

Information Retrieval in Clinical Chart Reviews

By

Cheng Ye

Dissertation

Submitted to the Faculty of the  
Graduate School of Vanderbilt University  
in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

May 10, 2019

Nashville, Tennessee

Approved:

Daniel Fabbri, Ph.D.

Bradley Malin, Ph.D.

Maithilee Kunda, Ph.D.

Yevgeniy Vorobeychik, Ph.D.

You Chen, Ph.D.

To my beloved wife, Ye Hong, my essential and unique source of inspiration and strength during my research life.

## ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest thanks and love to my wife, Ye Hong, for her unique, invaluable support of my life and my research since the first day we met. Without her encouragement and love, I would not make it in my research journey.

I want to express my most profound appreciation and gratitude to my academic advisor Dr. Daniel Fabbri, who brought me a life-changing experience to both my research and life. Dr. Daniel Fabbri not only directed me into an exciting research area but gave me enormous freedom in doing research. Without his invaluable guidance and generous support, I would have never been proud of myself, my research and my life.

I would also like to give sincere thanks to my committee members, Dr. Bradley Malin, Dr. Eugene Vorobeychik, Dr. You Chen and Dr. Maithilee Kunda, for their constructive advice to help improve the work in this dissertation.

I want to thank Joseph Coco for his excellent work in conducting crowdsourced chart reviews and his tremendous help in processing the result of chart reviews, one of the primary sources of the evaluation data sets of my dissertation.

Most importantly, I am particularly grateful to my family and all my friends at Vanderbilt University and Nashville, for their consistent support during my Ph.D. life.

## TABLE OF CONTENTS

	Page
DEDICATION . . . . .	ii
ACKNOWLEDGMENTS . . . . .	iii
LIST OF TABLES . . . . .	vii
LIST OF FIGURES . . . . .	xii
Chapter . . . . .	1
1 INTRODUCTION . . . . .	1
1.1 Clinical Chart Reviews . . . . .	1
1.2 Crowdsourcing Clinical Chart Reviews . . . . .	3
1.3 Challenges in Building Tools to Support Clinical Chart Reviews . . . . .	5
1.3.1 Challenge 1: High-Quality Clinically Similar Terms . . . . .	6
1.3.2 Challenge 2: Clinically Similar Terms Recommendation . . . . .	7
1.3.3 Challenge 3: Document Ranking . . . . .	9
1.4 Research Approaches and Contributions . . . . .	10
1.4.1 VBOSSA Crowdsourcing Platform for Clinical Chart Reviews . . . . .	10
1.4.2 Clinically Similar Terms Extraction . . . . .	12
1.4.3 Clinically Similar Terms Recommendation . . . . .	13
1.4.4 Document Ranking . . . . .	13
1.4.5 Generalizability of the Approach in this Dissertation in other Domains . . . . .	14
2 BACKGROUND AND RELATED WORK . . . . .	16
2.1 Electronic Medical Records (EMRs) . . . . .	16
2.2 Secondary Utilization of EMRs in Medical Research . . . . .	18
2.3 Labeling Datasets through Crowdsourcing . . . . .	19
2.4 Information Retrieval Systems . . . . .	20
2.5 Search Engine . . . . .	22
2.5.1 Overview of Search Engine . . . . .	22
2.5.2 Advanced Features of Search Engines . . . . .	25
2.5.2.1 Query Expansion and Recommendation . . . . .	25
2.5.2.2 Document Ranking . . . . .	26
2.5.3 Data Preparation in Information Retrieval . . . . .	26
2.5.4 Performance Metrics for Information Retrieval Evaluation . . . . .	28
3 THE VBOSSA PLATFORM FOR CROWDSOURCING MEDICAL DATA SETS . . . . .	33
3.1 Introduction . . . . .	33

3.2	Challenges in Building a Crowdsourcing Platform for Clinical Chart Reviews	35
3.2.1	Security and Privacy	35
3.2.2	Professional Clinical Crowds	35
3.2.3	Customizable Tools to Support Clinical Chart Reviews	36
3.3	VBOSSA Crowdsourcing System	37
3.4	Professional Workshop and Crowd Worker Pool	39
3.5	Tools to Support Crowd Workers	42
3.5.1	Text Search Engine and Document Ranking	43
3.5.2	Text Highlighting	44
3.5.3	Result Review&Comparison Tool	45
3.5.4	Pixel Selection Tool	46
3.6	An Example Use Case of the Framework	47
3.7	Overview of the Finished Crowdsourced Medical Research	48
3.8	Evaluation	53
3.8.1	Overall Result	53
3.8.2	Impact of Instructions and Training Sessions	54
3.9	Discussion	55
3.10	Conclusion	58
4	CLINICALLY SIMILAR TERMS EXTRACTION	59
4.1	Introduction	59
4.2	EMR-based word2vec embedding	59
4.3	EMR-subsets Similar Terms Extraction Method	61
4.4	Evaluation	67
4.4.1	User Preference Study	67
4.4.2	Information Retrieval Performance	69
4.4.3	Elbow Method	71
4.4.4	Time Efficiency Analysis	71
4.5	Result	72
4.5.1	User Preference Study	72
4.5.2	Information Retrieval Performance	73
4.5.3	Elbow Method	76
4.5.4	Time Efficiency Analysis	76
4.6	Discussion	77
4.7	Conclusion	79
5	CLINICALLY SIMILAR TERM RECOMMENDATION	80
5.1	Introduction	80
5.2	Usage Vector Space	81
5.3	Evaluation	88
5.3.1	Usage Vector Space	88
5.3.2	Datasets	89

