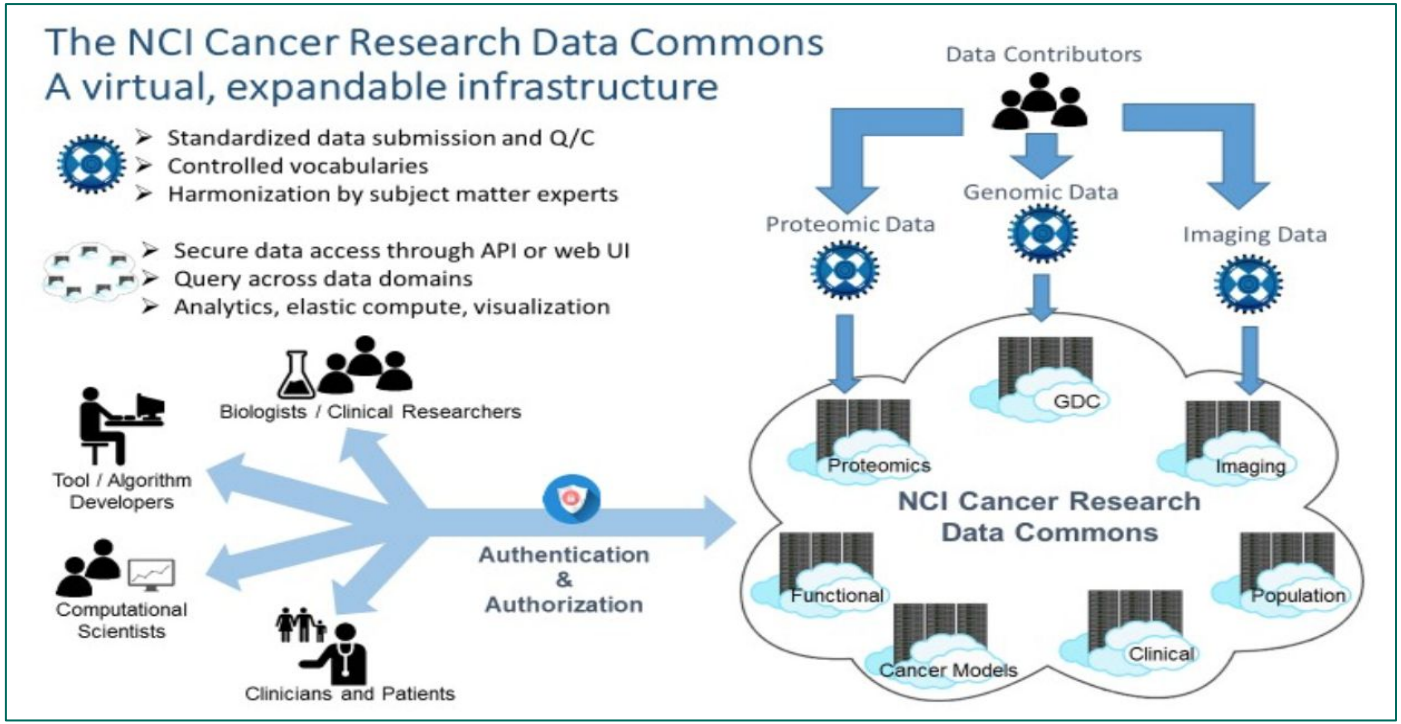
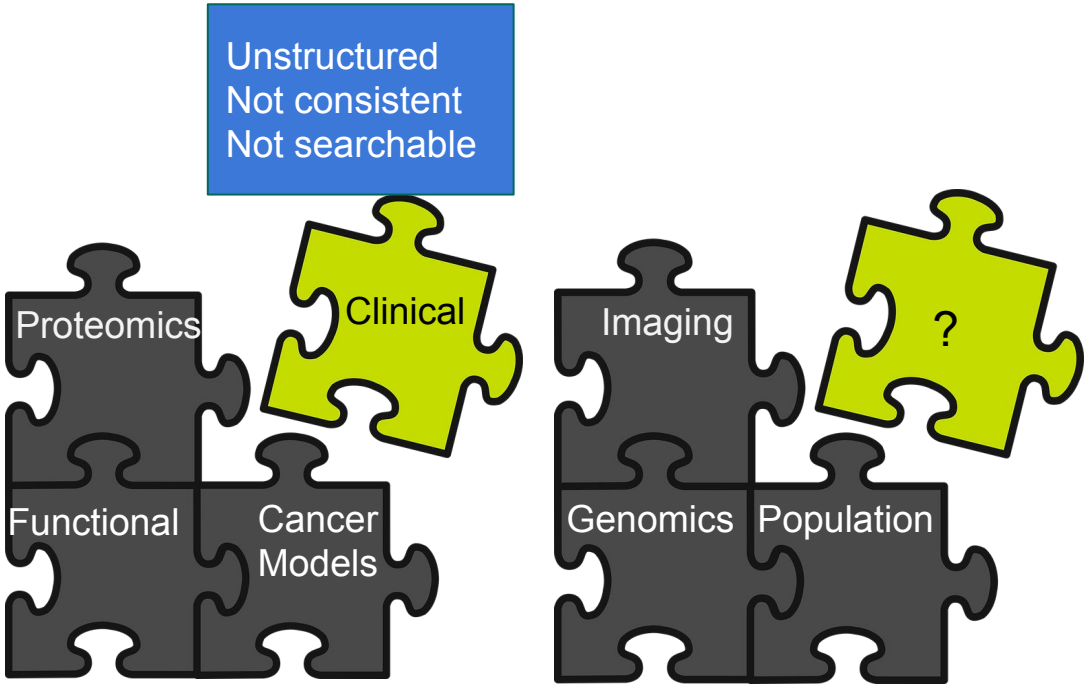
TCIA Data Harmonization

