MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Use Case Exercise 1:

Metadata extraction, standardization, and integration

The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation. This technical data deliverable was developed using contract funds under Basic Contract No. W56KGU-18-D-0004.

Use Case Exercise 1: Metadata extraction, standardization, and integration

Scenario: The Clinical Proteomic Tumor Analysis Consortium (CPTAC) program wants to make the data from its multiple studies readily available in an integrated, standardized format in the Cancer Research Data Commons (CRDC). Much of the patient case and sample data are available in supplemental tables in the original primary research papers, but with varying variable names and value types. This exercise is based on Li et al. 2023, Proteogenomic data and resources for pan-cancer analysis: Cancer Cell, which integrated and harmonized data from ten CPTAC source studies.

Task: Semi-automatically extract and integrate patient case metadata from the ten source studies in Li et al. Create a harmonized dataset in the Genomics Data Commons (GDC) data format for the 15 variables as shown in Table A below. Include only patients that have tumor samples with proteogenomic data and were not excluded in the source studies.

To illustrate, Table A shows data for one patient case from one paper and corresponding information in the GDC data format. Note, there is not a one-to-one mapping for some variables, and values for some GDC variables can be inferred from other variables in the paper table data.

Table A. Data from Dou et al. 2020 and corresponding data from the GDC for one patient case.

Dou et al. Table S1		GDC-formatted data	
Variable	**Value**	**Variable**	**Value**
Proteomics_Participant_ID	C3L-00006	case_submitter_id	C3L-00006
Age	64	age_at_diagnosis	23376
Gender	Female	gender	female
Race	White	race	white
Ethnicity	Not-Hispanic or Latino	ethnicity	not hispanic or latino
(none)	(none)	vital_status[1]	Alive[1]
Histologic_Grade_FIGO	FIGO grade 1	tumor_grade	G1
tumor_Stage-Pathological	Stage I	ajcc_pathologic_stage	Stage I
Path_Stage_Reg_Lymph_Nodes-pN	pN0	ajcc_pathologic_n	N0
Path_Stage_Primary_Tumor-pT	pT1a (FIGO IA)	ajcc_pathologic_t	T1a
Tumor_Focality	Unifocal	tumor_focality	Unifocal

Tumor_Size_cm	2.9	tumor_largest_dimension_diameter	2.9
Tumor_Site	Anterior endometrium	tissue_or_organ_of_origin	Endometrium
Histologic_type	Endometrioid	primary_diagnosis	Endometrioid carcinoma
Histologic_type; Tumor_Site	Endometrioid; Anterior endometrium	morphology	8380/3
Case excluded	No	(None, but presence in this dataset indicates the sample should be included)	

[1]Data not present in Dou et al. 2020 and therefore probably not possible to obtain from the original source.

Sources for the case data: The assembled patient case data in Li et al. is in the paper's Table S1. In this excel file, there is a data dictionary on the first sheet. Li et al. did not use all the standard GDC variables and values, and one part of this exercise is to generate the GDC-formatted output data from the data in Li et al. Table S1.

Table B below lists the ten source papers cited in the Li et al. paper ("*In the original publications investigating a single cancer cohort,*[11, 12, 13, 14, 15,16,17,18,19,20] *data were processed…")* and their associated files that contain patient case metadata. The variables and value types within these files vary. As noted in the last column, half of these tables have an associated data dictionary with variable descriptions.

Table B. The primary research source papers with patient case data used in Li et al.

First Auth.	Paper Link	Relevant table(s)	Has Data Dictionary
Clark, D.J.	Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma: Cell	Table S1	FALSE
Krug, K.	Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy	Supplemental_Table_1	TRUE
Vasaikar, S.	Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities: Cell	Table S1	TRUE
Wang, L.-B.	Proteogenomic and metabolomic characterization of human glioblastoma: Cancer Cell	Table S1	FALSE
Huang, C.	Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma: Cancer Cell	Table S1	TRUE
Satpathy, S.	A proteogenomic portrait of lung squamous cell carcinoma: Cell	Table S1	TRUE
Gillette, M.A.	Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma: Cell	Table S1	TRUE
McDermott, J.E.	Proteogenomic Characterization of Ovarian HGSC Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal Instability: Cell Reports Medicine	Table S1	FALSE

Cao, L.	Proteogenomic characterization of pancreatic ductal adenocarcinoma: Cell	Table S1	FALSE
Dou, Y.	Proteogenomic Characterization of Endometrial Carcinoma: Cell	Table S1	FALSE

Target output data: The target output for this task is in the file **Li_data_in_GDC_format.csv**. This file contains a subset of the data corresponding to the studies from Li et al. Table S1 converted to GDC format. This conversion was done by the MITRE team using a manually created script for extracting and formatting variables and values, and with some manual curation from information in Li et al. Table S1. The file includes 15 GDC identifier, demographic, and clinical variables. In addition, the file includes two variables that may be useful in assessing results: *tumor_code*, which has the sample's tumor code given in Li et al., and *study*, which indicates the primary research paper ("study") from the list above that each sample came from.

