breast cancer collections to TCIA [Data set] - TCIA DOIs - Cancer Imaging Archive Wiki. The Cancer Imaging Archive. 2019; doi: 10.7937/TCIA.2019.wgllssg1 Available from: http://wiki.cancerimagingarchive.net/display/DOI/DICOM+SR+of+clinical+data+and+measurement+for+breast+cancer+collections+to+TCIA

References

[Savadjiev] Savadjiev P, Chong J, Dohan A, Agnus V, Forghani R, Reinhold C, et al. Image-based biomarkers for solid tumor quantification. Eur Radiol. 2019; Available from: http://doi.org/10.1007/s00330-019-06169-w

[Alberich-Bayarri] Alberich-Bayarri Á, Hernández-Navarro R, Ruiz-Martínez E, García-Castro F, García-Juan D, Martí-Bonmatí L. Development of imaging biomarkers and generation of big data. Radiol med. 2017 Feb 21;1–5. Available from: http://doi.org/10.1007/s11547-017-0742-x

[Chennubhotla] Chennubhotla C, Clarke LP, Fedorov A, Foran D, Harris G, Helton E, et al. An Assessment of Imaging Informatics for Precision Medicine in Cancer. Yearb Med Inform. 2017 Aug;26(01):110–9. Available from: http://doi.org/10.15265/IY-2017-041

[Toga] Toga AW, Dinov ID. Sharing big biomedical data. Journal of Big Data. 2015 Jun 27;2:7. Available from: http://doi.org/10.1186/s40537-015-0016-1

[Neu] Neu S, Crawford K, Toga AW. Practical management of heterogeneous neuroimaging metadata by global neuroimaging data repositories. Front Neuroinform. 2012;6. Available from: https://www.frontiersin.org/articles/10.3389/fninf.2012.00008/full

[Lee] Lee JSH, Darcy KM, Hu H, Casablanca Y, Conrads TP, Dalgard CL, et al. From Discovery to Practice and Survivorship: Building a National Real-World Data Learning Healthcare Framework for Military and Veteran Cancer Patients. Clinical Pharmacology & Therapeutics. 2019;106(1):52–7. Available from: http://doi.org/10.1002/cpt.1425

[Hu] Hu H, Correll M, Kvecher L, Osmond M, Clark J, Bekhash A, et al. DW4TR: A Data Warehouse for Translational Research. Journal of Biomedical Informatics. 2011 Dec 1;44(6):1004–19. Available from: http://doi.org/10.1016/j.jbi.2011.08.003

[Bertagnolli] Bertagnolli MM, Sartor O, Chabner BA, Rothenberg ML, Khozin S, Hugh-Jones C, et al. Advantages of a Truly Open-Access Data-Sharing Model. 2017 Mar 22;376:1178–81. http://dx.doi.org/101056/NEJMsb1702054

[Clark] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013 Jul 25;26(6):1045–57. Available from: http://doi.org/10.1007/s10278-013-9622-7

[TCIA Collections] The Cancer Imaging Archive. TCIA Collections. Available at: https://www.cancerimagingarchive.net/

[Wilkinson] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016 Mar 15;3:160018. Available from: http://dx.doi.org/10.1038/sdata.2016.18

[Green] Green AK, Reeder-Hayes KE, Corty RW, Basch E, Milowsky MI, Dusetzina SB, et al. The Project Data Sphere Initiative: Accelerating Cancer Research by Sharing Data. The Oncologist. 2015 May;20(5):464-e20. Available from: http://doi.org/10.1634/theoncologist.2014-0431

[Tenenbaum] Tenenbaum JD, Sansone S-A, Haendel M. A sea of standards for omics data: sink or swim? J Am Med Inform Assoc. 2014 Mar;21(2):200–3. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932466/

[Deist] Deist TM, Dankers FJWM, Ojha P, Marshall MS, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. Radiotherapy and Oncology. 2020 Mar 1;144:189–200. Available from: http://www.thegreenjournal.com/article/S0167-8140(19)33489-9/abstract

