CANCER SUMMARY WORKFLOW

1. Purpose

- Generates the cancer summary dataset combining all the cancer incidents in GS data.
- Dataset helps to analyse risk factors for cancers.
- Updated version is produced every time it is executed.

2. Design

- High-level architecture: Figure 1 represents the flow of the process

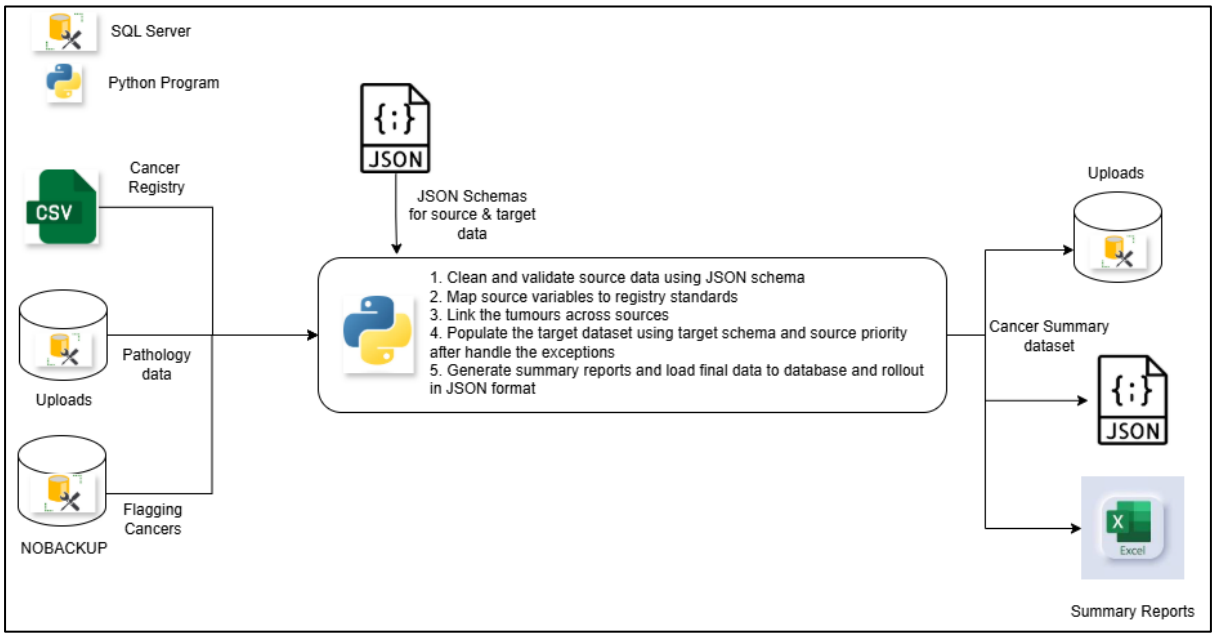

Figure 1: Cancer Summary workflow

- Cancer cases are considered only from confirmed/Registry/NHS Flagging as input to build the result dataset.
- The process combines NHS Flagging data and PHE (cancer registry) dataset for the GS cohort to produce Cancer Summary dataset with variables detailing the cancer.
- **Key pointers:**
    - The tumours across sources are linked based on StudyID, Diagnosis Date (+-60 days), ICD_CODE, MORPH_CODE, or LATERALITY
    - ICD_CODE is set to '**C56'** for all the ovarian tumours from histopathology report.

- For breast tumours from histopathology report, if
  - InvasiveCarcinoma='P', then ICD_CODE='**C50**'
  - InvasiveCarcinoma='N' and InsituCarcinoma='P', then ICD_CODE='**D05**'
- Table 1 shows the linking rules between different sources

SOURCE 1	SOURCE 2	LINKING COLUMNS
Histopathology breast report	Registry	StudyID, Diagnosis Date (+-60 days), ICD_CODE, LATERALITY
Histopathology breast report	Flagging	StudyID, Diagnosis Date (+-60 days), ICD_CODE
Histopathology ovarian report	Flagging	StudyID, Diagnosis Date (+-60 days), ICD_CODE
Histopathology ovarian report	Registry	StudyID, Diagnosis Date (+-60 days), ICD_CODE
Flagging	Registry	StudyID, Diagnosis Date (+-60 days), ICD_CODE, MORPH_CODE

Table 1: Tumour linking rules

- The decision to set the diagnosis date range to 60 days is based on the observations from Figure 2 and Figure 3.
- Most of the tumour diagnosis dates are lie within the range +/- 60 days when compared to Registry data.

