Deriving Initial Disease Episode from discrete diagnosis

Michael Gurley[1], Asieh Golozar[1], Robert Miller[1], Rimma Belenkaya[1], Shilpa Ratwani[1], Shanta Bethusamy[1], Andrew Williams[1], Joseph Sirintrapun[1], Christian Reich[1]
[1]OHDSI Oncology WG

Background

The OMOP Oncology Module extends the OMOP CDM and Standardized Vocabularies to support the comprehensive representation of cancer conditions, treatments, and disease abstraction required for addressing key research questions. One key component of the Oncology Module is the Episode model which enables representation and analysis of clinically relevant disease and treatment episodes and outcomes. The Episode model equips researchers with the foundation to research the dynamic of the disease and treatment. However, many Episodes are not typically recorded at the point of care, and therefore need to be inferred using a predefined and standardized algorithm. For example, the initial diagnosis of a cancer requires knowledge of the initial set of diagnostic measures, amongst them often a biopsy, in combination with a logic to determine the exact time point.

However, such inference algorithms need to be tested, compared, and validated. One of the conventions of the OMOP oncology extension is to represent a patient's 'date of initial diagnosis' by populating the EPISODE.episode_number column with the number '1'. The 'date of Initial diagnosis' time point is commonly used time point to indicate, "Date of initial diagnosis by a recognized medical practitioner for the tumor being reported whether clinically or microscopically confirmed.". [1] The 'date of initial diagnosis' can be distinguished from the 'date of initial clinical encounter' for a cancer diagnosis within a specific healthcare system.

Patients often receive care for diagnosis across multiple disconnected healthcare care systems and capturing 'date of initial diagnosis' within a single healthcare system often requires additional data collection efforts. Within many institutions, the tumor registry (TR) is an entity that fulfills this data collection effort for subpopulations of patients (often only patients receiving a first-course treatment at the institution). In the absence of TR data, ascertainment of the initial date of diagnosis from electronic health record (EHR) can be challenging and might require abstraction of such information from pathology reports or medical charts.

Here, we present results of initial assessment of concordance between 'date of initial diagnosis' captured within tumor registry and 'date of initial encounter' captured within EHR (including billing, problem list and encounter diagnoses) at specific institutions. We further assessed if

[1] NAACCR tumor registry file format documentation: http://datadictionary.naaccr.org/default.aspx?c=10&Version=21#390

this difference is impacted by the availability of information on biopsy (performed within or outside the institution) and other diagnostic procedures.

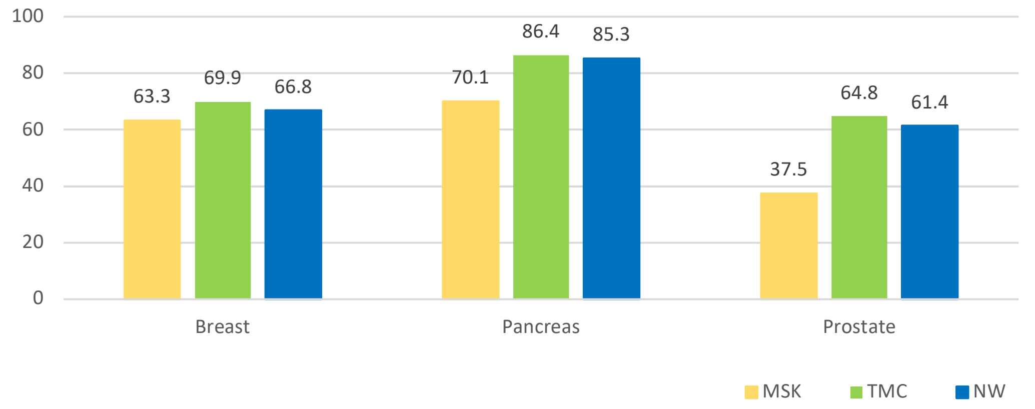
Methodology

Data for patients with prostate, pancreas, and breast cancer from three academic EHR data with linkage to the institutional TR, Northwestern University (NW), Memorial Sloan Kettering (MSK), Tufts Medical Center (TMC) were used in this study. Each patient population were identified using relevant vocabularies and definitions. Date of diagnosis from cancer registry and date of initial encounter in the EHR were retrieved and were used to quantify the difference between these two sources of information. Interval between closest biopsy (performed inside or outside institution) and closest pathology procedure to initial date EHR encounter and TR date of diagnosis were extracted.