[Meadows] Meadows B, Abrams J, Christian M, Silva J. The Common Data Element Dictionary—Developing a Standard Nomenclature for Reporting Cancer Clinical Trial Data. Proc AMIA Symp. 2001;972. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243546/

[NCI 2001a] National Cancer Institute. Automating Data Systems. 2001. Available at: http://web.archive.org/web/20010807114844/http://cancertrials.nci.nih.gov/system/html/datasystems.html

[NCI 2001b] National Cancer Institute. NCI Common Data Elements Dictionary (Version 2.0). 2001. Available at: http://web.archive.org/web/20010417093521/http://cii-server5.nci.nih.gov:8080/pls/cde_public/cde_java.show

[Hubbard] Hubbard SM, Setser A. The cancer informatics infrastructure: A new initiative of the national cancer institute. Seminars in Oncology Nursing. 2001 Feb 1;17(1):55–61. Available from: http://www.sciencedirect.com/science/article/pii/S0749208101800322

[Winget] Winget MD, Baron JA, Spitz MR, Brenner DE, Warzel D, Kincaid H, et al. Development of common data elements: the experience of and recommendations from the early detection research network. International Journal of Medical Informatics. 2003 Apr 1;70(1):41–8. Available from: http://doi.org/10.1016/S1386-5056(03)00005-4

[Nadkarni] Nadkarni PM, Brandt CA. The Common Data Elements for Cancer Research: Remarks on Functions and Structure. Methods Inf Med. 2006;45(6):594–601. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2980785/

[Vesteghem] Vesteghem C, Brøndum RF, Sønderkær M, Sommer M, Schmitz A, Bødker JS, et al. Implementing the FAIR Data Principles in precision oncology: review of supporting

initiatives. Brief Bioinform. 2019; Available from: http://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz044/5522017

[Pathak] Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc. 2011;18(4):376–86. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3128396/

[Zhang] Zhang G-Q, Cui L, Mueller R, Tao S, Kim M, Rueschman M, et al. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc. 2018 May 31;25(10):1351–8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6188513/

[Sansone] Sansone S-A, Rocca-Serra P. Review: Interoperability standards. Wellcome Trust Report. 2016 Oct 24; Available from: http://wellcome.figshare.com/articles/Review_Interoperability_standards/4055496

[Smedley] Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015 Jul 1;43(W1):W589–98. Available from: http://doi.org/10.1093/nar/gkv350

[Denny] Denny J, Glazer D, Grossman RL, Paten B, Philippakis A. A Data Biosphere for Biomedical Research. Medium - Benedict Paten. 2017. Available from: https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d

[Grossman] Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine. 2016 Sep 21;375:1109–12. Available from: http://doi.org/10.1056/NEJMp1607591

[Jensen] Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood. 2017 Jul 27;130(4):453–9. Available from: http://doi.org/10.1182/blood-2017-03-735654

[DMWG] NCI Genomic Data Commons (GDC) Data Model Working Group. Selecting Common Cross-Study Clinical Data Elements. Available from: https://gdc.cancer.gov/documentation/selecting-common-cross-study-clinical-data-elements

[Hess] Hess C, Ostrom E, editors. Understanding Knowledge as a Commons: From Theory to Practice. Cambridge, Mass: The MIT Press; 2006.

[Dearry] Dearry A. Towards a Cancer Research Data Commons. NCI BioMedical Informatics Blog. 2017. Available from: http://web.archive.org/web/20180222004839/http://ncip.nci.nih.gov/blog/towards-cancer-research-data-commons/

[Fedorov] Fedorov A, Clunie D, Ulrich E, Bauer C, Wahle A, Brown B, et al. DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research. Huisman H, editor. PeerJ. 2016 May 24;4:e2057. Available from: http://doi.org/10.7717/peerj.2057

[Elhalawani] Elhalawani H, Mohamed ASR, White AL, Zafereo J, Wong AJ, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. Scientific Data. 2017 Jul 4;4:170077. Available from: http://doi.org/10.1038/sdata.2017.77