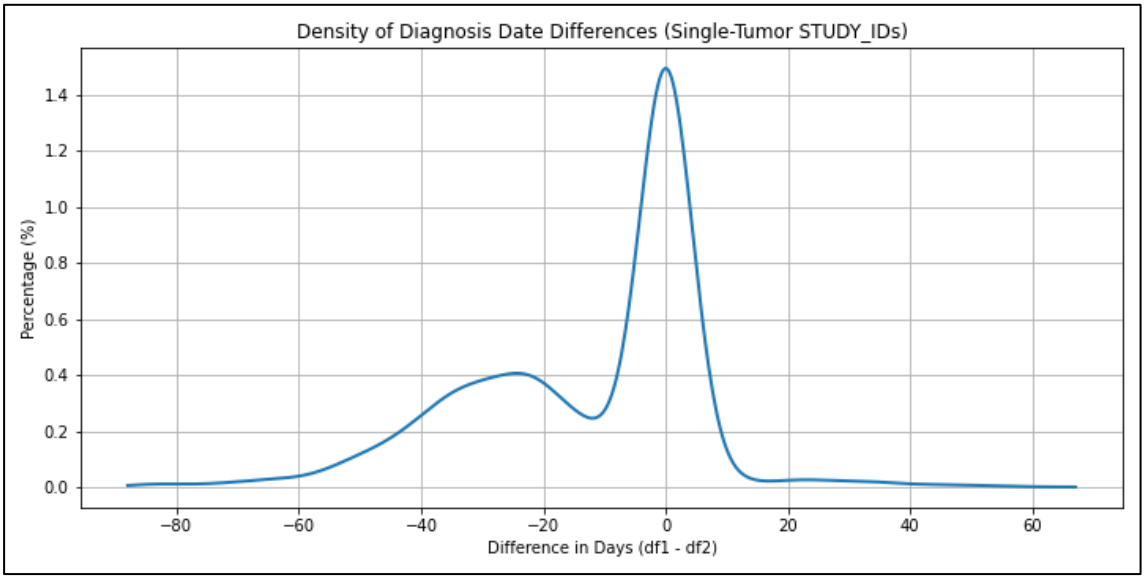

Figure 2: Registry vs Breast Path Report

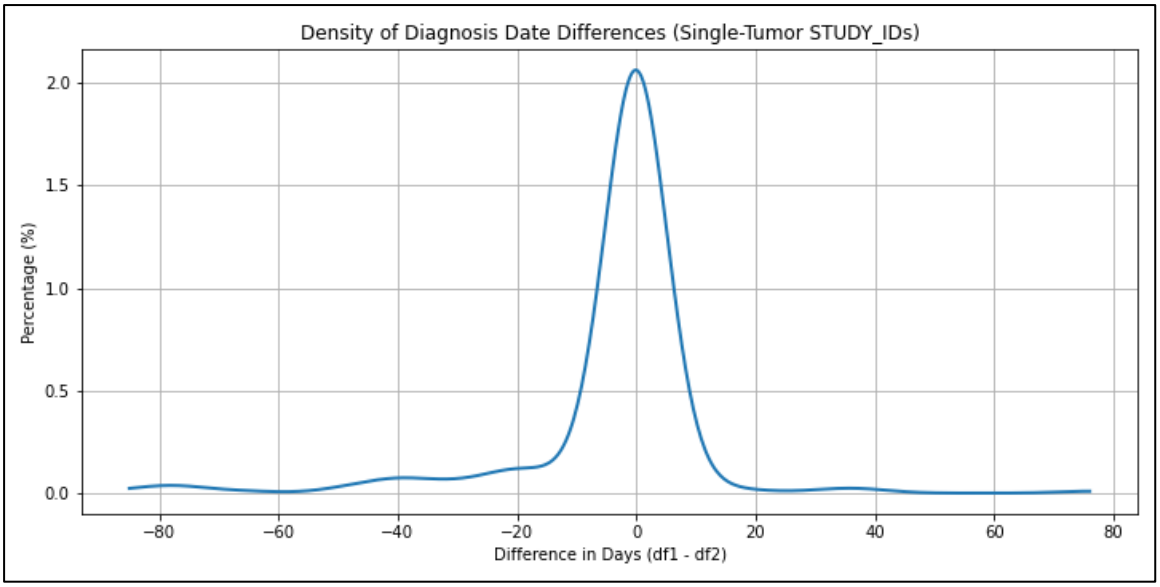

Figure 3: Registry vs Ovarian Path Report

- The same linking rule is applicable between **Legacy tumours** (existing cancer summary) and other data sources. This is only applicable in the first run
- Legacy data includes tumours from previous registry, GP confirmations, and other confirmed cancer sources
- **JSON schema** is used as the data dictionary for source and target dataset
- The schemas are used to validate the data, enforce data types, and makes it more readable to all users
- The target JSON schema (NewCancerSummary.json) contains **x-sourcePriority** field indicating the order in which data needs to be populated for every variable from source datasets
- All the source data variables are mapped to Registry standard
- The **conflict in ICD_CODE** and **MORPH_CODE** identified after linking the tumours are handled at the end of the process by always preferring the tumour from Registry.
- The **conflict in LATERALITY** identified between Registry and Histopathology breast data are handled using following rules:
  - If Registry laterality is '9' or 'M': replace with HistoPath laterality
  - If Registry laterality is 'B': split into multiple rows, one per HistoPath laterality (L/R)
  - Remove the matched HistoPath_BrCa rows after processing