Results

Data from 22,215 breast cancer, 2,071 pancreatic cancer and 12,369 prostate cancer patients with linked TR data from NW, MSK and TMC were included. The date of initial EHR encounter was within 30 days of the TR date of diagnosis for most patients (except for prostate cancer in MSK). However, we observed a substantial variability in the difference between TR and EHR in all databases and across tumor types (Fig 1 and 2). The difference between the two dates were less for pancreatic cancer, where majority of tumors are usually detected at advance stage and life expectancy is low.

Fig 1. % of tumors with EHR initial date of diagnosis within 30 days of tumor registry date of diagnosis



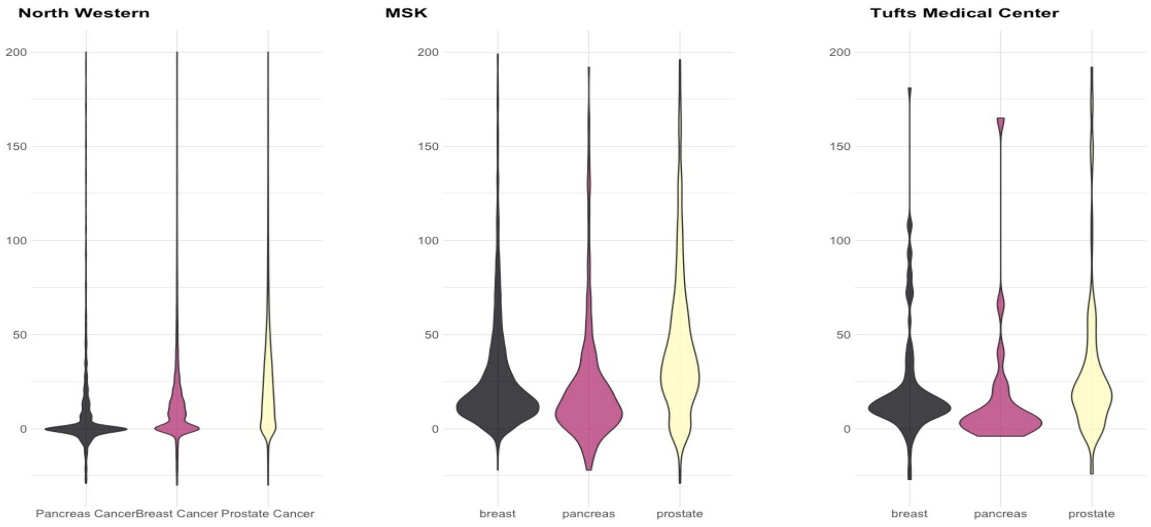


Fig 2. Difference in TR date of Diagnosis and EHR date of initial encounter across three tumor types

The difference between the two dates was less pronounced in patients who had a biopsy procedure performed within the institution across all tumor types. This figure was much variable in patients whose sole source of biopsy was from outside institutions (Fig 3)