[Grossberg] Grossberg AJ, Mohamed ASR, Halawani HE, Bennett WC, Smith KE, Nolan TS, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Scientific Data. 2018 Sep 4;5:180173. Available from: http://doi.org/10.1038/sdata.2018.173

[Fridsma] Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG Project: A Technical Report. Journal of the American Medical Informatics Association. 2008 Mar 1;15(2):130–7. Available from: http://doi.org/10.1197/jamia.M2556

[CDISC SDTM] CDISC. SDTM. [cited 2019 Aug 31]. Available from: http://www.cdisc.org/standards/foundational/sdtm

[DICOM] DICOM Standard. Available from: http://www.dicomstandard.org/

[Golbeck] Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thésaurus and Ontology. Journal of Web Semantics. 2003 Dec 1;1(1):75–80. Available from: http://doi.org/10.1016/j.websem.2003.07.007

[Sioutos] Sioutos N, Coronado S de, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics. 2007;40(1):30–43. Available from: Available from: http://doi.org/10.1016/j.jbi.2006.02.013

[Donnelly] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform. 2006;121:279–90. Available from: http://ebooks.iospress.nl/publication/9130

[FNL IDC] Frederick National Laboratory for Cancer Research. Imaging Data Commons (IDC). 2019. Available from: http://frederick.cancer.gov/workwithus/solicitations/s19-037

[Jett IDC] Jett S. The NCI Informatics Technology for Cancer Research (ITCR) Program and Imaging Data Commons. MICCAI; 2018 Sep; Granada, Spain. Available from: http://www.med.upenn.edu/sbia/assets/user-content/documents/MICCAI_Tactical_2018_slides/Jett_Stephen_NIH_NCI_ITCR_IDC_MICCAI_TACTICAL_2018.pdf

[Dearry IDC] Dearry, A. Award of the Imaging Data Commons: Bringing Multi-modal Imaging Data to the Cancer Research Community. NCI Center for Biomedical Informatics and Information Technology (CBIIT) Cancer Data Science Pulse Blog. 2019. Available from: http://datascience.cancer.gov/news-events/blog/award-imaging-data-commons-bringing-multi-modal-imaging-data-cancer-research

[Newitt multi] Newitt D, Hylton N. Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials. The Cancer Imaging Archive; 2016. Available from: http://wiki.cancerimagingarchive.net/x/EwA7AQ

[Newitt single] Newitt D, Hylton N. Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy. The Cancer Imaging Archive; 2016. Available from: http://wiki.cancerimagingarchive.net/x/ZIhXAQ

[Lingle] Lingle W, Erickson BJ, Zuley ML, Jarosz R, Bonaccio E, Filippini J, et al. Radiology Data from The Cancer Genome Atlas Breast Invasive Carcinoma [TCGA-BRCA] collection. The Cancer Imaging Archive; 2016. Available from: http://wiki.cancerimagingarchive.net/x/GQE2

[Bloch] Bloch BN, Jain A, Jaffe CC. Data From BREAST-DIAGNOSIS. The Cancer Imaging Archive; 2015. Available from: http://wiki.cancerimagingarchive.net/x/JQAo

[Li] Li X, Abramson RG, Arlinghaus LR, Chakravarthy AB, Abramson VG, Sanders M, et al. Data From QIN-Breast. The Cancer Imaging Archive; 2016. Available from: http://wiki.cancerimagingarchive.net/x/NoAaAQ

[Shah] Shah N, Feng X, Lankerovich M, Puchalski RB, Keogh B. Data from Ivy GAP. The Cancer Imaging Archive; 2016. Available from: http://wiki.cancerimagingarchive.net/x/jY9XAQ

[NBIA] NBIA REST API User Guide - Imaging - NBIA - National Cancer Institute - Confluence Wiki. Available from: http://wiki.nci.nih.gov/display/NBIA/NBIA+REST+API+User+Guide

[Källman] Källman H-E, Halsius E, Olsson M, Stenström M. DICOM Metadata repository for technical information in digital medical images. Acta Oncologica. 2009 Jan 1;48(2):285–8. Available from: http://doi.org/10.1080/02841860802258786