5.3.3	Experiments . . . . .	90
5.3.3.1	Semantic Preference Prediction Experiment . . . . .	91
5.3.3.2	Learning Curve Experiment . . . . .	92
5.4	Result . . . . .	93
5.4.1	Semantic Preference Prediction . . . . .	93
5.4.2	Learning Curve Analysis . . . . .	95
5.5	Discussion . . . . .	96
5.6	Conclusion . . . . .	102
6	DOCUMENT RANKING . . . . .	103
6.1	Introduction . . . . .	103
6.2	Document Ranking Metrics . . . . .	106
6.2.1	Negative Guarantee Ratio (NGR) . . . . .	106
6.2.2	Critical Document . . . . .	108
6.3	Document Ranking Methods . . . . .	110
6.4	Evaluation . . . . .	111
6.4.1	Chart Reviews . . . . .	111
6.4.2	Evaluation datasets for NGR analysis . . . . .	112
6.4.3	Evaluation datasets for Critical Document Analysis . . . . .	113
6.5	Result . . . . .	114
6.5.1	Negative Guarantee Ratio . . . . .	114
6.5.2	Critical Document Prediction . . . . .	116
6.6	Discussion . . . . .	117
6.7	Conclusion . . . . .	118
7	OVERALL CONCLUSION . . . . .	119
7.1	Summary . . . . .	119
7.2	The Impact of this Dissertation . . . . .	121
7.2.1	The impact of this Dissertation to the Healthcare . . . . .	121
7.2.2	The impact of this Dissertation to Other Domains . . . . .	122
7.3	The Scientific Contributions of this Dissertation . . . . .	122
	BIBLIOGRAPHY . . . . .	124

## LIST OF TABLES

Table	Page
3.1 Chart review projects support by the crowdsourcing-based information retrieval system. . . . .	54
3.2 Average accuracy and average agreement in nine baseline crowdsourcing chart review projects. . . . .	55
4.1 Data sets used for training word2vec embeddings. Vocabulary size is the number of distinct words in the data set appearing at least 50 times . . . . .	61
4.2 Framework of the user preference study. . . . .	67
4.3 Information retrieval performance evaluation data sets . . . . .	69
4.4 Distribution of positive labels in the evaluation data sets. . . . .	71
4.5 Form (a) records the overall preferences of similar terms extracted from different sources. Form (b) records the count and the percentage of selections of similar terms by User type and task type. Form (c) records the selections of each similar term extraction method. . . . .	72
4.6 Analysis of the impact of user type and task type on the preference of similar terms. User type (MD=0, Non-MD=1) and task type (Clinical=0, General=1) are the inputs of the multinomial logistic regression models. The significance levels are: **: p-Value < 0.001, *: p-Value < 0.05, one-tailed.	73
4.7 Average P@5 scores of each similar word extraction methods in the evaluation subsets. One-sided Mann-Whitney U test was applied to compare the P@5 scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with ** (p-Value < 0.001). . . . .	73

4.8	Average P@10 Scores of each similar word extraction methods in the evaluation subsets. One-sided Mann-Whitney U test was applied to compare the P@10 scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with ** (p-Value < 0.001). . . . .	74
4.9	Average AUC Scores of each similar word extraction methods in the evaluation subsets. One-sided Mann-Whitney U test was applied to compare the AUC scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with ** (p-Value < 0.001). . . . .	74
4.10	Average P@5 scores of each similar word extraction methods in the whole evaluation dataset. One-sided Mann-Whitney U test was applied to compare the P@5 scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with ** (p-Value < 0.001). . . . .	75
4.11	Average P@10 Scores of each similar word extraction methods in the whole dataset. One-sided Mann-Whitney U test was applied to compare the P@10 scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with ** (p-Value < 0.001). . . . .	75
4.12	Average AUC Scores of each similar word extraction methods in the whole evaluation dataset. One-sided Mann-Whitney U test was applied to compare the AUC scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with ** (p-Value < 0.001). . . . .	76
4.13	Average P@20 scores of searching “diabetes” and “seizure” with similar words defined by different similarity cutoff. . . . .	76



4.14	The median time (25th and 75th percentile time) medical researchers spent on reviewing one patient’s notes. One-sided Mann-Whitney U test was applied for the analysis. The significance levels are: **: p-Value < 0.001, *: p-Value < 0.05 one-tailed. . . . .	77
4.15	The similar terms for “cancer” provided by the EMR-subsets, EMR-News, EMR , and News similar term extraction methods. . . . .	78
5.1	Usage dimensions of clinical terms in each usage context. . . . .	88
5.2	Chart review tasks defined for the evaluation. . . . .	89
5.3	Candidate semantic sets of the chart review tasks. . . . .	93
5.4	Average ROC AUC scores (10-fold cross validation) achieved by supervised machine learning models in the label set constructed from the Diabetes dataset with importance cutoff 10. One-sided Mann-Whitney U test was applied. The significance levels are: {***: p-Value < 0.001, **: p-Value < 0.01, *: p-Value < 0.05 one-tailed} . . . . .	94
5.5	Average ROC AUC scores (10-fold cross validation) achieved by supervised machine learning models in the label set constructed from the AMI dataset with importance cutoff 40. One-sided Mann-Whitney U test was applied. The significance levels are: {***: p-Value < 0.001, **: p-Value < 0.01, *: p-Value < 0.05 one-tailed} . . . . .	94
5.6	Average ROC AUC scores (10-fold cross validation) achieved by supervised machine learning models in the label set constructed from the Crohn dataset with importance cutoff 1. One-sided Mann-Whitney U test was applied. The significance levels are: {***: p-Value < 0.001, **: p-Value < 0.01, *: p-Value < 0.05 one-tailed} . . . . .	95

5.7	Binomial logistic regression analysis of the impact of usage similarity in different usage contexts on workers' preference of similar terms generated by the EMR-subsets method in the AMI project with importance cutoff 40. The significance levels are: {***: p-Value < 0.001, **: p-Value < 0.01, *: p-Value < 0.05, one-tailed.}	99
5.8	Binomial logistic regression analysis of the impact of usage similarity in different usage contexts on workers' preference of similar terms generated by the EMR-subsets method in the Crohn project with importance cutoff 1. The significance levels are: {***: p-Value < 0.001, **: p-Value < 0.01, *: p-Value < 0.05, one-tailed.}	100
5.9	Binomial logistic regression analysis of the impact of usage similarity in different usage contexts on workers' preference of similar terms generated by the EMR-subsets method in the Diabetes project with importance cutoff 10. The significance levels are: {***: p-Value < 0.001, **: p-Value < 0.01, *: p-Value < 0.05, one-tailed.}	101
6.1	Statistical analysis of the behavior of different crowd workers in searching and reviewing documents in a chart review. Two-sided Mann-Whitney U test was applied to compare the activities of crowd workers. Results that are significantly different from worker 4 are marked with ** (p-Value < 0.001) and *(p-Value < 0.05).	104
6.2	Ranking and Learning-to-Rank methods defined for the evaluation.	110
6.3	Chart review tasks selected for evaluating the ranking methods.	112
6.4	Datasets for evaluating the critical document prediction of learning-to-rank methods	113

