MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Use Case Exercise 2:

Finding patient case, ‘omics, and image data

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 2: Finding patient case, 'omics, and image data

Assembling multi-modal cancer datasets requires searching for and integrating data from papers and multiple data repositories. For a researcher, this is currently a time-consuming and tedious process. The purpose of this exercise is to develop flexible, low- or no-code tools that will simplify and accelerate the process of finding and obtaining comprehensive cancer datasets.

For this exercise, we have provided a scenario and a set of tasks that involve finding patient cases, locating data files, extracting data elements, and finding descriptions of methods. We worked through some of the tasks using currently available tools and have provided information about our results to illustrate possible answers and sources of specific data. To inform tool development, we have also provided information about domain-relevant data sources and tools, and some tips about these tasks and data sources. Finally, to convey some specific challenges and opportunities, we included a list of time-consuming activities that are currently involved in finding cancer datasets.

Scenario: Researchers want to investigate potential druggable pathways for endometrial carcinoma, and to do so, want to assemble a large dataset of patient, 'omic, and image data. This is loosely based on Dou et al. 2023, Proteogenomic insights suggest druggable pathways in endometrial carcinoma.

Tasks:

1. Generate a list of case ids for patients with endometrial carcinoma that have tumor samples with whole genome sequence (WGS) data. In addition, identify which of these cases have mass spectrometry proteome data and which have available Computed Tomography (CT) imaging data.

2. Develop capabilities to locate, download, and/or extract the following kinds of data for patient cases. To develop and assess this, we recommend trying this with a few patient case ids from different studies, e.g., C3L-01311, C3L-01355, and TCGA-B5-A11S.
    a. Patient demographic and clinical data from at least two sources for each patient case id; collectively, these should include patient age, vital status, BMI, and whether the patient has diabetes (the latter two may not be available for all cases)
    b. Data about the tumor and normal tissue samples collected from patients, including the ids of samples and aliquots used in analyses

c. A scientific paper, or paper supplementary information document, with information about the specimen collection methods, and information about how specimens were preserved
d. Whole genome sequence (WGS) data for a tumor sample, BAM format (just provide URL for the data)
e. Somatic mutation data, preferably from an ensemble of mutation callers, in MAF format; also, information about the software methods used to generate this data. (Somatic mutation data consists of mutations that are found in a tumor sample but not in a normal sample from the same patient.)
f. Whether the tumor sample has mutations in the following genes, and if so, what types of mutations: PTEN and PIK3R1
g. Copy number variant data by segment (paired tumor and normal sample data; this is the change in copy number in a tumor sample versus a normal sample from the same patient); also, information about the software methods used to generate it
h. miRNA quantification data from a tumor sample and from a normal sample, if available
i. Histology slide image data (provide URL to view or to download)
j. An aggregated acetylome, acetyl site dataset with data for the patient case, if available
k. Proteome mass spec data for the case if available, including:
    - The aggregated protein assembly data from TMT mass spec with this case
    - The biospecimen data file associated with the proteomic data
    - The proteome mass spec run metadata
    - The files with the processed mass spectra in open standard format that contain data for this case. Also identify the name of the column with the data for this case

3. Identify additional endometrial cancer patient cases with other data types
    a. Additional endometrial cancer patient cases with data about *PTEN*, *PIK3CA*, and *PIK3R1* mutations in tumor samples and histology data (the mutation data can come from targeted gene sequencing or Whole Exome Sequencing (WES or WXS), so there are cases beyond the set with WGS data for tumor samples)
    b. Endometrial patient cases for which single cell RNA-seq data from a tumor sample is or will be available soon

Example Task Results: For Task 1, the excel file **Exercise 2 - Task 1 example results.xlsx** contains a list of endometrial carcinoma patient case ids and whether

WGS data, mass spectrometry proteome data, and CT imaging data are available. We generated this list by searching for and reading a few papers and supplementary information files, and by using CRDC repository user interfaces. This list includes patient cases from a few studies-- it is not comprehensive and there are more cases with these data types that you can discover in other papers and repositories.

