SMARTPATH: Knowledge Extraction of Pathology Reports for Large Scale Labeling of Breast Cancer images with Natural Language Processing

Maryam Panahiazar, PhD, University of California San Francisco; Dmytro S. Lituiev, PhD; Hari M. Trivedi, MD; April S. Liang; Dexter Hadley, MD, PhD

Introduction
Breast cancer is a leading cause of cancer death among women in the United States (Patel, 2017) Screening mammography has proven to be an effective tool for reducing breast cancer mortality by allowing early detection of suspicious findings such as masses, abnormal calcifications, architectural distortion, or asymmetries (Patel, 2017). Despite deep learning's numerous successes in the biomedical field, a large rate-limiting step for its application is the paucity of large-scale, freely-accessible, and robustly-annotated data (Litjens2016) Albarqouni2016) (Bekker2016) which is required for deep learning for image recognition tasks. Our aim was to develop a technique to automatically label greater than 10,000 pathology reports with laterality and final pathologic diagnosis to serve as ground truth for annotating mammograms. An ideal classifier would accurately identify the pathologic outcome from the report, function across a variety of pathology report types (surgical, core biopsy, fine needle aspirations) and reporting pathologists, and perform without limitations to report length.

Hypothesis
An ideal classifier would accurately identify the pathologic outcome from the report, function across a variety of pathology report types (surgical, core biopsy, fine needle aspirations) and reporting pathologists, and perform without limitations to report length.

Methods
We present a semi-automated framework for the extraction of labels from routine clinical records using natural language processing (NLP) techniques and classifiers. We evaluate the performance of traditional NLP and machine learning techniques. 33,342 unique breast pathology reports from 1997-2014 were extracted from an in-house pathology database following approval from the institutional review board. Of these, 10,420 reports were found to be relevant to mammograms within one year. 3,099 were then randomly selected and manually annotated with the following ground truth labels: left positive, right positive, bilateral positive, or negative.

We converted attributes into a set of attributes representing word occurrence information from the text contained in the pathology report. The N-gram tokenizer (n=1 to 3), which breaks text into separate words as well as combinations of consecutive pairs and triples of words., further we applied both TF and IDF normalization along the entire corpus. This preprocessing step both frequently used tokens (TF) and down-weights abundant words that carry no information (such as ‘the’, ‘a’, and ‘or’).

We evaluate several classical machine learning methods, such as Decision Table, PART (rule mining algorithm), SVM, Naive Bayes, and Logistic Regression. Ultimately, logistic regression outperformed the other techniques and it was characterized. The sample was divided for 10-fold cross-validation according to standard practices.

Results
We sought to understand how each of algorithms performed for each of the four selected labels. For all three labels including negative, left positive, right positive logistic regression NLC Outperform all classifiers. With the F-measure of %98 for negative, 93% for right positive and %93 for left positive For Bilateral traditional performed low which indicate the future direction of the research to overcome this limitation. But considering that the data is not equally distributed and the number of bilateral is not that many in the big data set give an open discussion for possibility of ignoring a few bilateral options.
Conclusion
The goal of this study was to develop a framework capable of accurately extracting histopathologic diagnosis and laterality from narrative pathology reports for labeling mammographic images for deep learning. We describe a labeling structure for ground truth and discuss the performance of our framework utilizing natural language. Our work suggests that pathology reports caring valuable information to extract the label for annotation the mammogram images. Our work indicates that an integrative approach of text mining and machine learning can provide a reliable tool using pathology reports for labeling mammography images. As future work is going to propose a hybrid platform to overcome the limitation for the less frequent label such as Bilateral. We will use this labeling mechanism to prepare training data for our image analysis deep learning algorithm.

Statement of Impact
Our work has implications for semi-automatic of extracting labels, which geared towards available labeled data for to use deep learning in clinical research.

Keywords
natural language processing, classifier, breast cancer, pathology report