Development of phenotype algorithms for common acute conditions using SHapley Additive exPlanation values

Konan Hara, MD, PhD[1,2], Ryoya Yoshihara[3,2], Tomohiro Sonoo, MD[4,2], Toru Shirakawa, MD[5,2], Tadahiro Goto, MD, MPH, PhD[2,6], and Kensuke Nakamura, MD, PhD[4]

[1] Department of Public Health, The University of Tokyo, [2] TXP Medical Co. Ltd., [3] Faculty of Medicine, The University of Tokyo, [4] Department of Emergency and Critical Care Medicine, Hitachi General Hospital, [5] Department of Social Medicine, Osaka University Graduate School of Medicine, [6] Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo

Background. Phenotype algorithms facilitate the use of electronic medical records (EMRs) in translational research. However, the development of interpretable phenotype algorithms that are easily reproducible by subsequent researchers is still a challenge. In this work, therefore, we address this problem by introducing an intuitive and reasonable feature selection method based on SHapley Additive exPlanation (SHAP) values. Furthermore, this method is evaluated by developing phenotype algorithms that can identify patients with eleven common acute conditions. The application of this method does not require subject-matter knowledge of the corresponding conditions to filter candidate features that crucially relate to the target phenotype.

Methods. EMRs and medical claims data were obtained from a tertiary hospital in Japan, where medical charts have been structured and stored through the Next Stage ER system (TXP Medical Co. Ltd., Tokyo, Japan). To perform this feature, the system, in contrast to conventional medical charts, provide prespecified forms, such as vital signs, chief complaints, past medical history, physical assessments, and clinical diagnoses, to be filled by the physicians. As a result, the inputted clinical information are saved as reliable structured data and, therefore, researchers can access them more efficiently than when natural language processing techniques are applied to unstructured EMRs to retrospectively extract structured information. The data were collected from April 01, 2018, to March 31, 2019, and were composed of information related to 5459 patients transported by ambulance to the emergency department. To develop a phenotype algorithm construction procedure that can be applied to all phenotypes, we created a dataset containing features that were chosen regardless of the phenotypes considered in this study. This approach resulted in 3902 features from EMRs (related to patients demographics, consultation information, vital signs, chief complaints, past medical history, physical assessments, clinical diagnoses, and laboratory test results) and claims (diagnostic codes, medication codes, and procedure codes). The target phenotypes and their corresponding prevalence in the cohort group of the study were acute myocardial infarction (AMI; 2.2%), angina (1.1%), clinical scenario 1 acute decompensated heart failure (CS1-ADHF; 2.1%), CS2-ADHF (1.7%), bacterial pneumonia (3.1%), aspiration pneumonia (1.7%), cerebral infarction (3.4%), intracerebral or subarachnoid hemorrhage (2.5%), fractures due to traffic injury (1.1%), fractures from falling (1.8%), and hip fracture (1.4%). Two physicians independently reviewed the EMRs to diagnose these phenotypes, and one of the authors of this work, who is an emergency physician, resolved the discrepancies of their diagnosis. For each phenotype, the cohort data was split into training (75%) and test (25%) sets, in which stratified randomization was applied on the gold-standard diagnosis. We, thereafter, developed phenotype algorithms to predict scores related to the propensity of having the target phenotype. To predict each phenotype, we trained an XGBoost model and calculated the SHAP and SHAP interaction values. Furthermore, to develop an interpretable and easily reproducible model, we applied the following procedure: (1) for k = 1, 2, …, the logistic regression model was applied to the k features with the highest global impact of SHAP (model 1) or SHAP interaction (model 2), and the area under the receiver operating characteristics curve (AUROC) using a 5-fold cross-validation technique was calculated; (2) k was increased by one and step 1 was repeated; (3) the procedure was stopped if AUROC calculated in step 1 did not exceed the maximum value of the precedent AUROCs for two consecutive iterations, and k that maximized the AUROC was selected. To evaluate the robustness and detect potential weaknesses of the proposed method, we conducted the procedure five times without specifying the seeds for the random number generators involved in it. Performance on the test set of models 1 and 2 was evaluated using AUROC and compared to that of the XGBoost model.

Results. The AUROCs of the XGBoost model, model 1, and model 2, were, respectively: 0.968, 0.976, and.968 for AMI; 0.958, 0.934, and 0.931 for angina; 0.989, 0.986, and 0.972 for CS1-ADHF; 0.991, 0.982, and 0.983 for CS2-ADHF; 0.975, 0.970, and 0.975 for bacterial pneumonia; 0.990, 0.986, and 0.987 for aspiration pneumonia; 0.983, 0.976, and 0.985 for cerebral infarction; 0.968, 0.968, and 0.968 for intracerebral or subarachnoid hemorrhage; 0.955, 0.950, and 0.931 for fractures due to traffic injury; 0.986, 0.943, and 0.910 for fractures from falling; 0.998, 0.998, and 0.998 for hip fracture. Based on the global impact of the SHAP value, the median of the number of features selected for each phenotype in the five trials was: 2 for CS2-ADHF, intracerebral or subarachnoid hemorrhage, and hip fracture; 3 for aspiration pneumonia; 4 for CS1-ADHF and cerebral infarction; 6 for AMI and angina; 8 for bacterial pneumonia; 10 for fractures due to traffic injury; 11 for fractures from falling. The interaction terms did not appreciably contribute to the performance of the phenotype algorithms. For CS1-ADHF, CS2-ADHF, bacterial pneumonia, cerebral infarction, intracerebral or subarachnoid hemorrhage, and hip fracture, models 1 and 2 stably achieved similar AUROCs values compared with the XGBoost model for the five trials.

Conclusion. With the proposed intuitive and reasonable feature selection methods, highly interpretable and easily reproducible phenotype algorithms that exhibit satisfactory performance were created for six of the eleven phenotypes. However, a fine-tuned model was difficult to obtain by the proposed methods when few positive cases of the phenotype were available, or the set of features related to the phenotype was large. Experts may want to consider removing unnecessary features by juxtaposing the SHAP summary plot and the coefficients of the logistic regression.