Amrita Basu, Ph.D.
Presidential Innovation Fellow
National Cancer Institute
amrita.basu@gsa.gov

Data Ecosystem



Motivation

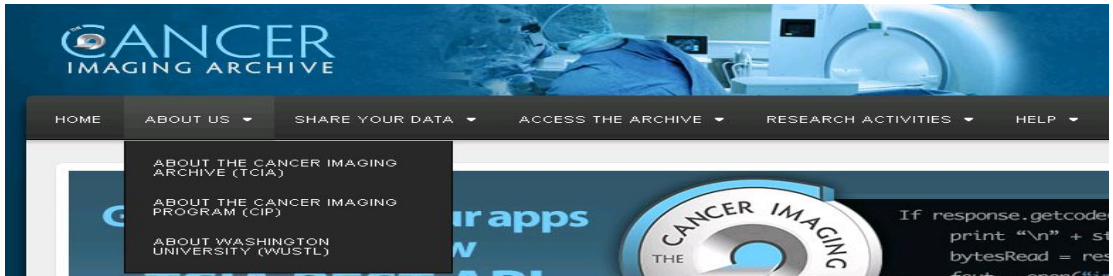

Unstructured
Not consistent
Not searchable
Proteomics
Clinical
Functional
Cancer Models
Imaging
?
Genomics
Population

TCIA User Base

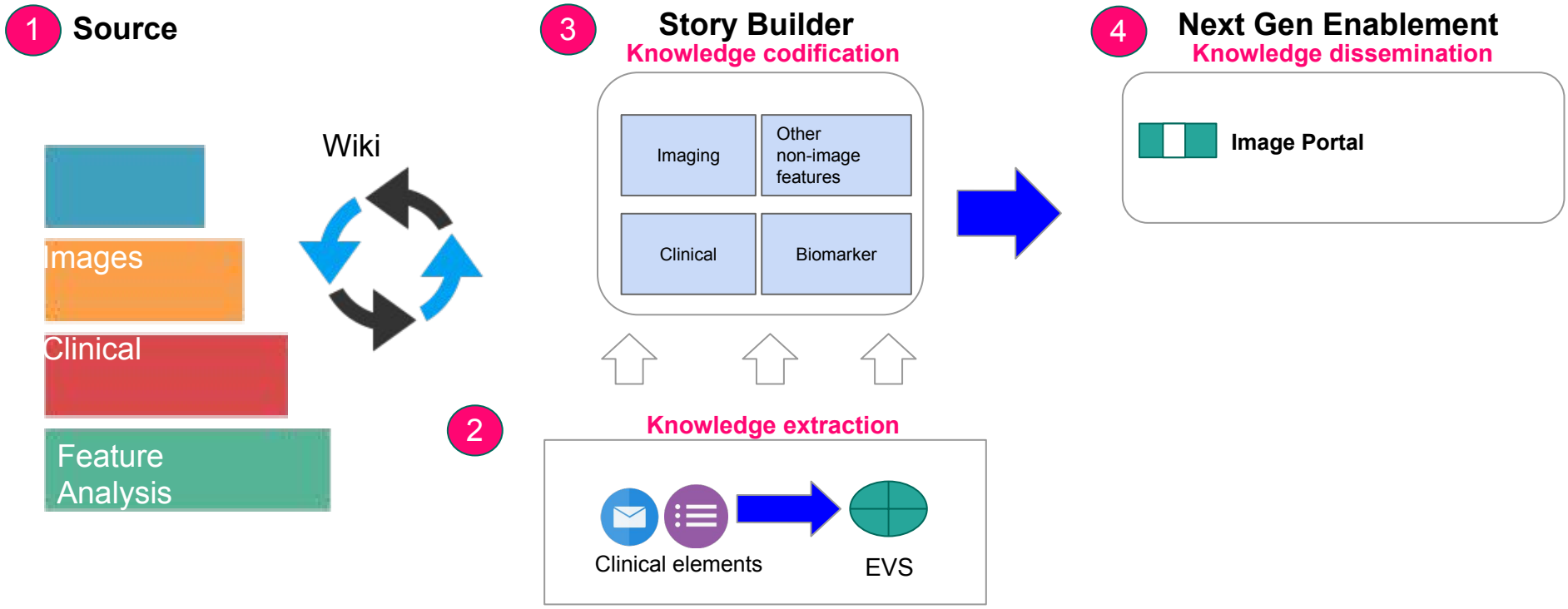
Support the TCIA user community's requests (from many QIN investigators and others) to enable clinical data queries for cohort selection.

The main types of data submissions to TCIA are:

1. large projects like TCGA/CPTAC/APOLLO where NCI has control over how clinical data is collected via contract obligations
2. NCI clinical trials where NCI has some control over how clinical data is collected
3. investigator collected data derived from NCI grants or tied to peer reviewed publications where NCI has less control over how clinical data is collected. Public databases like **TCIA** are coming into heavier use there is finally a driving purpose for researchers to try to harmonize.



