Curation and Evaluation of an Open Cohort for Early Detection of Pancreatic Cancer

Zhiwei Liang[1], Qiaoli Wang[1], Eshna Sengupta[1], Lauren Brais[1], Elizabeth Andrews[1], Vidya Madineedi[1],
Brian Wolpin[1,3], Michael Rosenthal[1-3]
[1]Dana-Farber Cancer Institute, Boston, MA; [2]Brigham and Women's Hospital, Boston, MA
[3]Harvard Medical School, Boston, MA

Background: Cohort studies are essential to developing and validating biomarkers of population risk. However, many cohorts are closed with structured enrollment of patients and do not account for the unpredictable entry and exit patterns that are common in clinical care. Compared to closed cohorts, building dynamic open cohorts entails more challenges, such as data sparsity, population heterogeneity, and temporal inconsistency of data.

Objective: To construct and validate an open cohort to develop digital biomarkers for pancreatic cancer.

Data: Utilizing the MGB Research Patient Data Repository (RPDR), we selected all patients who had outpatient vital sign measurements, excluding those with less than six months of contact. This resulted in a cohort of 1.1 million patients, including approximately 1,500 identified cases. We retrieved patients' demographics and EMR based records such as lab results, diagnostic features, and ICD codes.

Case Identification: Identification of cases was hindered by the lack of a reliable structured identifier of pancreatic cancer in the medical record. Considering the historical scope of our cohort which extended back to the 1970s, the robustness of our clinical data sources was time-varying. We first assessed ICD codes for PDAC and found a 30% positive predictive value. We manually reviewed all cases with at least 5 PDAC ICD codes within a six-month window for definitive pathology for PDAC. Using this manually verified data, we calculated the case counts for each calendar year, which noted a significant decline prior to 2008. This limitation was primarily due to insufficient data to accurately identify cases in earlier periods. We also noticed a recent drop in patients and their visits post 2020, which was caused by delayed updates in data sources. The cohort was thus limited to the period from Jan 2008 to Dec 2020.

Measurement Completeness and Accuracy: We detected considerable data missingness in some fields, with weight serving as an example. 27% of visit occurrences lacked measured weights within six months, and 23% of patients had weights recorded within six months of less than 10% of their visits. After manual comparison of EMR contents to the research registry results, we identified significant gaps in delivered data as a likely cause of this missingness. We have worked with the information technology team to source data directly from the primary data store to overcome this challenge. We have also developed strategies to perform data imputation where required. Investigation into the causes of inaccuracies is a necessary but time-consuming aspect of curating an open cohort.

Cohort Evaluation: To validate our cohort, we compared our pancreatic cancer incidence rates grouped by gender, race, ethnicity, and age at diagnosis against (Surveillance, Epidemiology, and End Results) SEER Incidence Data. Standard errors, confidence intervals, and number of cases were compared as well. Specifically for the age at diagnosis, we calculated and compared both the crude rates for 19 age bins and the age-adjusted rates for 3 age bins (0-40, 40-85, 85+). The comparison indicated that our cohort incidence rates were plausible for future analysis.