# Improving the Workflow of Curbside Consultation Using Unstructured Clinical Notes - A Natural Language and Machine Learning-based Approach

Wei-Hung Weng, MD[1,2*], Avni M. Khatri, BS[2], Kavishwar B. Wagholikar, MBBS, PhD[2,3], Adam B. Cohen, MD[4], and Henry C. Chueh, MD[2,3]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[2] Laboratory of Computer Science, [3]Department of Medicine, [4]Department of Neurology, Massachusetts General Hospital, Boston, MA

## Abstract

Inappropriate clinical referral leads to delayed disease diagnosis and management. To resolve the issue of inadequate referral, performing predictive analytics to identify medical specialties and experts using narrative clinical documents, which contain the most detailed but hidden information of the current clinical setting, can be an automatic approach to curate the knowledge of specialty and also expert profiles for the adequate referral. We develop a natural language processing, medical ontology and supervised learning algorithm integrated pipeline to categorize the clinical documents by the author and their medical specialty, and investigate the capability of the automatic medical specialties, medical experts tagging and prediction. The results show that machine can identify the medical specialty related to the content of document in a publicly available clinical notes from iDASH repository, with the best prediction accuracy of 0.899, precision of 0.901, recall of 0.855 and F1 measurement of 0.865 by the appropriate and optimized medical concept extraction method and supervised learning algorithm, which is bag-of-words representation with support vector machine with linear kernel. In the dataset from Massachusetts General Hospital (MGH), the best accuracy of 0.916, precision of 0.904, recall of 0.878 and F1 measurement of 0.890 were obtained with bag-of-words representation and regularized logistic regression. However, we also find that using concept-derived representation can reduce the size of feature space significantly but maintain the comparable prediction performance. In the future, the developed method and tools may augment the existing recommendation system for clinical consultation, and help the patients with undiagnosed diseases find the appropriate specialty and expert for their specific clinical condition.

## Background

### Motivation

- Sometimes clinicians encounter patients' clinical problems beyond their domain expertise.
- These clinical problems and dilemmas may leave unanswered, and result in misdiagnosis, delayed clinical care, and even lead to inappropriate treatment and management.

### Current Solution

- Clinicians usually refer the patients through clinical referral to relevant medical specialties and domain experts they are familiar with.
- Not easy to find the relevant domain expert for the clinical referral because clinicians can hardly find a correct clinical expert or even clinical specialty without a detailed understanding of both the domain experts' profiles and preferences, and also the key issues inside the raised clinical questions that usually noticed by only specialists.
- Patients may be referred to the wrong experts, and still resulting in delayed or inappropriate care.
- The curbside consultations can be inappropriate, time-consuming, and even unsafe for patients.

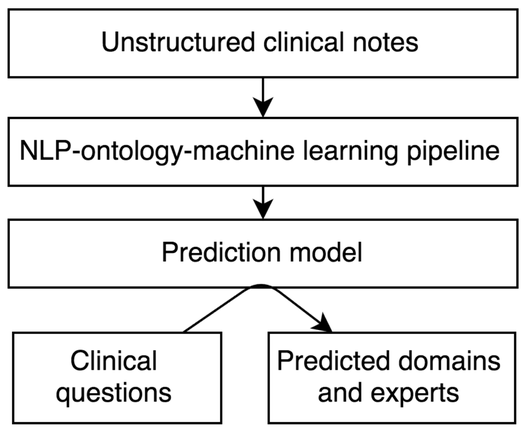
### Proposed Approach

- Expert recommendation systems (ERS) with natural language processing (NLP) and machine learning (ML) components [1].