1 Architecture and Roadmap



2 Overview

- **Goal**: Support a prototype clinical data portal to query across shared TCIA clinical data files
- **Task**: Selected 9 files: 5 lung cancer and 4 brain cancer clinical datasets to map to the GDC Clinical Data Elements
- **5 Lung clinical data files from TCIA web Portal:**

1. LungCTDiagnosis - All the images were diagnostic contrast enhanced CT scans. The images were retrospectively acquired, to ensure sufficient patient follow-up. All images were done at diagnosis and prior to surgery. The objective of the study was to extract prognostic image features that will describe lung adenocarcinomas and will associate with overall survival.
2. NSCLC-Radiomics - This collection contained images from 422 non-small cell lung cancer (NSCLC) patients. For these patients pretreatment CT scans, manual delineation by a radiation oncologist of the 3D volume of the gross tumor volume and clinical outcome data are available. This study was published in Nature Communications.
3. NSCLC-Radiomics-Genomics - This collection contained images from 89 non-small cell lung cancer (NSCLC) patients that were treated with surgery. For these patients pretreatment CT scans, gene expression, and clinical data are available. This study was published in Nature Communications.
4. TCGA LUAD.clinical.patient - The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) data collection is part of a larger effort to build a research community focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome Atlas (TCGA). Clinical, genetic, and pathological data resides in the Genomic Data Commons (GDC) Data Portal while the radiological data is stored on The Cancer Imaging Archive (TCIA).
5. TCGA LUSC.clinical.patient - The Cancer Genome Atlas Lung Squamous Cell Carcinoma (TCGA-LUSC) data collection is part of a larger effort to build a research community focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome Atlas (TCGA). Clinical, genetic, and pathological data resides in the Genomic Data Commons (GDC) Data Portal while the radiological data is stored on The Cancer Imaging Archive (TCIA).

Overview continued

- **4 Brain cancer clinical files from TCIA web portal**:

1. **ROI-Masks-Low-Grade-Glioma-Tumor** - This collection contains 406 ROI masks in MATLAB format defining the low grade glioma (LGG) tumour region on T1-weighted (T1W), T2-weighted (T2W), T1-weighted post-contrast (T1CE) and T2-flair (T2F) MR images of 108 different patients from the TCGA-LGG collection. From this subset of 108 patients, 81 patients have ROI masks drawn for the four MRI sequences (T1W, T2W, T1CE and T2F), and 27 patients have ROI masks drawn for three or less of the four MRI sequences.The ROI masks were used to extract texture features in order to develop radiomic-based multivariable models for the prediction of isocitrate dehydrogenase 1 (IDH1) mutation, 1p/19q codeletion status, histological grade and tumour progression.

   - Clinical data (188 patients in total from the TCGA-LGG collection, some incomplete depending on the clinical attribute), VASARI scores (188 patients in total from the TCGA-LGG collection, 178 complete) with feature keys, and source code used in this study are also available with this collection.

2. **REMBRANDT**- Rembrandt contains data generated through the Glioma Molecular Diagnostic Initiative from 874 glioma specimens comprising approximately 566 gene expression arrays, 834 copy number arrays, and 13,472 clinical phenotype data points. The file contained the pre-surgical magnetic resonance (MR) multi-sequence images from 130 REMBRANDT patients.

3. **MR-ImagingPredictors** –Study patients had been previously de-identified by the Cancer Genome Atlas (TCGA). Presurgical MR images of 75 patients with GBM with genetic data in the TCGA portal were rated by three neuroradiologists for size, location, and tumor morphology by using a standardized feature set. Interrater agreements were analyzed by using the Krippendorff α statistic and intraclass correlation coefficient. Associations between survival, tumor size, and morphology were determined by using multivariate Cox regression models; associations between imaging features and genomics were studied by using the Fisher exact test.

4. **OutComePredictors** - Correlates patient survival with morphologic imaging features and hemodynamic parameters obtained from the nonenhancing region (NER) of glioblastoma (GBM), along with clinical and genomic markers. Forty-five patients

Harmonization Approach

- Selected 108 GDC clinical data elements as the target for harmonization for fields from the 9 files
  - GDC Data Dictionary Viewer
    - **Demographics**
    - **Diagnosis**
    - **Exposure**
    - **Family History**
    - **Follow-up**
    - **Sample**
    - **Treatment**
- Used file column headings and data values to align each data element to the GDC, or TCGA if GDC CDE did not exist. The TCGA files contained CDE references, we retrieved the permitted data values from the caDSR
- Manual alignment of data values from each file with the GDC standard data values
- Manual transformation of the data
- Summary metrics, field level mappings created for each field

Mapping Results – 37 (15% of total 255) Fields found in 3 or more files

- **Values matched exactly or matched a subset** (includes upper/lower case differences and months/days/years conversions) (16)
  - Gender - GDC
  - Race - GDC
  - Ethnicity - GDC
  - ICD 10 - TCGA
  - Days to Death - GDC
  - Days to Follow-up - GDC
  - Days to Birth - GDC
  - Year of Pathologic Diagnosis - GDC
  - Tissue or organ of origin- GDC
  - Tissue Source Site - GDC
  - Patient BCR Barcode - TCGA
  - Laterality - GDC
  - Karnofsky Performance Status - TCGA
  - Karnofsky Status Timing - TCGA
  - Radiation Therapy Indicator - TCGA
  - Person Neoplasm Status (last known) - GDC