[dicom3tools] Clunie D. Dicom3tools Software. Available from: http://www.dclunie.com/dicom3tools.html

[Murphy] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar 1;17(2):124–30. Available from: http://doi.org/10.1136/jamia.2009.000893

[Bidgood] Bidgood WD. Documenting the information content of images. Proceedings of the AMIA Annual Fall Symposium. 1997;424–8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233520/

[Clunie SRClinTrials] Clunie DA. DICOM Structured Reporting and Cancer Clinical Trials Results. Cancer Informatics. 2007 May 12;4(CIN-ImSI-Clunie-et-al). Available from: http://journals.sagepub.com/doi/10.4137/CIN.S37032

[Clunie SR] Clunie DA. DICOM Structured Reporting. PixelMed Publishing; 2000. 394 p. Available from: http://pixelmed.com/srbook.html

[Noumeir] Noumeir R. Benefits of the DICOM Structured Report. J Digit Imaging. 2006 Dec 1;19(4):295–306. Available from: http://doi.org/10.1007/s10278-006-0631-7

[Hussein] Hussein R, Engelmann U, Schroeter A, Meinzer H-P. DICOM structured reporting: Part 2. Problems and challenges in implementation for PACS workstations. Radiographics. 2004 Jun;24(3):897–909. Available from: http://doi.org/10.1148/rg.243035722

[W3C XSLT 2.0] W3C. XSL Transformations (XSLT) Version 2.0. 2007. Available from: http://www.w3.org/TR/xslt20/

[PixelMed] Clunie D. PixelMed Publishing™ Java DICOM Toolkit. Available from: http://www.pixelmed.com/dicomtoolkit.html

[TID4200] National Electrical Manufacturers Association (NEMA). Digital Imaging and Communications in Medicine (DICOM) Standard PS3.16 - Content Mapping Resource - Breast Imaging Report Templates. Rosslyn, VA. Available from: http://dicom.nema.org/medical/dicom/current/output/chtml/part16/sect_BreastImagingReportTemplates.html

[dciodvfy] Clunie D. DICOM Validator - dciodvfy. Available from: http://www.dclunie.com/dicom3tools/dciodvfy.html

[DicomSRValidator] Clunie D. DicomSRValidator. PixelMed Publishing™ Java DICOM Toolkit. Available from:

http://www.dclunie.com/pixelmed/software/javadoc/com/pixelmed/validate/DicomSRValidator.html

[Horos] Horos Project – Free DICOM Medical Image Viewer. Available from: http://horosproject.org/

[dsr2html] OFFIS. DCMTK: dsr2html: Render DICOM SR file and data set to HTML/XHTML. Available from: http://support.dcmtk.org/docs/dsr2html.html

[CDISC SDTM v1.3] CDISC Submission Data Standards Team. Study Data Tabulation Model v1.3. 2012. Available from: http://www.cdisc.org/system/files/members/standard/foundational/sdtm/study_data_%20tabulation_%20model_v1.3.pdf

[CDISC SDTM IG HT] CDISC Submission Data Standards Team. Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.2. 2013. Available from: http://www.cdisc.org/system/files/members/standard/foundational/sdtmig/sdtmig_20v3.2_20noportfolio.pdf

[Pinnacle 21] Pinnacle 21 Community Downloads. Available from: http://www.pinnacle21.com/downloads

[CP 1838] DICOM Standards Committee. DICOM CP 1838 - Additions to Breast Imaging Report. National Electrical Manufacturers Association (NEMA); 2019. Available from: ftp://medical.nema.org/medical/dicom/final/cp1838_ft_BreastImagingReport.pdf

[NCI Ecosystem] NCI. Build a National Cancer Data Ecosystem - Cancer Moonshot Recommendation. National Cancer Institute. 2018. Available from: http://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/implementation/data-ecosystem

[Basu] Basu A, Warzel D, Eftekhari A, Kirby JS, Freymann J, Knable J, et al. Call for Data Standardization: Lessons Learned and Recommendations in an Imaging Study. JCO Clinical Cancer Informatics. 2019 Dec 1;(3):1–11. Available from: http://ascopubs.org/doi/10.1200/CCI.19.00056