Information about the GDC data format: The GDC provides definitions for the GDC variables and acceptable values in a data dictionary (see Info about the GDC Data Dictionary and View the GDC Data Dictionary). All the variables, other than *case_submitter_id*, are within Clinical under Demographic, Diagnosis, and Pathology Detail. Specifications for the value types are also available in gdcdictionary/src/gdcdictionary/schemas at develop · NCI-GDC/gdcdictionary · GitHub. The GDC also provides other tools and information that might be of use, including an API (GDC API Info), information about data submission (Submission - GDC Docs), and term search functionality (Search - GDC Docs).

The GDC variables *primary_diagnosis*, *morphology*, and *tissue_or_organ_of_origin* make use of terminologies from the WHO's International Classification of Diseases for Oncology (ICD-O-3; documented in the *International Classification of Diseases for Oncology, Third Edition*, which can be downloaded for free here.) Updated values for *primary_diagnosis* and *morphology* are available on the site SEER ICD-O-3 Coding Materials (cancer.gov) in this file: https://seer.cancer.gov/icd-o-3/sitetype.icdo3.d20220429.xlsx. The value terms for *primary_diagnosis* are in the column "Histology/Behavior Description". Values for *morphology* consist of codes that correspond to the terms under "Histology/Behavior Description" and are in the column "Histology/Behavior". Some additional synonyms for the *primary_diagnosis* terms and corresponding *morphology* codes, can be found here: Copy-of-ICD-O-3.2_MFin_17042019_web.xls. The values for *tissue_or_organ_of_origin* are text terms that correspond to ICD-O topography (site) codes, e.g. "Brain, NOS" corresponds to "C71.9". These text terms can be found in the book International Classification of Diseases for Oncology, Third Edition in the Topography section starting on page 45; the terms to use are in bold next to ICD-O topography codes. The website Site-Specific Modules | SEER Training (cancer.gov) provides modules with topography code

information sorted by broad cancer types; two websites that provide simplified lists of these terms for *tissue_or_organ_of_origin* are here and here. Note, the GDC data dictionary and schemas provide lists of enumerated values for these three variables, but some of these values are not in the ICD-O terminologies and some of the listed values appear to be errors; values from the official ICD-O-3 terminologies may be preferred.

Additional information about the target output data: The 15 GDC variables in the target output were selected because there is data for each of them in at least some of the primary source papers. However, not all variables have data in all the primary source papers, and consequently, the target output dataset based on Li et al. includes some data that will probably not be obtainable from the primary papers. On the other hand, there is richer data for some variables, such as *tissue_or_organ_of_origin*, in some primary source papers than in Li et al., which could yield different and more specific values than in the target output based on Li et al. The presence/absence of data for each of the target variables in each of these source papers is given in the file **presence_absence_var_data_in_papers.xlsx**; this is based on manual inspection of the source paper supplemental tables. Note that some of the variables in the target output data that are marked as present are not in the source paper tables but have values that can be inferred either from other variables in the table and/or from the paper text.

Part of this exercise is to semi-automatically identify cases with tumor samples with usable proteogenomic data. Consequently, the target output file contains only the patient cases in Li et al. Table S1 that have tumor samples (as opposed to only having normal tissue samples) with proteogenomic data and that are not marked as excluded. Li et al. Table S1 includes 117 cases that are marked as excluded (113 cases) and/or do not have tumor sample ids (4 additional cases). Of these, 68 cases occur in the primary source tables. For at least 66 of these 68 cases, the primary source files include some information that indicates the case should be excluded and/or lacks tumor sample data; the two remaining cases have reasons listed in Li et al., possibly based on subsequent reanalysis of the data.

Additional available data: There is data in the GDC for the patient cases in the target dataset, which is available for download in json and tsv formats via the “Clinical” button at https://portal.gdc.cancer.gov/projects/CPTAC-3 and https://portal.gdc.cancer.gov/projects/CPTAC-2. However, this GDC data is missing information that is present in Li et al. and the source papers, and the GDC data contains inconsistencies with the Li et al. data, formatting inconsistencies, and typos; consequently, for the target output, we chose to use the richer data in the Li et al. Table S1, converted to GDC format.

The GDC also contains data associated with multiple studies and papers beyond the set in this task, which could potentially be used as training data. However, as in the case of Li et al., there may be substantial mismatches in the GDC and paper data values.

Variations on this task:

- Start with just the source paper references, and automatically identify and extract the supplemental tables to be harmonized
- Start with a query to identify CPTAC program papers to gather data from. Note, there are additional, newer CPTAC papers with data beyond the set here
- Perform some of the steps in a natural language interface
- In addition to putting the data into the GDC format, map to additional formats such as the format used in the Li et al. paper and standard data elements used by other NCI-related resources (see caDSR II (cancer.gov))