- **Data values are not the same but are semantically equivalent** and can be converted or rolled up to support query (21)
  - Age at Diagnosis - GDC
  - TNM Stage (includes 8 fields T, N, M clinical and pathologic) - GDC
  - Histology/Diagnosis (required SME) - GDC
  - Vital Status - GDC
  - Anatomic Organ Subdivision - TCGA
  - Survival Time (although meaning is ambiguous)
    - Survival Time at last known Vital Status
    - Where present with Days to Death, the data values are different
  - Histologic Grade - GDC
  - Therapy Type - GDC
  - New Tumor after event diagnosis - TCGA
  - Primary Tumor Outcome - GDC
  - Progression or Recurrence - GDC
  - TP53 Mutation Status - TCGA
  - EGFR Mutation Status - TCGA

Observations by Cancer “Site”

- “Most Common” = Fields found in 3 or more files by Brain and Lung

Most Common Brain Cancer fields		Most Common Lung Cancer Fields		Most common overlapping fields across both Diseases
days_to_death		anatomic organ Subdivision		gender
EGFR		Clinical M Stage		Histologic type
gender		Clinical N Stage		Patient Age at Diagnosis
Histologic Type		Clinical T Stage		Vital Status
karnofsky_performance_score		gender		Patient Identifier
Patient Age at Diagnosis		Histologic Type		"Site" of Disease - Inferred
Patient Identifier		Overall.Stage		
PDGFRA		Pathologic M Stage		
progression_or_recurrence		Pathologic N Stage		
Survival Time		Pathologic T Stage		
therapy_type		Patient Age at Diagnosis		
TP53		Patient Identifier		
Vital Status		TNM Pathologic		
Site of Disease - Inferred		Vital Status		
		Site of Disease- Inferred		
14		15		6
5%		6%		2%

Preliminary results of analysis of 9 TCIA files vs GDC CDEs

Interpretation:

Common Fields – # fields found in the 9 TCIA files

- 140 Fields were found in only 1 file,
  - 78 were found in 2 or more files, etc.
  - Total of 255 fields

GDC Match – Number of GDC fields matching the Common Fields

Summary:

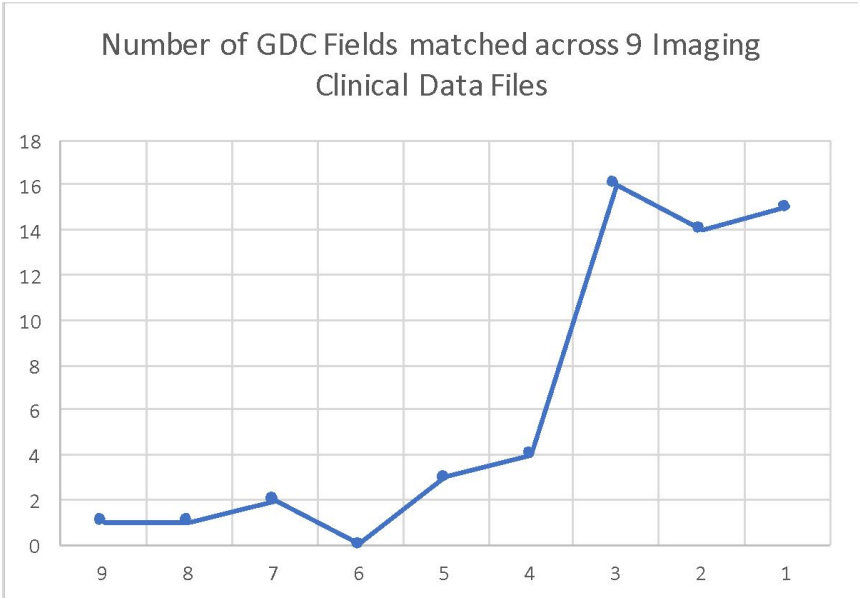
- 37 fields are found to be common across 3 or more files
- Of those 37, 23 matched GDC fields, or 62%

- Note: Recent mapping of coded elements to SDTM
  - 26 of the 37 fields have codelists
  - 7 (27%) of the SDTM codelists match fairly well
  - 12 (46%) do not have matching SDTM codelist
  - 14 (54% have codelists, but don't match well

#Files Containing the field	Common Fields	GDC match	% Match
1	140	16	11%
2	78	15	19%
3	24	14	58%
4	4	3	75%
5	2	2	100%
6	2	0	0%
7	2	2	100%
8	0	0	-
9	3	2	67%
Totals - All Field	**255**	**54**	**21%**
Fields in 3 or More Files	**37**	**23**	**62%**

Matched GDC CDEs across 9 Imaging files

e.g. There were 16 GDC fields that matched fields across 3 files



Harmonization Challenges

- Different data value encodings across the same field in different files (Except for TCGA files)
- Little field name or data value documentation
  - The meaning of some data values was hard to figure out e.g. Histologic Type
- Different data structures and representations
  - Some are indicators versus selecting from a controlled list of values (data modeled differently)
    - E.g. “Therapy Type” where “Radiation” is one of the choices vs “Radiation Therapy Indicator”
  - Days vs Months vs Years
- Mappings not one-to-one
  - Stratification granularity differences – 5 choices vs 3 choices
  - Semantic similarity across several fields: Progression, Progression or Recurrence, Recurrence, Time to new Tumor Event