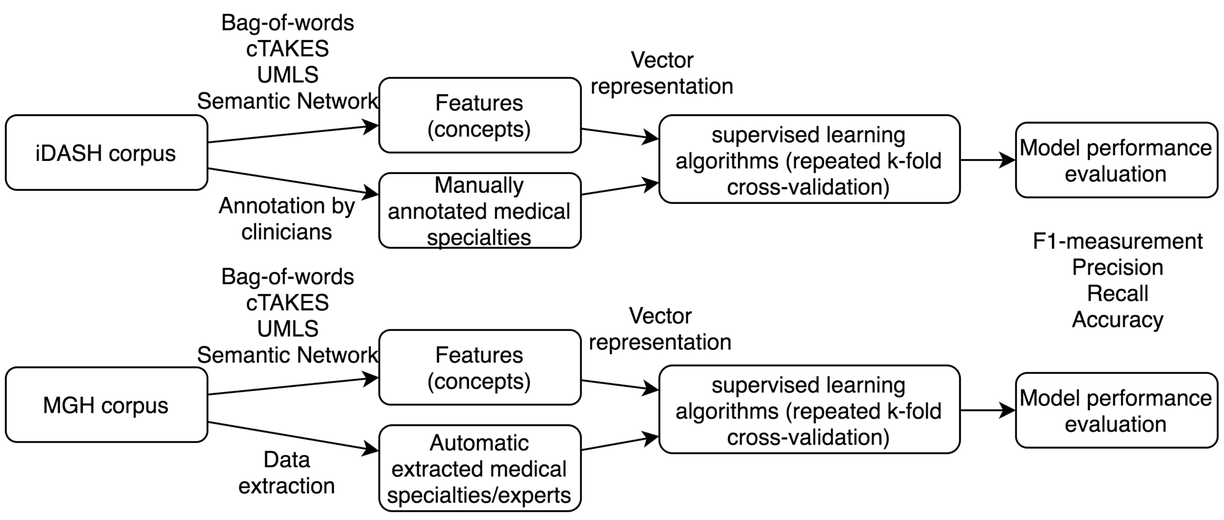
### Research Objective and Aims

- To use predictive analytics as a potential solution to predict suitable medical specialties and experts in order to solve the specific clinical dilemmas appropriately
  1. Use unstructured clinical documents
  2. Build a NLP-ML software to do modeling and prediction automatically
  3. Identify the suitable medical specialties and experts automatically by given clinical notes

## Materials and Methods




| Specialty | Number of Notes (iDASH) | Number of Notes (MGH) |
|---|---|---|
| ALL | 431 | 61276 |
| CARDIOLOGY | 116 | 20928 |
| ENDOCRINOLOGY | - | 12395 |
| NEUROLOGY | 97 | 10974 |
| PEDIATRICS | - | 4790 |
| GENERAL_SURGERY | - | 4388 |
| DERMATOLOGY | - | 4067 |
| PSYCHIATRY | 30 | 3734 |
| GASTROENTEROLOGY | 110 | 3188 |
| ORTHOPEDIC_SURGERY | - | 3053 |
| GERIATRIC_MEDICINE | - | 2092 |
| NEPHROLOGY | 22 | 1936 |
| PULMONARY_DISEASE | 56 | 763 |



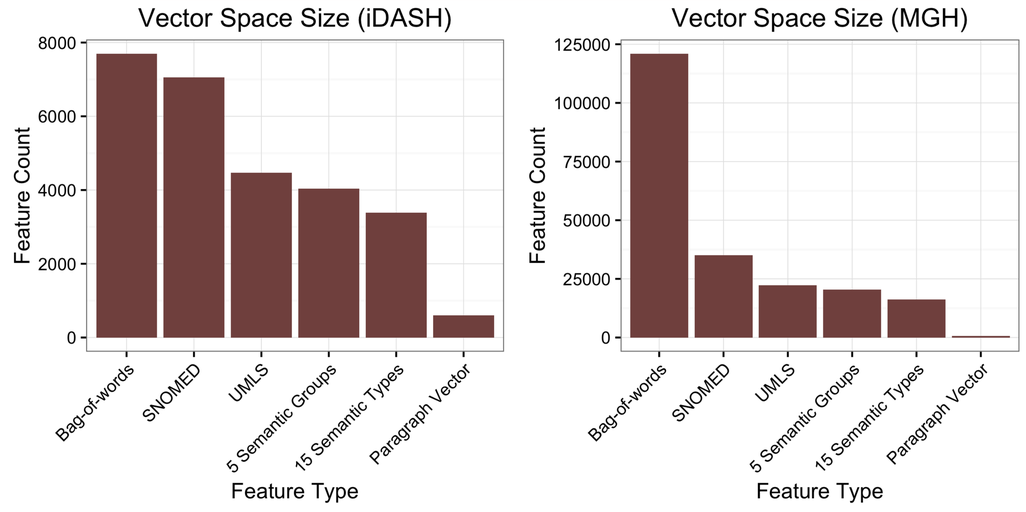

### NLP and Machine Learning Approach

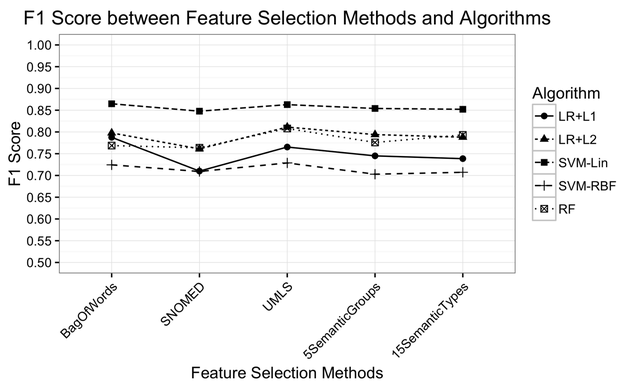
- We took the advantages from NLP, medical ontology, machine learning with the semantic approach using UMLS Metathesaurus and Semantic Network [2-4]
- Feature - Bag-of-Words / SNOMED concepts / UMLS concepts / Semantic groups and types
- Vector representation - Frequency count / tf-idf / paragraph vector [5]
- Algorithm - regularized logistic regression, SVM, random forest
- Metrics - Accuracy, precision, recall, F1 score