6.5	IR performances of different ranking methods in the Crohn project with document set size 200 and a positive ratio of 20%. One-sided Mann-Whitney U test was applied to compare the P@10 and NGR scores of ranking methods. Methods that significantly outperformed the baseline method (Index 1) are marked with ** (p-Value < 0.001) and * (p-Value < 0.05). . . . .	115
6.6	IR performances of different ranking methods in the AMI project with document set size 200 and a positive ratio of 20%. One-sided Mann-Whitney U test was applied to compare the P@10 and NGR scores of ranking methods. Methods that significantly outperformed the baseline method (Index 1) are marked with ** (p-Value < 0.001) and * (p-Value < 0.05). . . . .	115
6.7	IR performances of different ranking methods in the Diabetes project with document set size 200 and a positive ratio of 20%. One-sided Mann-Whitney U test was applied to compare the P@10 and NGR scores of ranking methods. Methods that significantly outperformed the baseline method (Index 1) are marked with ** (p-Value < 0.001) and * (p-Value < 0.05). . . . .	116
6.8	Average AUC scores (10-fold cross validation) of predicting critical documents with different feature spaces in the dataset constructed with minimum time difference cutoff as 30 seconds. . . . .	116
6.9	Top frequent document access patterns for making a decision in the Crohn project. . . . .	118

## LIST OF FIGURES

Figure	Page
1.1 An sample electronic medical record. . . . .	1
1.2 Pipeline to provide high-quality labels for medical research through chart reviews . . . . .	2
1.3 An example project deployed in the Amazon Mechanical Turk (AMT) crowdsourcing platform. . . . .	4
1.4 Workflow of the VBOSSA crowdsourcing platform for medical research. . .	5
1.5 Example of the requirements of clinically similar terms in different chart reviews. . . . .	8
1.6 Example text and usage contexts of clinical terms. . . . .	9
2.1 Example structure of an EMR system. . . . .	17
2.2 Tools to support clinical chart reviews . . . . .	21
2.3 Screenshot of the Google search engine, <a href="https://www.google.com/">https://www.google.com/</a> . Retrieved February,13, 2019. . . . .	23
2.4 Screenshot of the Bing search engine, <a href="https://www.bing.com/">https://www.bing.com/</a> . Retrieved February,13, 2019. . . . .	23
2.5 Example search engine for clinical chart reviews. . . . .	24
2.6 Example receiver operating characteristic (ROC) curves. . . . .	31
3.1 Structure of the framework for crowdsourcing medical data sets. . . . .	34
3.2 Privacy and security rules of the HIPAA act. . . . .	35
3.3 Example research question for analyzing if a patient with Crohn’s was clinically responsive to anti-TNF medication. . . . .	36
3.4 Vanderbilt PyBossa(VBOSSA) crowdsourcing platform. . . . .	37
3.5 Multiple-layer access control mechanism of the VBOSSA system. . . . .	38

3.6	Sensitive data de-identification mechanism of the VBOSSA system. . . . .	39
3.7	Agenda for crowdsourcing workshop with researchers. . . . .	40
3.8	Professional crowd worker pool of the VBOSSA crowdsourcing platform. .	41
3.9	Access control of the professional crowd worker pool of the VBOSSA crowdsourcing platform. . . . .	42
3.10	Customizable tools for different types of Crowdsourced chart reviews . . .	43
3.11	Search engine to support crowd workers . . . . .	44
3.12	Text highlighting tool to support crowd workers . . . . .	44
3.13	Result review&comparison tool to support medical researchers . . . . .	45
3.14	Medical image review&comparison tool to support medical researchers . .	46
3.15	Workflow of crowdsourcing medical data sets using the VBOSSA system. .	47
3.16	An example VBOSSA presenter with a text search engine. . . . .	48
3.17	User interface of the diabetes/seizure note relevance and patient condition note relevance chart review project. . . . .	49
3.18	User interface of the Acute Myocardial Infarction Chart Review project. . .	50
3.19	User interface of the Crohn’s Anti-TNF responsiveness chart review project.	51
3.20	User interface of the anesthesiology patients on dialysis chart review project.	52
3.21	User interface of the student patient interaction note comparison chart re- view project. . . . .	53
4.1	Similar terms of “cancer” from the “Clinic Note” EMR-subset embedding broken down by intra-subset similarity, inter-subsets similarity, and har- monic similarity. The harmonic similarity is used for ranking terms. . . . .	62
4.2	Example of expanded document quality analysis for “epilepsy.” The pro- portion of high similarity terms (i.e., terms that have similarities larger than 0.60 while 1.0 is the maximum value) decreases with similar term expansion.	65

4.3	Example of similarity cutoff computation. Since all terms have similarities larger than 0.40, the y-axis starts from 0.3. Similarity cutoff is at the “elbow” of the similarity curve (arrow). . . . .	66
4.4	Screenshot of the preference survey. An introduction is provided, followed by 14 questions that ask the participant to choose the best list to expand a keyword. List orders were randomized to hide source methods. . . . .	68
5.1	Usage frequencies of the similar Terms of “epilepsy” in different note sections. . . . .	81
5.2	Usage context information of an example medical note. . . . .	82
5.3	Usage counts of “EEG” distributed by the medical usage contexts of the example note shown in Figure 5.2. . . . .	84
5.4	Top 3 dimensions in each medical usage context of the usage vector of the clinical term “EEG.” . . . .	85
5.5	Usage vectors of “diabetes” and “hypertriglyceridemia” in the “Department” usage context. Only the non-zero dimensions are displayed. . . . .	86
5.6	Usage similarities of “diabetes” and “hypertriglyceridemia” in all usage contexts. . . . .	87
5.7	Workflow of learning and recommending clinically similar terms to users during a chart review by re-weighting the usage similarity vectors. . . . .	91
5.8	Average AUC scores (based on a 10-fold cross validation) achieved by logistic regression in the label sets constructed from the Diabetes Dataset. . . . .	94
5.9	Average AUC score (based on 10-fold cross validation) achieved by the logistic regression as a function of training data set sizes in the label set constructed from the Crohn Dataset with an importance cutoff of 1. . . . .	96