[Zhou] Zhou L, Hripcsak G. Temporal reasoning with medical data—A review with emphasis on medical natural language processing. Journal of Biomedical Informatics. 2007 Apr 1;40(2):183–202. Available from: http://doi.org/10.1016/j.jbi.2006.12.009

[Allen] Allen JF. Maintaining knowledge about temporal intervals. Commun ACM. 1983 Nov 1;26(11):832–843. Available from: http://hdl.handle.net/1802/10574

[Goralwalla] Goralwalla IA, Leontiev Y, Özsu MT, Szafron D, Combi C. Temporal Granularity: Completing the Puzzle. Journal of Intelligent Information Systems. 2001 Jan 1;16(1):41–63. Available from: http://dsg.uwaterloo.ca/ddbms/publications/ozsu/JIIS/jiis624-99.pdf

[Wiederhold] Wiederhold G. The Impossibility of Global Consistency. OMICS: A Journal of Integrative Biology. 2003 Jan 1;7(1):17–20. Available from: http://doi.org//10.1089/153623103322006517

[Haendel] Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. N Engl J Med. 2018 11;379(15):1452–62. Available from: http://doi.org/10.1056/NEJMra1615014

[Mate] Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H-U, Bürkle T, et al. Ontology-Based Data Integration between Clinical and Research Systems. PLOS ONE. 2015 Jan 14;10(1):e0116656. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116656

[CCDH] CBIIT. Center for Cancer Data Harmonization (CCDH). Available from: http://datascience.cancer.gov/data-commons/center-cancer-data-harmonization-ccdh

[Cangioli] Cangioli G, Chronaki C, Kalra D, Schulz S, Stroetmann V, Thiel R, et al. How fit is SNOMED CT for eHealth interoperability in Europe? In: Health-exploring complexity (HEC): Interdisciplinary Systems Approach to Health. Munich, Germany; 2016.

[ASSESS CT] ASSESS CT - Assessing SNOMED CT for Large Scale eHealth Deployments in the EU - ASSESS CT Recommendations. 2016. Available from: http://assess-ct.eu/fileadmin/assess_ct/final_brochure/assessct_final_brochure.pdf

[CBIIT Metadata] CBIIT. Metadata for Cancer Data. Available from: http://datascience.cancer.gov/resources/metadata

[NCI CDE Browser] NCI CBIIT. CDE Browser 5.3.5 [Internet]. [cited 2019 Aug 31]. Available from: http://cdebrowser.nci.nih.gov/cdebrowserClient/cdeBrowser.html

[Segrelles] Segrelles JD, Medina R, Blanquer I, Martí-Bonmatí L. Increasing the Efficiency on Producing Radiology Reports for Breast Cancer Diagnosis by Means of Structured Reports. Methods Inf Med. 2017; Available from: Available from: http://doi.org/10.3414/ME16-01-0091

[DICOMweb] DICOM Standards Committee. DICOMweb™. Available from: http://www.dicomstandard.org/dicomweb/

[Genereaux] Genereaux BW, Dennison DK, Ho K, Horn R, Silver EL, O'Donnell K, et al. DICOMweb™: Background and Application of the Web Standard for Medical Imaging. J Digit Imaging. 2018 May 10;1–6. Available from: http://doi.org/10.1007/s10278-018-0073-z

[IHE IID] Integrating the Healthcare Enterprise (IHE). Invoke Image Display - IHE Wiki. Available from: http://wiki.ihe.net/index.php/Invoke_Image_Display

[Murphy] Murphy SN, Herrick C, Wang Y, Wang TD, Sack D, Andriole KP, et al. High Throughput Tools to Access Images from Clinical Archives for Research. J Digit Imaging. 2015 Apr 1;28(2):194–204. Available from: http://doi.org/10.1007/s10278-014-9733-9

[Gollub] Gollub RL. Enabling technologies for research using clinically acquired medical image data: Clinical Image Bank and MI2B2. 2018 Mar 5. Available from: http://www.slideshare.net/imgcommcall/enabling-technologies-for-research-using-clinically-acquired-medical-image-data-clinical-image-bank-and-mi2b2

