Development of a Machine-Learning Model to Predict Mortality and Its Cause Using the National Health Insurance Service National Sample Cohort

Chungsoo Kim, Pharm.D.[1], Seng Chan You, M.D. MS. [2], Rae Woong Park M.D., Ph.D[1, 2]
[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Republic of Korea;
[2]Department of Biomedical Informatics, Ajou University School of Medicine, Republic of Korea

Is this the first time you have submitted your work to be displayed at any OHDSI Symposium?

Yes _____ No__V__

Abstract

In observational and clinical studies, death is considered the most important clinical outcome, with cause-specific mortality used to assess the effect of a medical intervention on a disease. However, many observational databases do not contain full information on mortality and cause of death. In this study, we have developed two machine-learning models to predict different categories of death status in the premature end of patient observation, as follows: overall death, and death due to cardiovascular disease, cerebrovascular disease, pneumonia, diabetes, liver disease, chronic lower respiratory disease, or hypertensive disease. By combining the predicted values from these individual models, the random forest model was developed to predict the cause of death. The area under the receiver operating characteristic curve of the cause prediction model was 0.9192. This model can be applied to any database in the format of the Observational Medical Outcome Partnership Common Data Model provided by Observational Health Data Sciences and Informatics. The study package for the development of this model is available at https://github.com/ABMI/CauseSpecificMortality.

Introduction

Mortality is an important clinical outcome that is widely used in retrospective observational studies and clinical trials. All-cause mortality is less sensitive to each disease condition and is highly affected by underlying diseases[1]. Therefore, it is important to consider cause-specific mortality and data, for which accurate inclusion of the patient's cause of death is absolutely necessary. However, data on deaths and cause of death are decreasing globally, presenting a stumbling block to research that uses death data. J.M.Reps et al. developed a machine-learning model based on the patient-level prediction (PLP) packages of the Observational Health Data Sciences and Informatics (OHDSI). While the model showed good performance (area under the receiver operating characteristic curve [AUROC], 0.986)[2], the need for a cause of death prediction model has not been met. The purpose of this study was to develop a machine-learning model that would predict death and the cause of death using the National Health Insurance Service-National Sample Cohort (NHIS-NSC) and to make it interoperable using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

Methods

First, we created two stacked PLP models that predicted death, which made it possible to classify the causes of death. Based on the patient's last visit date to health care providers, the machine-learning models gathered data on the patient's previous visits and predicted the death occurring within a certain interval (30, 90, 180, or 365 days) following the last visit. In addition, the predictive value of each model

was used to classify ten different categories of death status: non-death, death due to cancer, cerebrovascular disease, cardiovascular disease, diabetes, chronic lower respiratory disease, liver disease, hypertensive disease, pneumonia, and other causes.

We set the target cohort as patients who had visited the same medical institution for more than 1 year. Data from the final year of the database were excluded so as avoid misinterpretation of data as patient survival. The outcome cohort was set up as an all-cause mortality group to confirm the presence of death and eight cause-specific mortality groups to identify the cause of death. The definition of all diseases was based on the International Classification of Diseases, 10th revision (ICD-10), and the code was converted to SNOMED-CT. Lasso logistic regression and gradient boosting machine were used to develop the death prediction model. For cause prediction model, applied the random forest algorithm. Evaluations were performed in comparison with patient death records. All analyses were performed using R version 3.5.2, and all source codes are available at https://github.com/ABMI/CauseSpecificMortality.

Results

The number of patients in the target cohort was 174,748. Of the 42,614 patients with death records, 30,878 (72.46%) died within 30 days of the last visit; 35,708 (83.79%) died within 60 days; 37,040 (86.92%) died within 90 days; 38,862 (91.20%) died within 180 days; and 40,649 (95.39%) died within 365 days. Regarding the cause of death, 12,454 (29.23%) of those with a death record died from cancer, 3,528 (8.28%) from cardiovascular disease, 3,427 (8.04%) from cerebrovascular disease, 967(2.27%) from pneumonia, 1,904(4.47%) from diabetes, 3,429(8.05%) from liver disease, 1,198 (2.81%) from chronic lower respiratory disease, and 834 (1.96%) from hypertensive diseases. The mean AUROC of the cause prediction model was 0.9192 and (Table 1) and the mean AUROC of the death prediction model was 0.9796. (Table 2)

Table 1. Performance values of random forest model predicting cause of death.

Days after the last visit (days)	30	60	90	180	365	Mean
AUROC	0.925	0.9222	0.9194	0.9154	0.9139	0.9192
Accuracy	0.943	0.9345	0.9352	0.9269	0.9248	0.9329

Conclusion

In this study, we have developed a model for predicting death and the cause of death using the NHIS-NSC. In the cause prediction model developed by the random forest algorithm, it was possible to classify the causes of death with a high performance of AUROC 0.9 or higher. Further research will need to develop a machine-learning model that can accurately predict the cause of death.

Acknowledgement

This work was supported by the Bio Industrial Strategic Technology Development Program (20005021) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number: HI16C0992]

Table 2. The area under the receiver operating curve of the death prediction models.

Days after last visit (days)	Lasso logistic regression model						Gradient Boosting Machine model					
	30	60	90	180	365	Mean	30	60	90	180	365	Mean
Any death	0.985	0.985	0.984	0.983	0.981	0.984	0.985	0.987	0.986	0.985	0.984	0.985
Cancer death	0.994	0.994	0.994	0.994	0.994	0.994	0.995	0.994	0.995	0.995	0.995	0.994
Cardiovascular death	0.971	0.967	0.965	0.958	0.962	0.964	0.967	0.959	0.960	0.972	0.970	0.965
Cerebrovascular death	0.978	0.979	0.980	0.976	0.978	0.978	0.985	0.982	0.98	0.979	0.994	0.981
Pneumonia related death	0.974	0.970	0.969	0.965	0.967	0.969	0.979	0.967	0.972	0.970	0.966	0.970
Diabetes related death	0.981	0.983	0.983	0.978	0.978	0.981	0.984	0.984	0.983	0.985	0.984	0.982
Liver disease related death	0.990	0.999	0.990	0.985	0.989	0.991	0.989	0.988	0.993	0.993	0.99	0.990
Chronic lower respiratory disease related death	0.986	0.986	0.983	0.984	0.985	0.985	0.987	0.989	0.985	0.985	0.986	0.985
Hypertensive disease related death	0.965	0.952	0.959	0.963	0.959	0.959	0.956	0.957	0.961	0.960	0.963	0.959

References

1. Sasieni Peter D, Wald Nicholas J. Should a Reduction in All-Cause Mortality Be the Goal When Assessing Preventive Medical Therapies? Circulation. 2017;135(21):1985-7.
2. Reps JM, Rijnbeek PR, Ryan PB. Identifying the DEAD: Development and Validation of a Patient-Level Model to Predict Death Status in Population-Level Claims Data. Drug Safety. 2019.