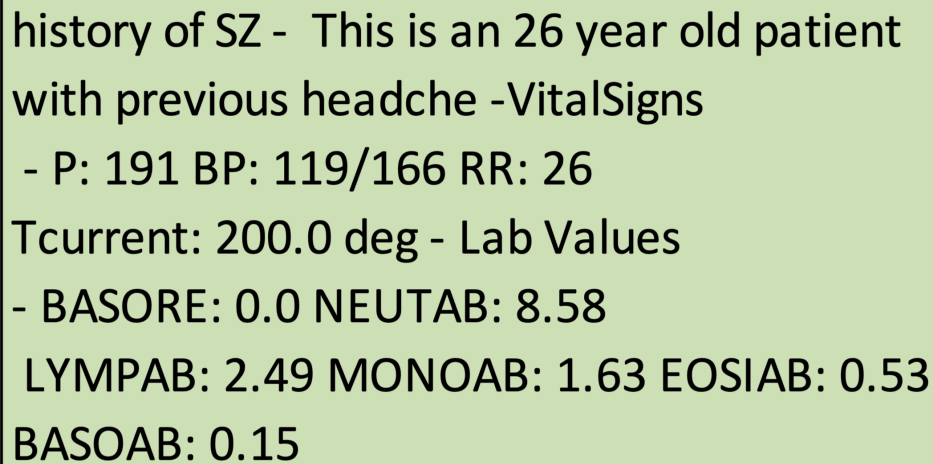
- 6.1 Time spent by two crowd workers in reviewing the same document list in a chart review task. The first worker made positive decision while the second worker made negative decision. . . . . 105
- 6.2 Comparison of two ranked document list (1-relevant document, 0-irrelevant document). Ranking 1 list has high P@5 and P@10 scores while Ranking 2 list has low P@5 and P@10 scores. The last relevant document of ranking 1 list exists in position 200 while the last relevant document of ranking 2 list exists in position 50. . . . . 107
- 6.3 Examples of critical document in a crowdsourcing chart review project. . . 109

## Chapter 1

### INTRODUCTION

#### 1.1 Clinical Chart Reviews

Clinical chart reviews [1] are one of the common components of medical research in which medical students, staff or nurses review thousands of unstructured medical notes for specific snippets that support or reject specific decisions. For example, Figure 1.1 shows a sample snippet from an electronic medical record created for a 26-year-old male patient by a physician in the Neuro-Epilepsy Department at 12:30 pm on January 1, 2016. This snippet was synthesized by the author, without any sensitive or real information of the patient. The snippet shown in Figure 1.1, supports the decision that *“This medical note mentions the history of the seizure of the patient”* but reject the decision that *“The patient was having a headache when the medical note was created”*.



history of SZ - This is an 26 year old patient  
with previous headche -VitalSigns  
- P: 191 BP: 119/166 RR: 26  
Tcurrent: 200.0 deg - Lab Values  
- BASORE: 0.0 NEUTAB: 8.58  
LYMPAB: 2.49 MONOAB: 1.63 EOSIAB: 0.53  
BASOAB: 0.15

Figure 1.1: An sample electronic medical record.

Since unstructured medical text dominates the EMRs [2, 3], it is difficult to identify important information, such as complex labels of patients, automatically from EMRs for medical researches or supervised machine learning tasks. As the example snippet shown in



Figure 1.1, clinical text is filled with misspellings (e.g., *headche*), medical acronyms (e.g., *NEUTAB* is the acronyms of *absolute neutrophil count*) [4], and abbreviations (e.g., *SZ* is the abbreviation of *seizure*) which make disambiguation difficult for software scripts [5] or even natural language processing techniques [6].

Chart reviews have been widely used to support medical research in analyzing unstructured and complex medical data, such as the analysis of severe sepsis [7], the evaluation of the quality of health service to specific patient cohorts [8], and the improvement of the quality of mental health screening in pediatric primary care [9].

The resulting data of chart reviews, such as labels of patients and text snippets that support specific decisions, are invaluable resources for both medical researches and supervised machine learning projects such as medical image processing [10, 11, 12] and medical natural language processing [13], which otherwise would be limited by smaller training data sets or training data sets without much high-quality labels (Figure 1.2). For example, in a chart review task that focuses on the Crohn’s anti-TNF Responsiveness of a certain patient cohort, the resulted label of a patient is positive if the patient was clinically responsive to anti-TNF medication. Otherwise, the label is negative. With the produced labels, supervised machine learning models could be trained to predict if a patient with Crohn’s will be clinically responsive to anti-TNF medication or not.



Figure 1.2: Pipeline to provide high-quality labels for medical research through chart reviews

However, chart reviews are also one of the most time-consuming and expensive steps in doing medical research, since scrolling through vast amounts of unstructured medical text

to produce labels or identify snippets is slow and require medical knowledge. For example, at Vanderbilt University Medical Center, it currently costs \$109 per hour for a service which pays a nurse to review patient charts and produce labels, where a large part of this fee goes to project management and other overhead. Moreover, examining a patient chart may take hours or even days for complex data sets. While some researchers have employed software scripts to infer labels from text data automatically [5], the messiness and complexity of semi-structured medical notes [14] make verifying the accuracy of the results difficult. More problematic is that the medical notes are filled with misspellings, medical acronyms, and abbreviations which make disambiguation difficult with natural language processing techniques [6]. Consequently, as the size and complexity of EMR systems keep growing [15, 16], efficient strategies, and tools are needed to help medical researchers efficiently find relevant information within unstructured medical data to support specific decisions in medical researches.