Observations/Lessons Learned

1. **Very difficult and time consuming to harmonize data after its been collected**
   - Field names and values can be ambiguous, definitions are not available for the fields, sometimes the documentation does not match the data
2. **“Data files” can be in different formats, need different approaches to support harmonization**
   - Data file type (csv, txt) vs Vs Summarized data (not individual patient) in Word document
3. **CDEs help make the meaning of the data clear**
   - “morphology” vs “International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code”
4. **“Query fields“ can be derived to enable cross-file query, bring back the original data values**
   - E.g. No “type of therapy” data element in file X, but other data elements describing particular kinds of therapy are provided e.g. if there is a value for Chemotherapy_Agent_Name, we can derive that the patient was given chemotherapy
5. **Identifying query use cases would help define minimal set of required clinical elements**
6. **Query interface should support viewing hierarchu for fields such as Histology**
7. **Query interface should provide context sensitive to support**
   - query to accomodate disease specific features
8. **SMEs are valuable to help with mapping e.g. Histologic type mapping difficulties**
   - Differences in granularity “adenocarcinoma” vs “Lung Papillary Adenocarcinoma”, synonyms, “Mucinous (Colliod) Carcinoma” = Lung Mucinous Adenocarcinoma”, combined histology: “Papillary Type AND Adenocarcinoma, Bronchiolo-alveolar Features”, Solid Type And Acinar

Observations from curators about looking for matching CDEs

- Curation style matters: How the CDE creator choose to semantically define the data variable matters; the order in which the concepts are selected and applied can limit what is returned based on the query search order.
- Data variable creation method: Tool vs manually curated; Most caDSR content is manually created and use specific datatypes and max size limits (ex. (Char(5), Number(3,2)) rather than the generic “java” representation typically associated with model content. Therefore, a match on the question may not occur due to the datatype/CDE representation type.
- Data Source: higher level of matches among data sets from same source (TCGA sources vs non-TCIA sources). Also, mapping limited based on level of supporting documentation provided to describe the variable, especially when acronym and abbreviations used in variable name/XLS column heading.
- Equivalent response format: Coded vs Text response; Semantics of the question matched but the response was coded differently, e.g. indicator versus choosing from a list of answers
- Pre-analysis constraints: Mapping performed in specific order; when multiple CDEs found, priority given to GDC CDEs (limited to clinical GDC CDEs), then TCGA CDEs, then any caDSR CDE, helped to guide mapping process.

Conclusions and suggestions

- Matching to GDC generally provided well defined, target data values for harmonization
- It may be beneficial to create a “Key” of the transformed values at the beginning of each “row” to use for query/searching, but return the actual data values in the search results
  - Use protocol information to populate some of the fields
  - Use correlated fields to derive data values for query
    - Merge Clinical and Pathologic Stage for query
    - Derive indicator fields
- Availability of standards in easily consumable formats that describe the fields (like information found in CDEs) up front before data collection would be beneficial
  - Need to provide a template with ‘typed’ columns to support researcher data submission using common, valid data formats/standards
- Provide a data transformation tool to researchers to simplify and streamline data harmonization and transformation
- Identify fields/data elements that would support common/high priority queries for combining data across clinical data files

Questions?

- Contact:
- Denise Warzel warzeld@mail.nih.gov
- Aras Eftekhari aras.eftekhari@nih.gov (Brain/Lung Mapping)
- Erin Muhlbradt muhlbradtee@mail.nih.gov (CDISC/SDTM)
- Janice Knable janice.knable@nih.gov (Lung Mapping)

3 Testing

1)User-based, iterative feedback model

2)Requirements are gathered, user input, and prototype

3)Really important design concept is the USER

4 Prospective Data

1. Who are the primary users and what are their incentives to submit data?
2. Burden on us vs user? Middle ground?
3. Flexibility for user
4. Evaluation of submission strategies