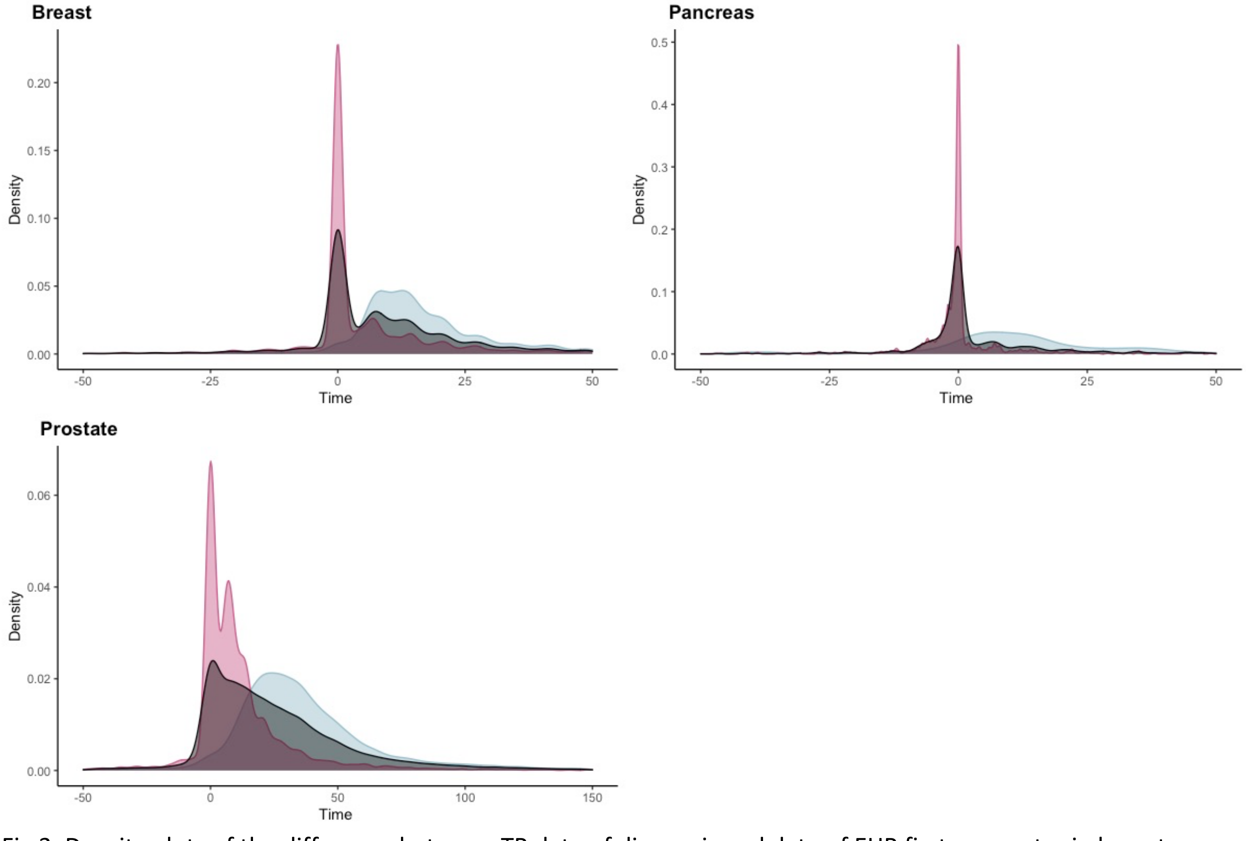


Fig 3. Density plots of the difference between TR date of diagnosis and date of EHR first encounter in breast, pancreas, and prostate cancer in NW in all patients (black), those with biopsy within the institutions (pink) and outside institution (blue).

Conclusion

The observed differences between the initial EHR encounter and TR date of diagnosis supports the need for more comprehensive approaches in identifying date of cancer diagnosis from EHR. This analysis supports amendment of the OMOP oncology extension documentation to counsel that implementers should not populate the EPISODE.episode_number field without high confidence that source systems track ‘date of initial diagnosis’ at a level equivalent to a tumor registry. Validated curation/NLP derivation effort using inside and outside pathology procedures, oncology progress notes, surgical oncology progress notes, radiology exam reports and radiation oncology progress notes might help identify the date of diagnosis.