## 1.2 Crowdsourcing Clinical Chart Reviews

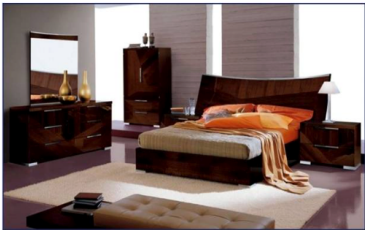
To overcome the limitations (i.e., slow and expensive) of generating labels and evidence through chart reviews, a concept, called “**Crowdsourcing**”, has been introduced to medical research fields. The core idea of crowdsourcing a labeling task or a evidence-identification task is asking a group of people (called **crowd workers**) doing a list of sub-tasks and then synthesize the results of sub-tasks to get the final result.

Crowdsourcing [17] has been proven to be a much cheaper way to get labels for a large number of data points compared to labeling data points by researchers themselves. Public crowdsourcing platform like Amazon Mechanical Turk (AMT) [18, 19, 20] and open source crowdsourcing software like PyBossa [21] have enabled researchers to ask questions to crowds of workers and quickly receive labeled responses.

Figure 1.3 shows an example AMT crowdsourcing project. In this project, the crowd workers review the question (i.e., choose the best category for a given image) and select

the correct label (e.g., bed) from multiple options. The resulted labels could be used to train supervised machine learning models (e.g., random forests or deep neural networks) to classify the images. Crowdsourcing has already been applied to support many research fields, such as bioinformatics [22], citizen science [23] and computer science [24, 25, 26, 27, 28].

**Choose the best category for this image** [View Instructions ↓](#)



Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

kitchen  
 living  
 bath  
 bed  
 outside

You must ACCEPT the HIT before you can submit the results.

Figure 1.3: An example project deployed in the Amazon Mechanical Turk (AMT) crowdsourcing platform.

The author and other members from the Hail Lab of the Department of Biomedical Informatics (**DBMI**), of Vanderbilt University Medical Center (**VUMC**), have developed a crowdsourcing system, the VBOSSA system, for medical data sets. The VBOSSA system contains 20<sup>+</sup> tools based on classical machine learning models, deep learning, and semantic embedding. The VBOSSA system helps to simplify the process of medical data review and paper preparation, and supported 60<sup>+</sup> projects and 60<sup>+</sup> medical researchers since 2016. As shown in Figure 1.4, the VBOSSA system provides a lightweight workflow to simplify the process of conducting a crowdsourced clinical chart review. We present the design and evaluation of this system in Chapter 3.

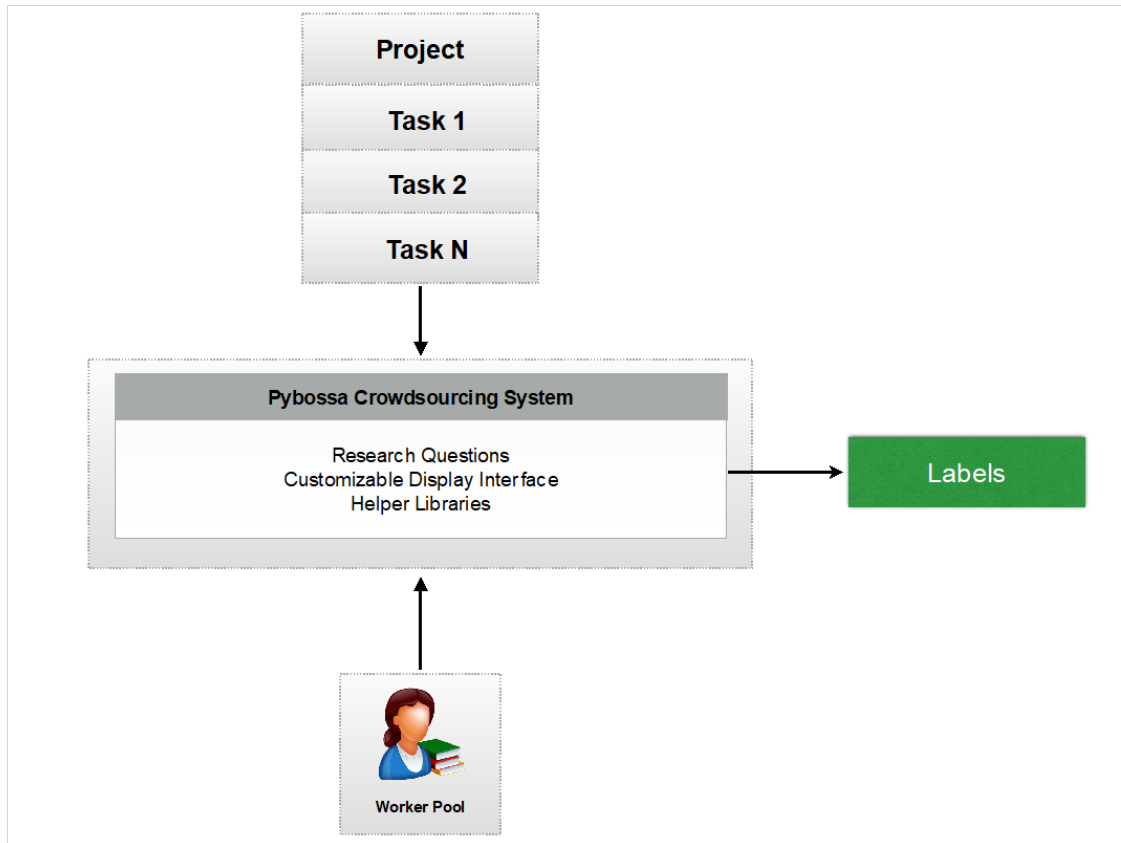


Figure 1.4: Workflow of the VBOSSA crowdsourcing platform for medical research.