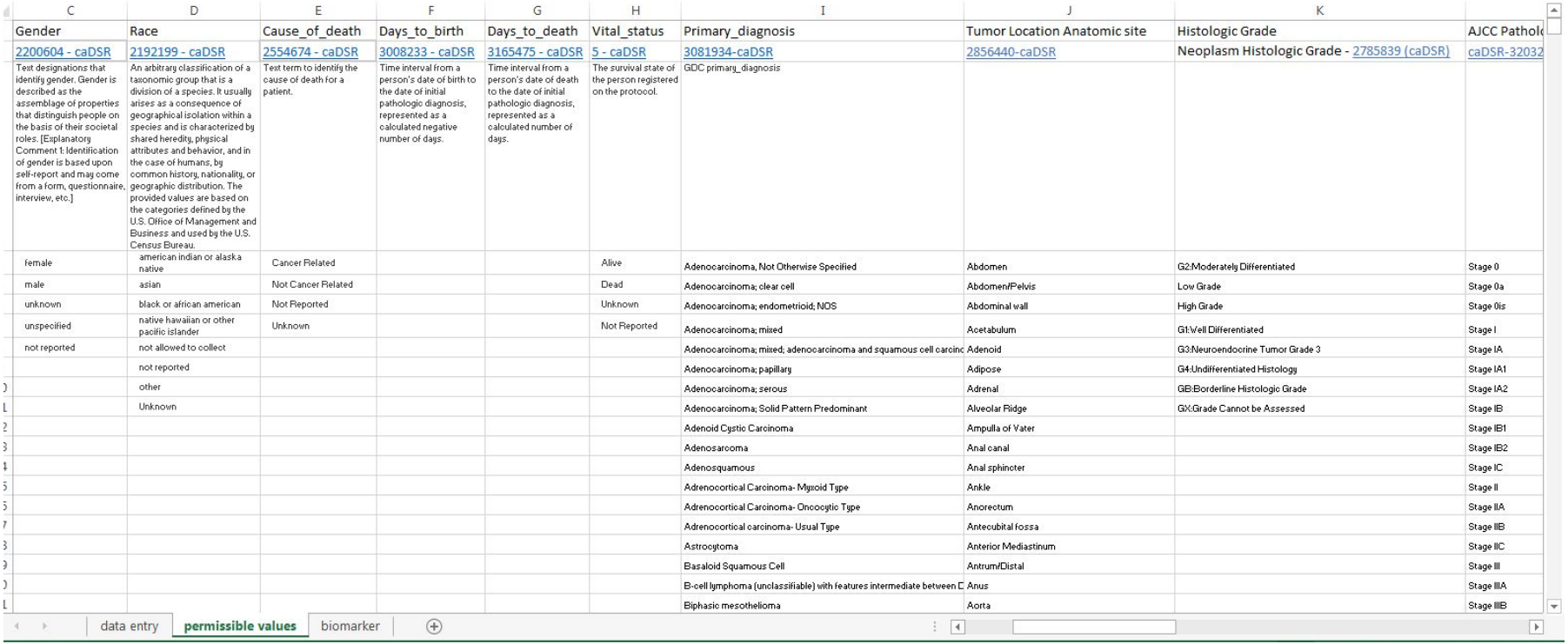
Constrained spreadsheet for submission

	A	B	C	D	E	F	G	H	I	J	K
1	Collection name	Site id	Patient id	Ethnicity	Gender	Race	Cause of death	Vital Status	Days to birth	Days to death	Primary diagnosis
2	NSCLC-Radiomics	12	1	not hispanic or latino	male	black or african american	Cancer Related	Alive	770	100	Non-Small Cell Lung Cancer (NOS)
3	NSCLC-Radiomics	12	2	Unknown	female	asian	Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
4	NSCLC-Radiomics	12	3	not hispanic or latino	male	american indian or alaska nativ	Cancer Related	Unknown	770	100	Non-Small Cell Lung Cancer (NOS)
5	NSCLC-Radiomics	12	4	not hispanic or latino	male	asian	Cancer Related	Unknown	770	100	Non-Small Cell Lung Cancer (NOS)
6	NSCLC-Radiomics	12	5	not reported	unspecified	black or african american	Not Reported	Alive	770	100	Non-Small Cell Lung Cancer (NOS)
7	NSCLC-Radiomics	12	6	not hispanic or latino	male	american indian or alaska nativ	Not Reported	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
8	NSCLC-Radiomics	12	7	not hispanic or latino	female	american indian or alaska nativ	Not Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
9	NSCLC-Radiomics	12	8	Unknown	male	not reported	Cancer Related	Alive	770	100	Non-Small Cell Lung Cancer (NOS)
10	NSCLC-Radiomics	12	9	not hispanic or latino	female	black or african american	Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
11	NSCLC-Radiomics	12	10	Unknown	male	black or african american	Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
12	NSCLC-Radiomics	12	11	Unknown	male	black or african american	Not Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
13	NSCLC-Radiomics	12	12	not hispanic or latino	female	black or african american	Not Cancer Related	Alive	770	100	Non-Small Cell Lung Cancer (NOS)
14	NSCLC-Radiomics	12	13	not hispanic or latino	male	black or african american	Cancer Related	Alive	770	100	Non-Small Cell Lung Cancer (NOS)
15	NSCLC-Radiomics	12	14	not hispanic or latino	unspecified	american indian or alaska nativ	Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
16	NSCLC-Radiomics	12	15	not hispanic or latino	male	american indian or alaska nativ	Cancer Related	Alive	770	100	Non-Small Cell Lung Cancer (NOS)
17	NSCLC-Radiomics	12	16	not hispanic or latino	male	asian	Not Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
18	NSCLC-Radiomics	12	17	not hispanic or latino	unknown	black or african american	Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
19	NSCLC-Radiomics	12	18	not hispanic or latino	male	black or african american	Not Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
20	NSCLC-Radiomics	12	19	not hispanic or latino	male	black or african american	Not Cancer Related	Dead	770	100	Non-Small Cell Lung Cancer (NOS)
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											
32											

data entry | permissible values | biomarker

Tumor Location Anatomic site | Histologic Grade | AJCC Pathologic Stage | TNM Pathology | AJCC Pathologic N | AJCC Pathologic M | AJCC Pathologic T | AJCC Clinical T | AJCC Clinical N | AJCC Clinical M | AJCC Overall Clinical stage | Ann Arbor Stage | Biomarker name