### Data Source

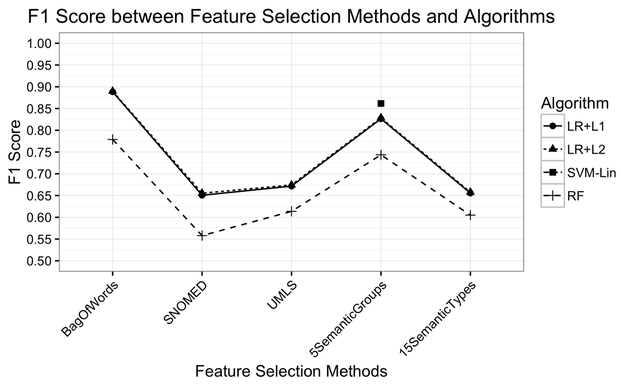
- 431 clinical documents from the iDASH repository [6], annotated the specialties manually by two clinicians
- 61276 MGH clinical notes from Partners HealthCare RPDR (IRB:2016P000011)

## Result

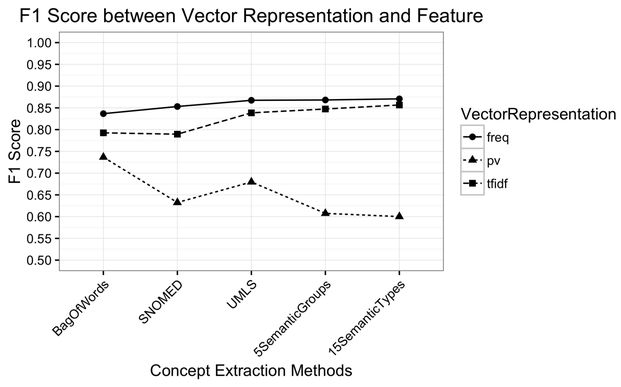


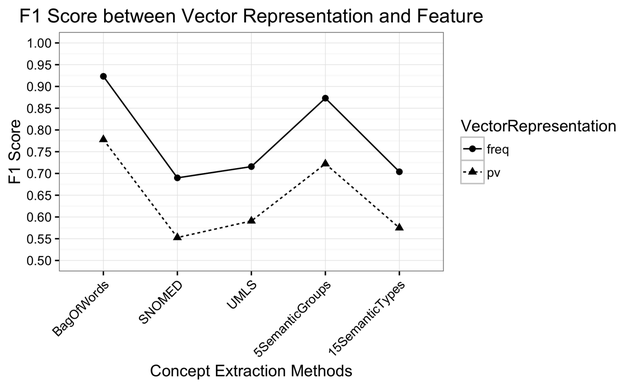

F1 Score between Feature Selection Methods and Algorithms




F1 Score between Vector Representation and Feature




F1 Score between Vector Representation and Feature




### Vector Space Reduction

- Using UMLS-derived or semantic restriction can reduce the feature space >40% in iDASH dataset, and >75% in MGH dataset

### Replicability

- Replicable in iDASH and MGH dataset
- More data / features usually yield better performance

### Model Performance

| | iDASH | MGH |
|---|---|---|
| Accuracy | 0.899 | 0.916 |
| Precision | 0.901 | 0.904 |
| Recall | 0.855 | 0.876 |
| F1 Score | 0.865 | 0.890 |

The best model performance in two dataset

- Algorithm selection plays important role. SVM with linear kernel and regularized logistic regression performed better
- Clinical feature selection method has a higher impact than learning algorithm selection in larger dataset
- Specific semantic groups-derived concepts has compatible performance with bag-of-words, even with the great reduction of feature space
- cTAKES parsed UMLS-derived concepts have better performance than SNOMED-derived concepts
- Frequency count matrix yields better performance than paragraph vectors

### Next Steps

- Prediction
- Portability
- More features (n-grams)
- Feature combination
- Preserving sequential information (e.g. recurrent neural network)

## Conclusion

To solve the clinical problem of triaging the patients to the correct specialty and expert, we proposed an ERS with NLP-ML approach using clinical documents. We have developed an NLP-ML pipeline for clinical document classification, and applied it on tagging specialty of clinical documents. In the iDASH and MGH datasets, our document specialty tagging model reached the performance of accuracy, precision, recall and F1 score around 0.9 with bag-of-words and regularized logistic regression approach. We have also found that the UMLS and Semantic Network-derived approach yields comparable performance even if a huge reduction of feature space. Our method and tools may augment the existing recommendation system for future clinical curbside and formal consultation.

## Reference

[1] D'Avolio LW, Nguyen TM, Farwell WR, Chen Y, Fitzmeyer F, Harris OM, et al. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). Journal of the American Medical Informatics Association. 2010 Jul 1;17(4):375–82.
[2] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 Sep;17(5):507–13.
[3] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004 Jan 1;32(90001):267D–270.
[4] McCray AT. An Upper-Level Ontology for the Biomedical Domain. Comparative and Functional Genomics. 2003;4(1):80–4.
[5] Le, QV and Mikolov, T. Distributed Representations of Sentences and Documents. Paper presented at the meeting of the ICML, 2014.
[6] The iDASH dataset used in this project was downloaded from iDASH repository (https://idash-data.ucsd.edu) supported by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54HL108460.