CNS
2018 ANNUAL MEETING
HOUSTON, TEXAS
OCTOBER 6-10, 2018

Using Natural Language Processing in Outcomes Modeling: A Text-Based Model Predicts Non-Home Discharge After Meningioma Surgery with Greater Discrimination Than a Model with 57 Discretely Collected Clinical Variables

Whitney Muhlestein; Lola Blackwell Chambless MD

Introduction

Building databases for predictive modeling is time and resource-intensive. Here, we use natural language processing (NLP) and machine learning (ML) to predict non-home discharge after craniotomy for meningioma from the text of preoperative notes and radiology reports. The NLP model outperforms a model built from 57 variables thought to be clinically relevant.

Methods

We built a database of 597 patients surgically treated for meningioma at our institution between 1995 and 2015. Age, sex, and number surgery, and text from the preoperative note and radiology report was collected for each patient. 57 total preoperative variables were collected in a second database. Text was represented via transverse frequency-inverse document frequency and used to create a linear model to predict operative time from the text alone, with the predictions acting as a feature in final model training, or used as features directly. 32 ML algorithms were trained to predict non-home discharge and the top performing algorithms combined to form an ensemble model. Area under the curve (AUC) was calculated for the NLP and 57-variable ensembles to compare discriminative ability, and word clouds generated to visualize which words best predict non-home discharge.

Results

The NLP model predicted non-home discharge with an AUC of 0.80 and 0.76 on internal and external validation, while the 57-variable model had an AUC of 0.77 and 0.74. Text in the preoperative note that predict non-home discharge include: “progressive,” “large,” and “decline.” Text in the preoperative radiology report that predict with non-home discharge include: “large,” “edema,” and “effacement.”

Conclusions

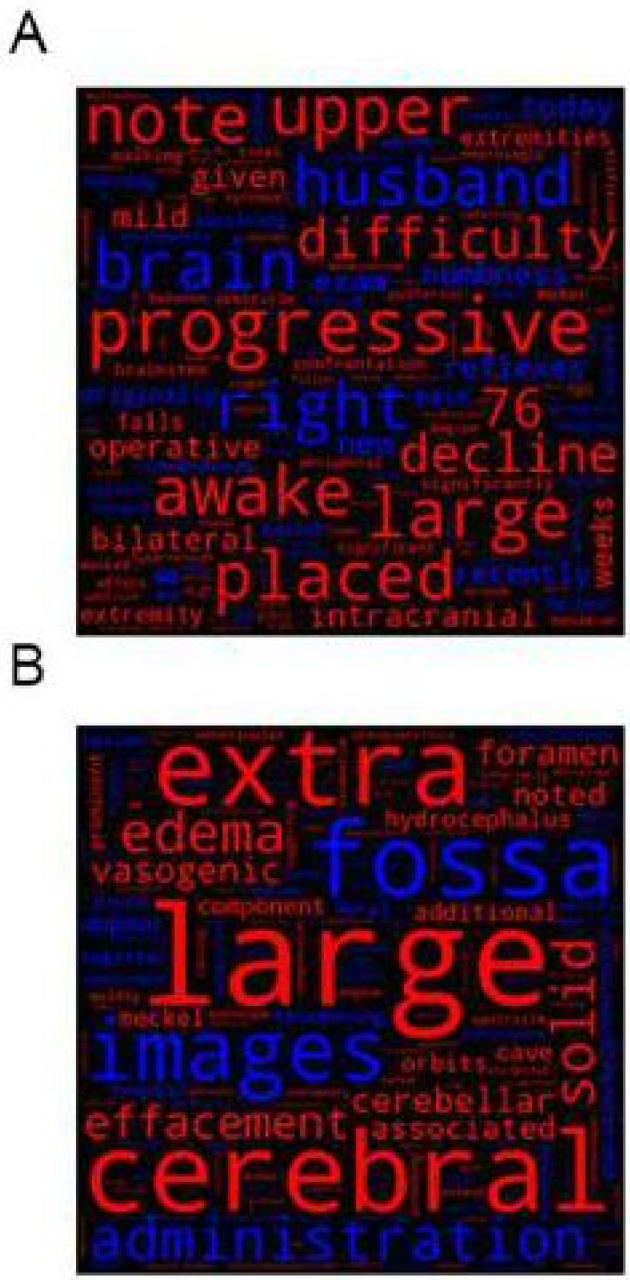
A NLP model outperforms a model requiring the collection of 57 discreet variables. Using free text as a primary data source for outcomes modeling may allow researchers to build predictive models more efficiently and at less cost, while also reducing bias inherent to databases with pre-selected variables.

Learning Objectives

By the conclusion of the session, participants should be able to:

(1) Demonstrate how natural language processing can be used to harness underutilized data sources in the EMR.
(2) Recognize that unstructured data sources, including free text, can be used in outcomes modeling to improve predictions.

[Default Poster]



Word clouds demonstrating the relative importance of specific words and phrases to the NLP model from the (A) preoperative note and (B) preoperative radiology report. Red font color denotes words and phrases associated with non-home discharge, while blue font color denotes association with home discharge. Font size is proportional to how influential the word or phrase is in either direction (e.g. a large, red word is highly associated with non-home discharge; a small blue phrase is associated with home discharge, though not as strongly).