### 1.3 Challenges in Building Tools to Support Clinical Chart Reviews

The other challenge for doing chart reviews is building efficient tools to help medical researchers uncover relevant information quickly from complex medical data sets. In a typical medical research project, patient charts are managed as a collection hundreds, if not thousands, of clinical documents, each of which may include tens of pages of information. In this case, even finding the specific snippets related to a patients diabetes care history or cancer medication adherence is nontrivial. While keyword search can help find some content, variations in terminology (e.g., “ca” is the abbreviation of “cancer”) and other clinical semantics make finding all relevant data challenging [29]. For example, (e.g., “Keppra” and “levetiracetam” are the same medication for treating epilepsy with different names.

Moreover, identifying all relevant snippets related to seizures in a single note including thousands of sentences, is time-consuming and requires extensive skimming. For these reasons, while the VBOSSA crowdsourcing system already reduce the cost (e.g., reducing from \$109 per hour to \$20 per hour) and time (e.g., saving the medical researchers an average 70 hours per chart review) of doing chart reviews, there are still plenty of room to future improve the efficiency of when doing crowdsourced chart reviews. Specific tools are needed to assist crowd workers in finding relevant content quickly, such as search engine, text highlighting and document ranking. We present these tools and the their back-end methodologies in Chapter 4, 5 and Chapter 6. In the rest of this section, we present the challenges in building efficient tools to support chart reviews.

### 1.3.1 Challenge 1: High-Quality Clinically Similar Terms

EMR search engines [30] have been proven to be one of the most efficient tools to assist medical researchers in finding relevant content from unstructured text in chart reviews. Query expansion [31, 32, 33], text highlighting [34, 35, 36] and document ranking [37, 30] are the three core features of an EMR search engine. A query expansion method takes the original search term, expands it into multiple terms, and returns documents containing any of the expanded terms. Similarly, a text-highlighting method highlights text within a note that include a search term or similar terms to quickly focus the reviewer on the important information. A document ranking method ranks the documents of a patient chart by specific metrics such as the number of keywords or similar terms in a document.

Underpinning the EMR search engines for supporting chart reviews is the need for high quality clinically similar terms for a given keyword. Clinically similar terms are terms that have similar medical meanings or similar usages in medical applications. For example, “Keppra” is a similar term of “epilepsy”, since “Keppra” is a medication for treating “epilepsy”. Clinically similar terms could be used to expand a search query, to enhance the text highlighting and document ranking features. For example, in a chart review project,

expanding the search of “epilepsy” with “Keppra”, or highlighting “Keppra” in medical notes provide the medical researcher additional information of the treatment of the patient. Similarly, documents that contain both “Keppra” and “epilepsy” could have higher ranks in specific chart reviews that focus on the treatments of epilepsy.

Previous work developed different types of clinically similar terms generators, including (i) ontologies, such as SNOMED-CT [38], UMLS [39], and (ii) EMR-based vector space models, such as the word2vec embeddings [37, 40]. However, little research has been done on providing high quality clinically similar terms to chart reviews and systematically evaluating the quality and quantity of similar term across various chart reviews. Also, as the needs of clinically similar terms vary with chart review tasks and user roles, specific methods are needed to adjust the clinically similar terms for different types of chart reviews.

### 1.3.2 Challenge 2: Clinically Similar Terms Recommendation

Adjusting clinically similar terms to the needs of specific chart reviews and medical researchers is essential to reduce the complexity of doing chart reviews. For example, as shown in Figure 1.5, when searching for “epilepsy” in a chart review task that focuses on the diagnosis of “epilepsy,” users may prefer “EEG” and “brain.” However, when searching for “epilepsy” in a task that focuses on treatment, users may prefer medications, such as “Keppra” or “Vimpat.” However, current query recommendation methods may recommend all *EEG, brain, Keppra, and Vimpat* given “epilepsy”, which is not suitable for varying tasks. However, previous work [41, 42, 43, 44, 45, 37, 46, 30] provide static clinically similar terms as query expansion lists to support chart reviews, which may not be suitable for chart reviews, in which the users might require different similar terms for the same search terms.

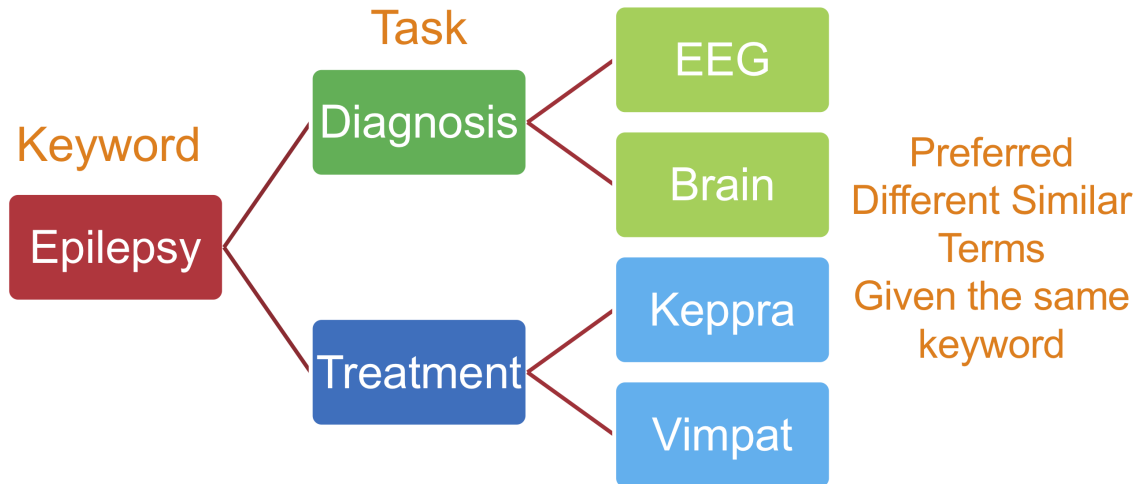


Figure 1.5: Example of the requirements of clinically similar terms in different chart reviews.