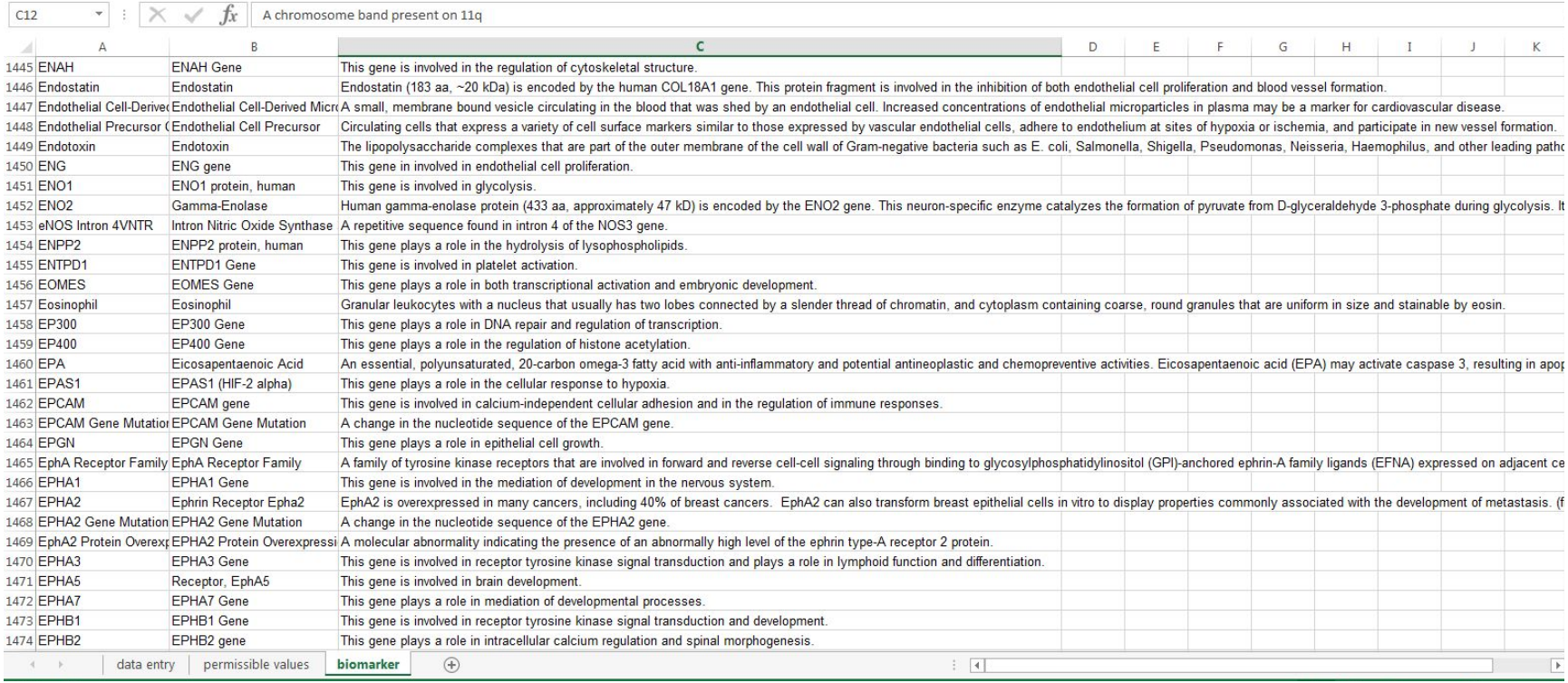
Constrained spreadsheet for submission



C	D	E	F	G	H	I	J	K	
Gender	Race	Cause_of_death	Days_to_birth	Days_to_death	Vital_status	Primary_diagnosis	Tumor Location Anatomic site	Histologic Grade	AJCC Pathol
2200604 - caDSR	2192199 - caDSR	2554674 - caDSR	3008233 - caDSR	3165475 - caDSR	5 - caDSR	3081934-caDSR	2856440-caDSR	Neoplasm Histologic Grade - 2785839 (caDSR)	caDSR-32032
Text designations that identify gender. Gender is described as the assemblage of properties that distinguish people on the basis of their societal roles. [Explanatory Comment 1: Identification of gender is based upon self-report and may come from a form, questionnaire, interview, etc.]	An arbitrary classification of a taxonomic group that is a division of a species. It usually arises as a consequence of geographical isolation within a species and is characterized by shared heredity, physical attributes and behavior, and in the case of humans, by common history, nationality, or geographic distribution. The provided values are based on the categories defined by the U.S. Office of Management and Business and used by the U.S. Census Bureau.	Text term to identify the cause of death for a patient.	Time interval from a person's date of birth to the date of initial pathologic diagnosis, represented as a calculated negative number of days.	Time interval from a person's date of death to the date of initial pathologic diagnosis, represented as a calculated number of days.	The survival state of the person registered on the protocol.	GDC primary_diagnosis			
female	american indian or alaska native	Cancer Related			Alive	Adenocarcinoma, Not Otherwise Specified	Abdomen	G2:Moderately Differentiated	Stage 0
male	asian	Not Cancer Related			Dead	Adenocarcinoma; clear cell	Abdomen/Pelvis	Low Grade	Stage 0a
unknown	black or african american	Not Reported			Unknown	Adenocarcinoma; endometrioid; NOS	Abdominal wall	High Grade	Stage 0is
unspecified	native hawaiian or other pacific islander	Unknown			Not Reported	Adenocarcinoma; mixed	Acetabulum	G1:Well Differentiated	Stage I
not reported	not allowed to collect					Adenocarcinoma; mixed; adenocarcinoma and squamous cell carcinc	Adenoid	G3:Neuroendocrine Tumor Grade 3	Stage IA
	not reported					Adenocarcinoma; papillary	Adipose	G4:Undifferentiated Histology	Stage IA1
	other					Adenocarcinoma; serous	Adrenal	GB:Borderline Histologic Grade	Stage IA2
	Unknown					Adenocarcinoma; Solid Pattern Predominant	Alveolar Ridge	GX:Grade Cannot be Assessed	Stage IB
						Adenoid Cystic Carcinoma	Ampulla of Vater		Stage IB1
						Adenosarcoma	Anal canal		Stage IB2
						Adenosquamous	Anal sphincter		Stage IC
						Adrenocortical Carcinoma- Myxoid Type	Ankle		Stage II
						Adrenocortical Carcinoma- Oncocytic Type	Anorectum		Stage IIA
						Adrenocortical carcinoma- Usual Type	Antecubital fossa		Stage IIB
						Astrocytoma	Anterior Mediastinum		Stage IIC
						Basaloid Squamous Cell	Antrum/Distal		Stage III
						B-cell lymphoma (unclassifiable) with features intermediate between D	Anus		Stage IIIA
						Biphasic mesothelioma	Aorta		Stage IIIB