For Task 2, the excel file **Exercise 2 - Task 2 information for case C3L-01311.xlsx** has information that we identified for one patient case id. This information includes file names, URLs, and in some cases directions for manually accessing files. It also includes values we identified from data files, excerpts of methods descriptions and their sources, and in some cases, our rationale for finding and selecting data. We generated this by using information in papers and their supplementary files, and by using data repository user interfaces and APIs. This is not comprehensive and there are additional sources and alternative data files for some of these sub-tasks. Also, this is not intended as a template for your output.

To date, we have not generated results for Task 3, but could do so if you think that would be useful.

Data Sources: These tasks require clinical, 'omics, and image data that are distributed over multiple papers and data repositories. To get started, we have provided a list of some resources with endometrial cancer data for these tasks. You are welcome to use additional resources too.

Publications with supplemental files

- Dou et al. 2023, Proteogenomic insights suggest druggable pathways in endometrial carcinoma: Cancer Cell
- Dou et al. 2020, Proteogenomic Characterization of Endometrial Carcinoma: Cell
- Li et al. 2023, Proteogenomic data and resources for pan-cancer analysis: Cancer Cell
- Levine et al. 2013, Integrated genomic characterization of endometrial carcinoma | Nature
- Liu et al. 2018, An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics: Cell

Cancer Research Data Commons (CRDC) repositories

- Genomic Data Commons (GDC)
- Proteomic Data Commons (PDC)
- Imaging Data Commons (IDC)

- Cancer Data Service (CDS)

Each of these CRDC repositories has a user interface for researchers to find data using filters and basic keyword search. In addition, there are multiple alternative ways to access data in the GDC and PDC. Both utilize knowledge graphs (GDC Data Model, GDC Data Dictionary Information, Viewer - GDC Dictionary, PDC data dictionary) which can be queried with APIs (GDC API Info, PDC - Query and API). There are also software packages for querying and accessing GDC data (a GDC list here). These include python tools and guides such as Genomic Data Commons · GitHub, GitHub - broadinstitute/gdctools, and GDC python query documentation and the R packages Bioconductor - GenomicDataCommons and Bioconductor - TCGAbiolinks. In addition, CRDC data can be accessed via the Cancer Genomics Cloud platform by Seven Bridges and the Broad's FireCloud via Terra; the Terra software is open-source.

Other data repositories and resources

- NCI Human Tumor Atlas Network (HTAN) (info about additional data access options)
- cBioPortal for Cancer Genomics (info about APIs and data access options)
- The Cancer Imaging Archive (TCIA) (info about data access)
- LinkedOmics
- UCSC Xena
- National Center for Biotechnology Information (NCBI) (APIs and software packages available)
    - DataBase of Genomes And Phenotypes (dbGAP) (access to some data is restricted)
    - Gene Expression Omnibus (GEO)
    - Sequence Read Archive (SRA)
    - BioProject and BioSample
- European Genome-Phenome Archive (EGA) (European version of dbGAP)

Additional Information and Tips:

1. Dou et al. 2023 analyzes cases from three cohorts, which they term “independent cohort”, “exploratory cohort” and “TCGA cohort”. For Task 1, your set of cases, at minimum, should include cases from all three cohorts. The three recommended patient case ids in Task 2 include one case from each of these cohorts.
2. You can start by using data repositories or information from papers. You are free to use Dou et al. 2023 to find cases and information about where to find datasets. Note though that the information in this paper will not be sufficient, and at least one

of the paper's pointers to datasets is not currently correct (possibly because datasets are typically posted after the papers are published, and sometimes different decisions are ultimately made about where to post the data and how to label it).

