Machine Learning for Healthcare 2020 – Clinical Abstract, Software, and Demo Track

The use of natural language processing to improve identification of patients with peripheral artery disease

E. Hope Weissler, MD;[1] Jikai Zhang, BS;[2] Steven Lippmann, PhD;[3] Shelley Rusincovitch, MMCi;[2] Ricardo Henao, PhD;[2] W. Schuyler Jones,MD[3,4]

1. Division of Vascular Surgery, Department of Surgery, Duke University Medical School, Durham, NC; 2. Duke Forge, Duke University Medical School, Durham, NC; 3. Department of Population Health Sciences, Duke University Medical School, Durham, NC; 4. Division of Cardiology, Department of Medicine, Duke University Medical School, Durham, NC

Background: Peripheral artery disease (PAD) is estimated to affect between 8–12 million Americans. It leads to decreased blood flow to the legs; when severe, it causes constant pain and tissue loss due to ischemia. Regardless of severity, PAD is associated with increased risks of amputations, heart attacks, hospitalizations, and death. PAD is under-recognized, under-treated, and under-studied,[1] and improved PAD diagnosis, care, and investigation require efficient and accurate identification of patients with PAD. Current identification methods rely on diagnosis codes, procedure codes, or lists of patients diagnosed and/or treated by specific clinicians in specific locations and/or ways. Each of these methods is problematic because clinical data systems are not primarily designed to facilitate identification of broad patient phenotypes. Diagnosis codes perform poorly for PAD identification and use of procedure codes or patient lists from specific clinical contexts limits identification to a restricted subgroup of PAD patients.[2] The aim of this project was to apply natural language processing (NLP) to the unstructured clinical notes created by healthcare providers, such as patient discharge summaries. We expected that this approach would more efficiently and accurately identify patients with PAD in the Duke University Health System (DUHS) compared with a diagnosis code-based approach.

Methods: The baseline patient cohort was derived from the DUHS electronic health record system for patients with encounters associated with PAD diagnosis codes between January 1, 2015 and March 31, 2016. This yielded 15,406 patients; among a subset of 2,500 patients randomly selected for PAD status confirmation by a clinician, only 32.6% actually had PAD. We first used this subgroup of 2,500 patients to construct a diagnosis code-based model to predict the presence of PAD using a Least Absolute Shrinkage and Selection Operator (LASSO) approach. This algorithm was then applied to the remaining 12,801 patients. Charts of patients with a 45% or greater predicted probability of PAD were clinically adjudicated in order to classify PAD ground truth using four criteria: ankle-brachial indices indicative of PAD, imaging documentation of significant atherosclerotic stenosis or occlusion, prior peripheral revascularization, or prior amputation due to symptomatic PAD. The clinical narratives of 80% of the adjudicated cohort of 6,865 patients served as the training set for our NLP approach, using a Label-Embedding Attentive Model (LEAM). The LEAM framework was modified to account for numerous notes per patient with varying relevance to the categorization task (therefore, a hierarchical-LEAM, called PAD-ML). To evaluate PAD-ML, we report areas under receiver operating characteristic (AUC) and precision recall curves. with the standard deviations over 10-fold cross validation. To compare model performance, we apply the DeLong test on the median-AUC fold for the LEAM model and the same fold for the LASSO model. To evaluate the model's clinical interpretability, we review words identified as highly relevant by PAD-ML.

Results: Patients had a median of 16 clinical notes each (interquartile range (IQR) 6-39) with a median of 419 words (IQR 239-697) per note. The median (standard deviation) of the area under the ROC curve for the hierarchical-LEAM was 0.888 (0.009) versus 0.801 (0.017) for the original LASSO approach (DeLong p value <0.0001). The median (standard deviation) of the area under the PR curve was 0.909 (0.008) versus 0.816 (0.012) for the original LASSO approach. The top ten words most strongly associated with a prediction of PAD were: atherectomy, aortofemoral, aorto, rsfa, anthropometrics, msfa, aortogram, stenting, bifemoral, and iliacs.

Conclusion: Using an NLP approach to clinical notes, we were able to meaningfully improve our ability to identify patients with PAD compared to an algorithm consisting of diagnosis codes and administrative flags. Because PAD-ML was trained on a full-spectrum PAD cohort of patients from across DUHS regardless care context, it is also able to identify the full range of PAD patients, greatly increasing its clinical and investigational utility. We plan to use PAD-ML to

[1] Hirsch AT, Criqui MH, Treat-Jacobson D, et al. Peripheral arterial disease detection, awareness, and treatment in primary care. JAMA.2001; 286:1317–1324.

[2] Hong Y, Sebastianski M, Makowsky M, Tsuyuki R, McMurtry MS. Administrative data are not sensitive for the detection of peripheral artery disease in the community. Vasc Med. 2016;21:331–336.

identify PAD patients for care-improvement interventions within DUHS, as well as for research collaborations with other institutions.

References.