A machine-learning algorithm to identify hepatitis C in health insurance claims data

Mohammed A. Khan[2, 1], Jae Eui Soh[2, 1], Matthew Maenner[2], William W. Thompson[2], Noele P. Nelson[2]

[1]Emory University, Atlanta, Georgia, United States, [2]Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Objective

We developed a machine learning-based algorithm to identify patients with chronic hepatitis C infection in health insurance claims data.

Introduction

Hepatitis C virus (HCV) infection is a leading cause of liver disease-related morbidity and mortality in the United States. Monitoring the burden of chronic HCV infection requires robust methods to identify patients with infection. Insurance claims data are a potentially rich source of information about disease burden, but often lack the laboratory results necessary to define chronic HCV infection. We developed a machine learning-based algorithm to identify patients with chronic HCV infection using health insurance claims alone and compared it a previously developed ICD-9 code-based algorithm.

Methods

We obtained insurance claims, demographics, enrollment information, and hepatitis C laboratory results from the IBM MarketScan® Commercial Claims and Encounters databases. We defined chronic HCV infection cases as a patient with one or more positive HCV RNA result and required controls to have a negative HCV antibody result and no positive HCV RNA or antibody results. Patients were required to be continuously enrolled in a health insurance plan during the six months before and after the first positive or negative test result (index date). Outpatient and inpatient insurance claims for the six months before and after the index date were included in the analyses. The study period spanned from 2011 to 2014.

Subjects were randomly divided into a training sample (80%) and test (20%) sample. We trained a random forest classifier using age, sex, region, Charlson comorbidity index, and variables defining the presence and frequency of 67 ICD-9 diagnosis codes and CPT procedure codes related to HCV and liver disease. We up-weighted cases to account for the low prevalence of infection in our sample. We generated forests of 1,000 trees for all models. The initial model included all variables. Permutation-based variable importance scores from this initial model were used to select variables for the final model. The previously developed algorithm defined chronic HCV infection as either two claims with codes for chronic hepatitis infection > 60 days apart after an HCV RNA test claim or three claims with codes for chronic HCV infection on different dates after an HCV RNA test claim. We compared the predicted classification to HCV laboratory result- defined classification and calculated percent agreement, Kappa, sensitivity, specificity, positive predictive value, and negative predictive value. We then applied the final classifier to all individuals continuously enrolled in commercial and/or Medicare supplemental insurance to estimate the prevalence of chronic HCV infection in this population in 2014. Analyses were performed in SAS version 9.4.

Results

We identified 5,780 (5.6%) cases with chronic HCV infection and 97,831 controls with negative HCV test results. The training dataset consisted of 82,888 individuals with approximately six million inpatient and outpatient claims. The final model included 23 variables related to hepatitis C (e.g., number of HCV RNA test claims), liver disease (e.g., cirrhosis diagnosis code), and comorbidities. In the training dataset, percent agreement, Kappa, sensitivity, specificity, positive predictive value, and negative predictive value were 99.2%, 0.92, 92.3%, 99.6%, 93.2%, and 99.5%, respectively. The presence of a CPT code for HCV RNA testing had the highest variable importance score. The test dataset included 20,723 individuals with approximately 1.5 million inpatient and outpatient claims. In the test dataset, percent agreement, Kappa, sensitivity, specificity, positive predictive value, and negative predictive value for the final classifier were 98.9%, 0.89, 89.9%, 99.4%, 89.0%, and 99.4%, respectively. Percent

ISDS Annual Conference Proceedings 2019. This is an Open Access article distributed under the terms of the Creative Commons AttributionNoncommercial 4.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

agreement, Kappa, sensitivity, specificity, positive predictive value, and negative predictive value for the previously developed algorithm were 96.3%, 0.50, 35.0%, 99.9%, 96.7%, 96.3%, respectively. Among the 35.6 million individuals with continuous commercial and/or Medicare supplemental insurance in 2014, 317,932 (0.9%) were classified as having chronic HCV infection.

Conclusions

Our machine learning-based algorithm was able to identify chronic hepatitis C cases in commercial health insurance claims data with relatively high estimates for percent agreement, Kappa, sensitivity, specificity, positive predictive value, and negative predictive. Future analyses and models will explore the ability of the algorithm to estimate the prevalence of HCV infection in different populations covered by different health plan types (e.g., commercial, Medicaid, Medicare, or no insurance) and for populations where laboratory testing data is not available or collected.

ISDS Annual Conference Proceedings 2019. This is an Open Access article distributed under the terms of the Creative Commons AttributionNoncommercial 4.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.