3. Using all the data repositories listed is not necessary to complete the tasks, but each has some data that could potentially be useful. For Tasks 1 and 2, we recommend using at least GDC, PDC, and IDC along with a few of the papers listed above and their supplemental files; NCBI repositories and other publications have information about additional patient cases not included in our example task results. For Task 3a, GDC, cBioPortal, and NCBI repositories have useful information. Work on Task 3b should utilize HTAN; there is also relevant data in the NCBI repositories.
4. All the data repositories listed provide a way to search by cancer, organ type, and/or cancer type, and we recommend doing this. Endometrial carcinoma is sometimes called Uterine Corpus Endometrial Carcinoma, with the abbreviation UCEC and its primary site is the uterus. Note, in some resources, multiple variables and multiple terms are used to label cases of endometrial cancer, and consequently searching will likely require querying more than just a single variable for a single value.
5. Some sub-tasks ask for methods used to produce datasets, which are important for selecting among alternate versions of data and for identifying datasets from similar analytical pipelines that are appropriate to integrate. Sample collection and preparation methods are particularly important for some kinds of data, e.g., transcriptomic data from RNA sequencing is sensitive to the methods used to preserve samples. Information about methods can be found in papers, paper supplementary documents, sample metadata, and websites documenting repository pipelines. You can choose how to provide information about the methods, e.g., provide a reference for a source with this information, a snippet or section of text from a paper, or a text response created with retrieval augmented generation.
6. Some of the data types and files have restricted access. Gaining access to them should not be necessary to complete these tasks; providing URLs to where the data can be obtained is sufficient.
7. Some of the data files are large, particularly sequencing read data, imaging data, and mass spec data. We recommend that you don't download these. Collectively, the other data files may also be large, and a storage space approach may need to be considered.
8. While some data types include case ids (e.g., "C3L-01311") in their associated metadata, not all data types do. For many patient cases, there are multiple biospecimen samples, including samples from normal tissue as well as tumor

samples, and each biospecimen sample has its own id (e.g., “C3L-01311-01”). Furthermore, there are often multiple aliquots from a given sample, e.g. biospecimen sample “C3L-01311-01” has aliquots with ids “CPT0077770013” and “CPT0077770003”. Different aliquots of a sample get used in different lab assays, e.g., one aliquot may be used for whole genome sequencing while another is used for RNA sequencing (RNA-Seq). Some data types, such as sequencing read data and some proteomic data files, include aliquot ids but not sample and case ids in their metadata. Consequently, mapping between case ids, biospecimen sample ids, and aliquot ids may be necessary. The data to map between these is available via the GDC and PDC APIs, is in biospecimen data files from GDC and PDC, and sometimes is in paper supplemental tables.

9. For the proteome data: Unaggregated proteome TMT mass spec data files are multiplexed, with data from nine samples (typically from nine different patient cases) in one file; the data for each sample is labelled with a probe name, e.g. “128N”. Information about which samples are in each multiplex mass spec run and what probe they are labelled with in the run is in run metadata files and should be accessible via the PDC API. To locate unaggregated TMT data for a specific case, you will need to map across run metadata, biospecimen sample and/or aliquot ids, and case ids.

Appendix: Time-consuming activities researchers currently perform to find data

- Finding papers about relevant datasets; getting associated supplemental data and information files.
- Finding data repositories that may have the sought-after data.
- Chasing down links in papers and websites to relevant data; finding alternatives when links no longer work or are incorrect.
- Searching the content of papers, paper supplemental files, and websites to find the methods that were used to create datasets, so that the data can be used and integrated appropriately.
- Finding and understanding the meaning of variables and values in datasets, to determine whether this data is useful for the analyses of interest.
- Identifying which variables and values to search on in each data repository to comprehensively obtain the data of interest.
- Cross-linking data, including dealing with id formats that vary, use of multiple ids for the same entity, and the ids of interconnected entities, e.g., aliquots taken from biospecimen samples from patient cases, which are variously associated with data files.
- Finding information about and selecting among different versions of similar data, e.g., datasets with mutation calls from different software platforms; different

versions of clinical data that may have incongruous values for things like vital status.

- Navigating data repository user interfaces with numerous filter options and screens, which sometimes have significant learning curves and sometimes do not provide ways to perform all the queries a researcher wants.
- Writing code for queries and data retrieval, including identifying software packages and APIs for repositories; understanding each repository's data structure sufficiently to write queries; learning to write code in the format needed for each repository; debugging.
- Selecting and setting up ways to manage and store metadata and data, which can be large.