When conducting chart reviews with the VBOSSA crowdsourcing system, we interviewed crowd workers, medical researchers as well as analyzed the activity log of chart reviews, to better understand the users’ needs for clinically similar terms. The main observation of the requirement for clinically similar terms in different tasks can be quantified by how whom, and where those terms are used in EMRs. This is exemplified by Figures 1.6 (a) and (b), which show how context differs for “epilepsy” with respect to “EEG” and “Keppra.” In this case, we could infer that a medical researcher may prefer the medication information of a patient when searching for “epilepsy” and “Keppra”, and therefore, recommend similar terms that are medications for treating epilepsy, such as “Vimpat”. However, such relevant contextual information, such as the clinical department of the author of a note does not always exist within the text, and therefore can’t be captured by training text-based word embeddings. Therefore, specific methods are needed to capture and leverage such context information of clinical terms to better fulfill the requirement for clinically similar terms in chart reviews.

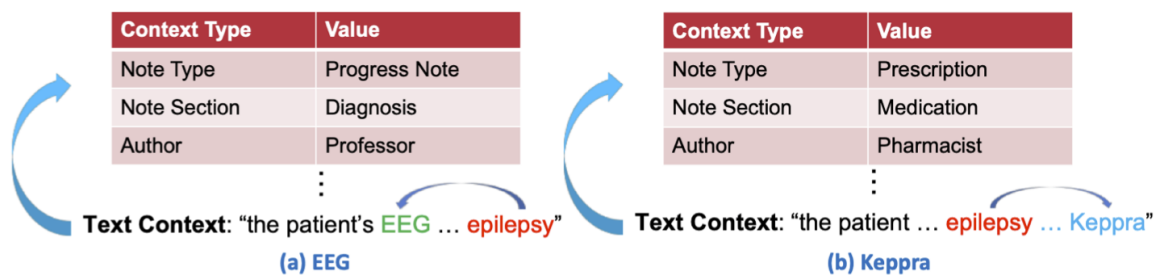


Figure 1.6: Example text and usage contexts of clinical terms.

### 1.3.3 Challenge 3: Document Ranking

In some chart reviews, the medical researchers make specific decisions without reviewing all medical notes of a patient chart. Therefore, specific ranking strategies or automatic ranking algorithms are needed to reduce the number of documents for making decisions. A ranking strategy, such as ranking by the dates of documents, is based on the medical researchers' experience in writing medical notes and doing chart reviews. Specific ranking strategies are effective in certain chart reviews, such as "If the patient was well treated in the latest visit?", in which ranking by the dates is the most efficient ranking strategy. An automatic document ranking algorithm is based on some mathematical metrics, such as the TF-IDF values of documents with respect to the keywords. The automatic document ranking algorithm first represents the document with a list of features, such as the number of keywords, length of the document, or the TF-IDF of the document and then ranks the document before returning to the users.

Document ranking has been thoroughly researched in areas that are outside the medical research fields, such as web page ranking [47], and item ranking in online retailers [48]. However, little research has been done in document ranking methods for support clinical chart reviews. Moreover, there exist specific needs for document ranking methods in chart reviews. The requirements for specific document ranking methods are based on two interesting observations in the analysis of the activity log of chart reviews.

First of all, we noticed that even if a text snippet contains a search term or its similar



terms, a user may determine the text snippet or the whole document that includes the snippet are not relevant to the research goal. For example, as shown in Table 4.4 a note may reference “diabetes” in the past medical history, but the primary purpose of the note is a recent leg injury. In this case, ranking medical notes by the number of similar terms may provide irrelevant documents to the users.

Second, we noticed that the crowd workers have their document ranking strategies (e.g., rank by note types or by date) during chart reviews. The activity log of finished chart review tasks shows that while the search engine and text highlighting significantly reduce the time for reviewing a patient chart, some crowd workers never or rarely used a search engine, but rather ranked the documents by their types and dates.

Consequently, specific ranking metrics and ranking methods are needed to better support chart reviews. For example, instead of ranking documents by their importance in a chart review, it might be better to rank and filter out non-important documents while letting the users apply their ranking strategies to identify essential documents. In this case, specific ranking methods to filter out non-important documents and corresponding ranking metrics are needed for developing and evaluating such methodology.

## 1.4 Research Approaches and Contributions

### 1.4.1 VBOSSA Crowdsourcing Platform for Clinical Chart Reviews

A crowdsourcing clinical chart review platform is needed as the first step to improve the efficiency of chart review. (Figure 1.4). A crowdsourcing clinical chart review platform not only speeds up the current chart review process more cheaply and easily, but also provides us enough feedback to develop and evaluate medical search engines for chart reviews. With a crowdsourcing system, the total time a medical researcher spent in retrieving important information can be reduced to:

1. Exports the content of the medical data or the indexes of the medical data.

2. De-identifies the medical data before showing to the crowd workers.
3. Recruits crowd workers with enough medical knowledge.
4. Provides instructions or training sessions to crowd worker.
5. Reviews and verifies the returned results of the crowdsourcing.

If we further provide appropriate mechanisms to support the medical researchers in the first three steps, the medical researcher could focus on the last step of a crowdsourced chart review. While waiting for the result of a crowdsourced chart review, the medical researcher can utilize the time for doing other valuable research projects, which do not require doing clinical chart reviews.

In Chapter 3, we describe the crowdsourcing framework for medical data sets, including the Vanderbilt PyBossa crowdsourcing system (Section 3.1), a pool of professional medical crowd workers and professional workshops (Section 3.4) and efficient tools to support chart reviews (Section 3.5). As shown in Figure 1.4, a crowdsourcing framework for medical research includes two main components:

1. An internal crowdsourcing system that has specific mechanisms to secure sensitive medical data sets, protect the privacy of patients and support crowd worker to find relevant information fast.
  - (a) Specific mechanisms to specify workers attributes, roles, and access controls;
  - (b) A de-identification routine to perturb identifiers and meet ethical and legal requirements.
2. A professional crowd worker pool and a professional workshop:
  - (a) A pool of professional workers who have the pre-tested medical knowledge to review patient charts. Also, the crowd workers should be medical students, nursing students, or faculty who have the authorization to access sensitive medical data sets in an organization.

- (b) A professional design workshop for each crowdsourcing project, including the researcher, medical personnel, computer science researchers and anthropologists, to develop crowdsourcing questions, recruit appropriate workers and verify the result.

