Prediction of Pancreatic Cancer in a Population-Based Cohort Using Machine Learning

W. Chen W,[1] Y. Zhou,[1] F. Xie,[1] R.K. Butler,[1] C. Jeon,[2] E. T. Luong,[1] Lustigova,[1] B.U. Wu.[3]
[1]Kaiser Permanente Southern California Research and Evaluation, Pasadena, CA; [2]Oschin Comprehensive Cancer Institute, Cedars-Sinai Cancer Center, Los Angeles, CA; [3]Center for Pancreatic Care, Department of Gastroenterology, Los Angeles Medical Center, Southern California Permanente Medical Group, Los Angeles, CA.

Background: Pancreatic cancer (PC) is the 3rd leading cause of cancer deaths and the incidence and mortality rates are expected to increase in 2018-2040. Machine learning approaches based on electronic health records (EHR) have been attempted to predict PC, but most have focused on high risk populations (e.g. new onset of diabetes). We aimed to develop and validate a prediction model using high-dimensional clinical data extracted from the EHR of a large integrated health care system.

Methods: In this retrospective cohort study, health plan enrollees 50-84 years of age without a history of PC and at least one clinic-based visit (index visit) in 2008-2017 were identified. More than 500 potential predictors were extracted. Ten imputation datasets were generated to handle missing data. 'Random survival forests' were built to identify the most relevant predictors. Age was forced into the model. To avoid model over-fitting, 5-fold cross validation was conducted, yielding a total of 50 training and validation samples. Discrimination and calibration were evaluated based on c-index and Greenwood-Nam-D'Agostino calibration test, respectively.

Results: The cohort consisted of 1,801,931 patients (mean age 61.6 years). The estimated 18-month incidence rate was 0.83 (95% confidence interval 0.77-0.89) per 1000-PYs of follow-up. The three models containing age, abdominal pain, weight change and two biomarkers appeared most often (ALT change/HgA1c – 11 times, rate of ALT change/HgA1c – 11 times, and rate of ALT change/rate of HgA1c change – 9 times) out of the 50 training samples . The discrimination and calibration measures were comparable among the three models (c-index: mean 0.77 for all three models and SD 0.01-0.02; calibration test: p-value 0.19-0.45 and SD 0.17-0.54).

Conclusion: We developed and internally validated a PC risk prediction model which can be implemented to help providers evaluate the risk of PC at the time of a clinic visit in the general population.

Publish in Pancreas? No