[mi2b2] mi2b2 Home - mi2b2 - i2b2 Community Wiki [Internet]. 2014 [cited 2017 Apr 26]. Available from: http://community.i2b2.org/wiki/display/mi2b2/mi2b2+Home

[MGH/HST] MGH/HST Martinos Center for Biomedical Imaging. Medical Imaging Informatics Bench to Bedside. Available from: http://www.martinos.org/lab/mi2b2

[Klann] Klann JG, Phillips LC, Herrick C, Joss MAH, Wagholikar KB, Murphy SN. Web services for data warehouses: OMOP and PCORnet on i2b2. J Am Med Inform Assoc. 2018 Oct 1;25(10):1331–8. Available from: http://academic.oup.com/jamia/article/25/10/1331/5061849

[ISO 12052] ISO 12052 - Health informatics -- Digital imaging and communication in medicine (DICOM) including workflow and data management. 2017. Available from: http://www.iso.org/standard/72941.html

[ISO 14199] ISO 14199 - Health informatics -- Information models -- Biomedical Research Integrated Domain Group (BRIDG) Model. 2015. Available from: http://www.iso.org/standard/66767.html

[ISO 11179] ISO/IEC JTC1 SC32 WG2. ISO/IEC 11179 Information Technology -- Metadata registries. Available from: http://metadata-standards.org/11179/

[Ngouongo] Ngouongo SMN, Löbe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: Does it cover healthcare standards in empirical research? Journal of Biomedical Informatics. 2013 Apr 1;46(2):318–27. Available from: http://doi.org/10.1016/j.jbi.2012.11.008

[MAD] Metadata Automation DREAM Challenge. 2019. Available from: http://www.synapse.org/#!Synapse:syn18065891/wiki/588180

[MADResults] Metadata Automation DREAM Challenge - Results. 2020. Available from: http://www.synapse.org/#!Synapse:syn18065891/wiki/603790

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

[Mayo] Mayo CS, Kessler ML, Eisbruch A, Weyburne G, Feng M, Hayman JA, et al. The big data effort in radiation oncology: Data mining or data farming? Advances in Radiation Oncology. 2016 Oct 1;1(4):260–71. Available from: http://www.sciencedirect.com/science/article/pii/S2452109416300550

Tables

Table R-1. Standard NCI Thesaurus Data Elements and Preferred Names common across imaging collections.

Data Element					Value			In GDC
Name	**NCIt Code**	**SRT Code**	**DICOM SR Code**	**DICOM Header Attribute**	**Name**	**NCIt Code**	**SRT Code**	
Subject ID	C69258			(0010,0020) (0012,0040)				No
Anatomic Site	C13717				Brain	C12439		Yes
					Breast	C12971		
Clinical Course of Disease	C35461	None found			No Evidence of Disease	C40413	None found	No
					Recurrent Disease	C38155	None found	
Age	C69260		121033	(0010,1010)				Other
Race	C17049	S-0004D			American Indian or Alaska native	C41259	S-0004B	Yes
					Asian	C41260	S-00051	
					Black or African American	C16352	S-0004E	
					Native Hawaiian or	C41219	None found	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

					other Pacific Islander			
					Unknown	C17998		
					White	C41261	S-0003D	
Sex	C28421		121032	(0010,0040)	Female	C16576		Other
					Male	C20197		
Laterality	C25185	G-C171			Left	C25229	T-04030	No
					Right	C25228	T-04020	
Diagnosis	C15220		111042		Anaplastic Astrocytoma	C9477		Yes
					Astrocytoma	C60781		
					Benign	C14172	M-80000	
					Breast Carcinoma	C4872	D0-F0357	
					Breast Fibroadenoma	C3744	M-90100	
					Breast Fibrocystic Change	C3039	D7-90310	
					Ductal Breast Carcinoma In Situ	C2924	M-85002	
					Glioblastoma	C3058		
					Invasive Breast Carcinoma	C9245	D0-F0377	
					Invasive Ductal Carcinoma	C4194	M-85003	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

