Oncology Analysis Report

Application outline
Palantir Foundry Workshop application backed by Ontology Manager, Pipeline Builder and Graph Data Explorer + python preprocessing scripts.

Data Cohort
I chose a cohort made up of patients with breast or lung cancer that had available open access PDF clinical reports. This cohort was chosen because it had good differentiation across several datapoints which I chose for maximum impact on drawing insights about survival rates.

The cohort was made up of ~2200 cases across ~2200 documents. From this I inferred that with approximately one document per patient that I simply needed 500 of the PDFs to account for ~500 patients.

Data Preprocessing
Next I did some data cleaning which had two purposes; to increase the quality of the data that the LLM would draw upon and also reduce the total number of files. I did this using a Python script which iterated through all files and made copies of PDFs that contained 6 or more keywords which were essentially the names of desirable datapoints.
After some investigation I found that not all PDFs had patient demographic included in the content. Since gdc.cancer.gov provides filters for age, sex and vital status I used the manifest export function to create lists of sub cohorts which were fed into a separate Python script. This then annotated the PDFs with demographic data.

Palantir Foundry
I was able to repurpose a Foundry template and add some tweaks to accommodate this project. The pipeline has a number of stages:

- Split text data into chunks
- Extract case_id from filename
- Create a unique chunk ID
- Extract demographic annotations into distinct columns/table
- Generate chunk summaries
- Extract entities

An LLM block is given demographic fields and case_id as well as content to give it the best chance of cross referencing useful information.

Next, Ontology Manager is used to define the entity, chunk and demographic objects as well as a join table so they can be linked in a vertex graph.
An AIP Logic function specifies in detail how an LLM should use the transformed data to create meaningful responses to queries.
Finally a Workshop application gives an interface for queries as well as some data visualisation.

Challenges
Due to the restricted compute resources on a dev account the rebuild time while developing a pipeline can be quite slow. This can make testing a bit tricky. Luckily only new outputs are generated each time it seems.

Before I added the demographic table I was finding that the LLM was not able to infer useful stats about patient demographics. Since each pipeline transform block cannot compare all input files simultaneously but instead processes them one by one, an aggregate table is required. How to Load this extra table as extra input for queries was not obvious. I managed to do it by querying it with the tool function only available in Chain of Thought Mode.

The prompt for the logic function also required fine tuning to get it to actually compare cohorts. I had to be very specific about each stage to make the most out of the response.

For some reason the LLM in the Workshop app seems to give worse answers than when testing in the logic function. It seems to truncate the number of chunks referenced in the front end app, perhaps a dev account restriction?

If I had more time:

- I would look into how to squeeze more detail out of the LLM responses, as the thought process logs show quite detailed analysis.
- Create an object link between entities and cohorts to see visually if some conditions are associated with certain life situations.
- Import some other data types such as nucleotide variation data to match oncogene mutations with outcomes.

Notes

All files used were open access from https://portal.gdc.cancer.gov

userQuestion *

· "Show me survival trends for [cancer type] patients by treatment type"

Function output (x) Use LLM

Result

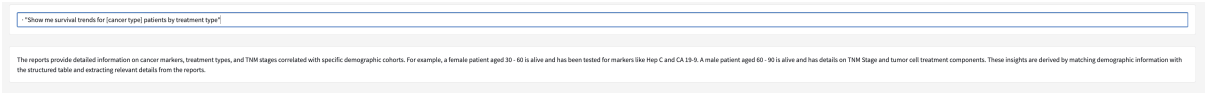
Successfully ran

The provided summaries and content offer insights into various patient demographics and their associated cancer markers, treatments, and diagnoses. Here's a detailed analysis:

1. **Demographic Cohorts:**
  - **Age Group 30 - 60, Female, Alive:**
    - Cancer markers include Hep C, CA 19-9, AFP, PSA, and B/T Cell Markers.
    - Treatment details are not explicitly mentioned, but the demographic is associated with a report on cancer markers.
  - **Age Group 60 - 90, Male, Alive:**
    - This cohort appears multiple times across different summaries, indicating a focus on studies related to NSCLC and CRC.
    - TNM stage and treatment type are mentioned, with a specific case having a TNM stage of 2 and no treatment.
    - Another report mentions tumor cell treatment components with over 50% involvement.
  - **Age Group 60 - 90, Female, Alive:**
    - Cancer markers include Hep C, CA 19-9, AFP, PSA, and B/T Cell Markers.
    - Treatment details are not explicitly mentioned, but the demographic is associated with a report on cancer markers.

2. **Cancer Markers and Treatments:**
  - **Cancer Markers:**
    - Hep C, CA 19-9, AFP, PSA, and B/T Cell Markers are frequently mentioned across different demographics.
  - **Treatments:**
    - TNM stage and treatment type are specified for a male patient aged 60 - 90, with a TNM stage of 2 and no treatment.
    - Tumor cell treatment components are mentioned for a male patient aged 60 - 90, with over 50% involvement.

3. **Correlations and Examples:**
  - The male cohort aged 60 - 90 is prominently featured in studies related to NSCLC and CRC, indicating a focus on these cancer types.
  - The female cohort aged 30 - 60 is associated with a variety of cancer markers, suggesting a broad diagnostic approach.



userQuestion *

· "What factors are most associated with positive outcomes in this dataset?"

Function output (x) Use LLM

Result

Successfully ran

The reports focus on female patients aged 60-90 who are alive, discussing treatment decisions for stage II, node-positive breast cancer. They emphasize considering clinical factors like tumor size, lymph node involvement, and steroid receptor status, rather than relying solely on HER-2/neu gene amplification. The Allred score is used to determine ER/PR status, and the PathVysion HER-2 DNA Probe Kit is mentioned for patient selection.

Hide