data entry | permissible values | biomarker

Constrained spreadsheet for submission



C12 | A chromosome band present on 11q

	A	B	C
1445	ENAH	ENAH Gene	This gene is involved in the regulation of cytoskeletal structure.
1446	Endostatin	Endostatin	Endostatin (183 aa, ~20 kDa) is encoded by the human COL18A1 gene. This protein fragment is involved in the inhibition of both endothelial cell proliferation and blood vessel formation.
1447	Endothelial Cell-Derive	Endothelial Cell-Derived Micr	A small, membrane bound vesicle circulating in the blood that was shed by an endothelial cell. Increased concentrations of endothelial microparticles in plasma may be a marker for cardiovascular disease.
1448	Endothelial Precursor (	Endothelial Cell Precursor	Circulating cells that express a variety of cell surface markers similar to those expressed by vascular endothelial cells, adhere to endothelium at sites of hypoxia or ischemia, and participate in new vessel formation.
1449	Endotoxin	Endotoxin	The lipopolysaccharide complexes that are part of the outer membrane of the cell wall of Gram-negative bacteria such as E. coli, Salmonella, Shigella, Pseudomonas, Neisseria, Haemophilus, and other leading patho
1450	ENG	ENG gene	This gene in involved in endothelial cell proliferation.
1451	ENO1	ENO1 protein, human	This gene is involved in glycolysis.
1452	ENO2	Gamma-Enolase	Human gamma-enolase protein (433 aa, approximately 47 kD) is encoded by the ENO2 gene. This neuron-specific enzyme catalyzes the formation of pyruvate from D-glyceraldehyde 3-phosphate during glycolysis. It
1453	eNOS Intron 4VNTR	Intron Nitric Oxide Synthase	A repetitive sequence found in intron 4 of the NOS3 gene.
1454	ENPP2	ENPP2 protein, human	This gene plays a role in the hydrolysis of lysophospholipids.
1455	ENTPD1	ENTPD1 Gene	This gene is involved in platelet activation.
1456	EOMES	EOMES Gene	This gene plays a role in both transcriptional activation and embryonic development.
1457	Eosinophil	Eosinophil	Granular leukocytes with a nucleus that usually has two lobes connected by a slender thread of chromatin, and cytoplasm containing coarse, round granules that are uniform in size and stainable by eosin.
1458	EP300	EP300 Gene	This gene plays a role in DNA repair and regulation of transcription.
1459	EP400	EP400 Gene	This gene plays a role in the regulation of histone acetylation.
1460	EPA	Eicosapentaenoic Acid	An essential, polyunsaturated, 20-carbon omega-3 fatty acid with anti-inflammatory and potential antineoplastic and chemopreventive activities. Eicosapentaenoic acid (EPA) may activate caspase 3, resulting in apop
1461	EPAS1	EPAS1 (HIF-2 alpha)	This gene plays a role in the cellular response to hypoxia.
1462	EPCAM	EPCAM gene	This gene is involved in calcium-independent cellular adhesion and in the regulation of immune responses.
1463	EPCAM Gene Mutatio	EPCAM Gene Mutation	A change in the nucleotide sequence of the EPCAM gene.
1464	EPGN	EPGN Gene	This gene plays a role in epithelial cell growth.
1465	EphA Receptor Family	EphA Receptor Family	A family of tyrosine kinase receptors that are involved in forward and reverse cell-cell signaling through binding to glycosylphosphatidylinositol (GPI)-anchored ephrin-A family ligands (EFNA) expressed on adjacent ce
1466	EPHA1	EPHA1 Gene	This gene is involved in the mediation of development in the nervous system.
1467	EPHA2	Ephrin Receptor Epha2	EphA2 is overexpressed in many cancers, including 40% of breast cancers. EphA2 can also transform breast epithelial cells in vitro to display properties commonly associated with the development of metastasis. (f
1468	EPHA2 Gene Mutation	EPHA2 Gene Mutation	A change in the nucleotide sequence of the EPHA2 gene.
1469	EphA2 Protein Overexp	EPHA2 Protein Overexpressi	A molecular abnormality indicating the presence of an abnormally high level of the ephrin type-A receptor 2 protein.
1470	EPHA3	EPHA3 Gene	This gene is involved in receptor tyrosine kinase signal transduction and plays a role in lymphoid function and differentiation.
1471	EPHA5	Receptor, EphA5	This gene is involved in brain development.
1472	EPHA7	EPHA7 Gene	This gene plays a role in mediation of developmental processes.
1473	EPHB1	EPHB1 Gene	This gene is involved in receptor tyrosine kinase signal transduction and development.
1474	EPHB2	EPHB2 gene	This gene plays a role in intracellular calcium regulation and spinal morphogenesis.

data entry | permissible values | biomarker

Feedback

Retrospective:

1) What would you like to be able to query? Your most common uses cases (think across datasets, within a single dataset, and multiple attributes i.e. I want to know patient ids of those patients that were diagnosed within the last five years with lung adenocarcinoma)

2) Will this type of interface be useful for your research?

Prospective:

3) What clinical elements are not currently in the template and would be useful for the TCIA user? Any other comments on the template?

Thank You!

Justin Kirby
John Freymann

Ulrike Wagner
Ed Helton
Smita Hastak

Denise Warzel
Aras Eftekhari
Janice Knable
Sherri De Coronado

Fred Prior
Lawrence Tarbox

Ashish Sharma (NCI ITCR U24 - 1U24CA215109-01)

Anthony Kerlavage
Eve Shalley
Juli Klemm