					Invasive Lobular Breast Carcinoma	C7950	M-85203	
					Mixed Neoplasm	C6930	M-893FF	
					Stromal Hyperplasia	C35857	M-72430	
					Unknown	C17998		
Estrogen Receptor Status	C16150		111475		Estrogen Receptor Negative	C15493	R-40759	No
					Estrogen Receptor Positive	C15492	G-A200	
					Estrogen Receptor Status Unknown	C15495		
HER2/Neu Status	C16152				HER2/Neu Negative	C68749	R-40759	No
					HER2/Neu Positive	C68748	G-A200	
					HER2/Neu Status Unknown	C68750		
Progesterone Receptor Status	C16149		111476		Progesterone Receptor Negative	C15497	R-40759	No
					Progesterone Receptor Positive	C15496	G-A200	
					Progesterone Receptor Status Unknown	C15498		
Vital Status	C25717		LN:11323-3		Alive	C37987	F-05036	Yes
					Dead	C28554	F-04DA1	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

					Lost to follow up	C48227	F-00FBE	
					Unknown	C17998		

Notes:

1. In general, DICOM does not explicitly code the concept of “unknown” but rather omits sending the concept at all.
2. DICOM has its own coding scheme for use when concepts are not found in external lexicons
3. In some cases, the concept is encoded in DICOM SR in the “header” instead of or in addition to the content tree.
4. The receptor status value concepts defined in NCI Thesaurus are pre-coordinated with the test, whereas in the DICOM SR Breast Imaging Report Template generic codes are used and presumed to be post-coordinated with the name of the name-value pair.
5. Old-style SNOMED RT (SRT) IDs were used as codes in DICOM until recently and hence in the project described here, even though the concepts are from SNOMED CT; in future, SCT numeric concept IDs will be used.
6. For some CDEs, the GDC Common Cross-Study list uses similar but sufficiently different concepts that they are marked as “Other” in this table, e.g., Sex versus Gender, Age versus Age At Diagnosis.

Figures

Figure R-MP-1. High level approach to data element mapping.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

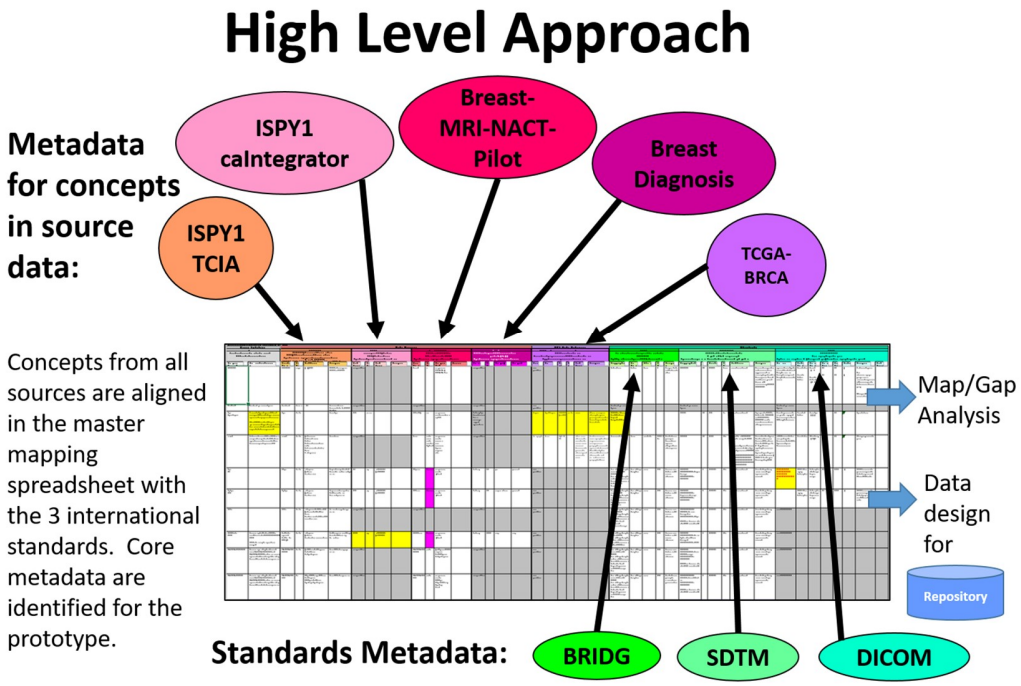

Figure R-DB-.1 i2b2 query for triple-negative breast cancer.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1