- NOTE: Always refer the log file and go through every tab of Summary report generated after every run to understand the discrepancies and observations while during creation of cancer summary dataset
  - **Log file path** - \\epidemiology\EPIDEMIOLOGY\SHARED\CancerEpidem\BrBreakthrough

\DeliveryProcess\Logs\Generate_Cancer_Summary_Log_2025-12-10_14:00:00.log

- **Summary Report location** - https://github.com/UK-Generations-Study/Schema_and_Derivation_Utils/tree/main/CancerSummary

3. Input

- Sources: SQL Server
  - Databases – NOBACKUP, Mailing, Uploads
  - Tables – People, casummary_v1, HistoPath_BrCa_GS_v1, OvCa_HistoPath_II, FlaggingCancers
  - JSON schemas – JSON schema with all the source dataset variables their constraints and descriptions.
  - The target schema contains variables, descriptions and source priorities to populate data. (https://github.com/UK-Generations-Study/Schema_and_Derivation_Utils/tree/main/CancerSummary/json_schemas)
- Frequency: Once every 6 months.
- Preconditions: All the source data tables should be up to date in NOBCAKUP database.

4. Output

- Destinations: SQL Server
  - Database – Uploads
  - Table – NewCancerSummary_v3
  - JSON dataset for RDS rollout - \\rds\data\DGE\DUDGE\Shared\GENERATIONS\NewCancerSummary\CancerSummary_v20250123
  - The JSON file in RDS will be ending with the rolled out version in the format ‘v_yyyymmdd’
- Frequency: Once every 6 months.
- Use cases: Analysts & Staff Scientists for research, Operations team for requesting Pathology samples.

5. Dependencies

- Upstream dependencies: Pipeline to update the NOBACKUP database.
- Downstream dependencies: Pathology sample collection.

6. Instructions for Access & Execution

- Entire codebase, JSON schemas, and summary report is available in GitHub: https://github.com/UK-Generations-Study/Schema_and_Derivation_Utils/tree/main/CancerSummary
- Access: Permission to login to Safe Haven (sutepidemts01) with write and execute permission on the Uploads database in SQL Server
- Execution: Manual execution of python program as it is an ad-hoc update to the dataset.
    - Login to Safe Haven (sutepidemts01) server using Windows Authentication
    - Open Spyder IDE and open N:\CancerEpidem\BrBreakthrough\SHegde\Schema_and_Derivation_utils\CancerSummary\scripts\Generate_CaSummary.py
    - Run the program and no parameters/arguments required.
    - Refer to Conifg.py file for the data source mapping, paths required for the program, and data validation mappings
    - There are multiple modules supporting different functionalities of the process. They are found in N:\CancerEpidem\BrBreakthrough\SHegde\Schema_and_Derivation_utils\CancerSummary\scripts\
    - **Modules:**
        - DDL.sql: Query to create Cancer Summary table in SQL Server database
        - Generate_CaSummary: Master script to generate the cancer summary dataset
        - config: All the settings, paths, rules, and constant variables needed for the pipeline
        - Clean_and_Validate: To clean the source datasets and validate using the JSON schema. Includes data type conversion and renaming fields to match the schema

        - Map_and_Derive_Stage: Mapping all the source variables to Registry standard and derive Stage variable for Histopathology breast data. The rules for mapping are part of config module
        - Derive_Morph_Code: Rules for deriving ICD-10 MORPH_CODE for Histopathology breast data when not available
        - Get_Legacy_Tumours: Mapping and transforming Legacy dataset to work with new cancer summary flow. This will be required only for the first time
        - Link_Tumours: Multiple functions to get the linking rules and link tumour across sources. Also includes populating target dataset using JSON schema and generate an intermediate file with all the tumours from every source before merging
        - ICD_to_Site_Mapping: Mapping of ICD codes as per https://icd.who.int/browse10/2019/en#/II standards. Contains further grouping of ICD codes as per GLOBOCAN 2020 database https://pubmed.ncbi.nlm.nih.gov/33538338/
        - Handle_Exceptions: Functions to handle conflicts in tumour linking columns across sources
        - SummaryReports: Code to generate Summary Reports as an excel file
- Automation: NA

7. Monitoring

- Metrics: Log file and Summary reports
    - Need to monitor log file for schema validation errors
    - Need to monitor logs for intermediate failures
    - Summary reports explaining the findings from the dataset (Completeness, Data source contribution, count per cancer site)
- Alerts: Inform data management team in case of any errors or unexpected output.
- Dashboards: NA

8. Troubleshooting

- Common issues:

    - Invalid source and target data found during schema validation
    - Unsupported columns and data types
    - Program runtime failures
- Steps to resolve: Refer the log file detailing every step and try to identify the issue.
- Contacts: Data management team.