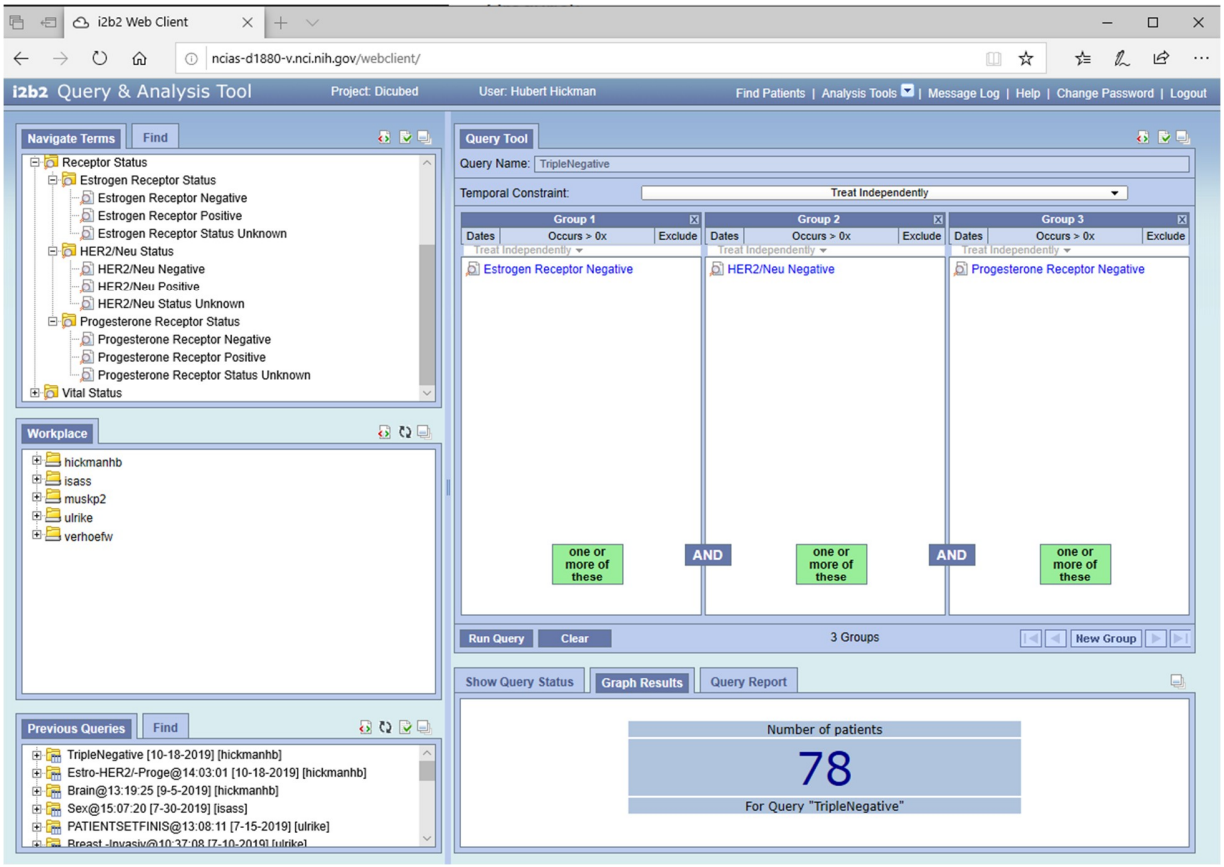

Figure R-DB-2. i2b2 linkage to TCIA images.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2020 doi:10.20944/preprints202008.0474.v1