#### 1.4.2 Clinically Similar Terms Extraction

A patient chart contains hundreds, if not thousands, of unstructured clinical documents, each of which may include tens of pages of information. Therefore, finding the specific paragraphs related to a patient's medical conditions (e.g., diabetes care history or cancer medication adherence) is time-consuming.

A user study of some chart review projects shows that although keyword search can help find some content, only half of the crowd workers ever used a search engine, even when searching can save significant time. However, variations in terminology (e.g., “dm” is used as the abbreviation of “diabetes”) and other clinical semantics make finding all relevant data challenging [29] (e.g., “ca” is used as the abbreviation of “cancer” in medical notes but used as the abbreviation of “California” in News ).

To provide an efficient search engine to support crowd worker, in Chapter 4, we present the EMR-subsets method to extract high quality clinically similar terms from multiple word2vec embeddings trained with the subsets of an EMR system. We evaluate the method with user studies, information retrieval analysis, and time efficiency analysis. The result showed that the extracted similar terms outperformed the baseline methods information retrieval performance (e.g., increasing the average P@5 from 0.48 to 0.60). Additionally, the extracted similar terms were preferred by most users and reduced the average time to answer a question.

### 1.4.3 Clinically Similar Terms Recommendation

After developing a method to extract high-quality clinical similar terms from multiple EMR-based word embedding, specific methods are needed to adjust to users' preferred similar terms in different chart review tasks. In Chapter 5, a usage vector space model is presented, in which each word is represented by a vector that captures how it is used in different medical contexts (e.g. ordering a prescription vs describing family history). By asking chart review users which terms they prefer for their given task, the similar terms can be weighted based on the implied medical context. This usage vector space model is compared against weighted, word to vector (word2vec) models in three chart review tasks. The area under the curve (AUC) is measured to determine how well similar terms are predicted.

The usage vector space outperformed the baseline word2vec embedding (e.g., AUC 0.80 vs. AUC 0.60) in all three chart review tasks. Additionally, the usage vector space significantly reduced the number of labels required to learn and predict the preferred similar terms of users (e.g., in one instance, reducing the labeling effort from 500 to 12).

### 1.4.4 Document Ranking

Since little research has been done in the user-centered document ranking approach, especially in a crowdsourcing chart review environment, further research is needed to better understand the users' needs and propose appropriate ranking metrics for evaluating ranking methods for chart reviews.

In Chapter 6, we first analyze how crowd workers interacted with the EMR search engines and how they applied ranking during chart reviews. After that, we proposed two novel ranking metrics, the negative guarantee ratio (NGR), and critical document, which are critical for developing the next generation of EMR search engine to support chart reviews. In this end, we tested the IR performance of a serial of ranking and learning-to-rank

methods using the proposed ranking metrics. The evaluation shows that traditional ranking and learning-to-rank approach are not efficient enough to support clinical chart reviews. Therefore, specific methods are needed to better rank documents in chart reviews.

#### 1.4.5 Generalizability of the Approach in this Dissertation in other Domains

In this section, we summarize our approach using general computer science language and discuss its generalizability in other domains besides the healthcare domain, such as law, finance, retailing and social media.

Given a database  $D$  and an input  $I$ , an information retrieval tool  $T$  identified relevant data points from  $D$  using some relevance metrics and ranks the relevant data points by their relevance to the input  $I$ . The challenges of building tools to support information retrieval presented in section 1.3 not only exist in the healthcare domain but also exist in other domains:

1. Unstructured data [49]. Unstructured data also exists and dominates the other domains. For example, the customers' comments in the online retailers' websites, which also contain typos, personal abbreviations and so on.
2. Requirements for similar terms[50]. Similar terms are also required to support information retrieval in other domains. For example, "tax" is similar to "corporation" in law documents that talked about corporate income tax but is similar to "person" in law documents that talked about personal income tax.
3. Complex contexts. There also exist complex contexts of words in other domains. For example, "Apple" may be relevant to "Orange" in food retailing context but may be relevant to "iPhone" in electronic retailing context.
4. Requirement for recommending dynamic queries with limited examples. For example, we recommend "banana" when given "Apple" and "Orange" and recommend "iPhone" when given "Apple" and "Social Media".

Therefore, we can generalize our approach to other domains in three steps:

1. Given a new domain, we first identify the main contexts of words in that domain. For example, different acts could be the contexts of words in the law domain.
2. We then train word embeddings per main-context and extract similar terms from the word embeddings using the EMR-subsets method (Chapter 4).
3. We then build a usage vector space based on the sub-contexts in the new domain. For example, we may consider the law of each state of the U.S.A as the sub-contexts in the law domain.

After getting the word embeddings and usage vector space of the new domain, we can support the information retrieval in that domain by supporting query construction, text highlighting and query recommendation using the same approach presented in this dissertation.

## Chapter 2

### BACKGROUND AND RELATED WORK

#### 2.1 Electronic Medical Records (EMRs)

**Electronic Health Records (EHRs)**, or **Electronic Medical Records (EMRs)** [51, 52, 53] include digital data that contain the detailed structured information about patients (e.g., ages and genders), unstructured medical text (e.g., discharge summaries), and multimedia files (e.g., CT images). EMRs allows for the analysis of healthcare processes [54, 55, 56, 57, 58, 52, 59], medical record usage[60], and support clinical research [61].The aggregated data are a rich resource for medical machine learning[62, 63, 64].

In general, EMRs and EHRs refer to the same type of medical data and are often used interchangeably [65]. However, EHRs are used more frequently when referring to the health data of patients in their whole medical history, such as the data generated by patients themselves (e.g., health data generated from a wearable device such as Fitbit [66]) and data from different medical providers (e.g., different hospitals). In contrast, EMRs are used more frequently when referring to the health data generated from a single medical data source, such as the database of one independent medical provider (e.g., the Vanderbilt University Medical Center). In this dissertation, we use the terms “EMRs” when referring the medical data generated by Vanderbilt University Medical Center (VUMC). Figure 2.1 shows the structure of an EMR system, which contains multiple databases, users and patients.

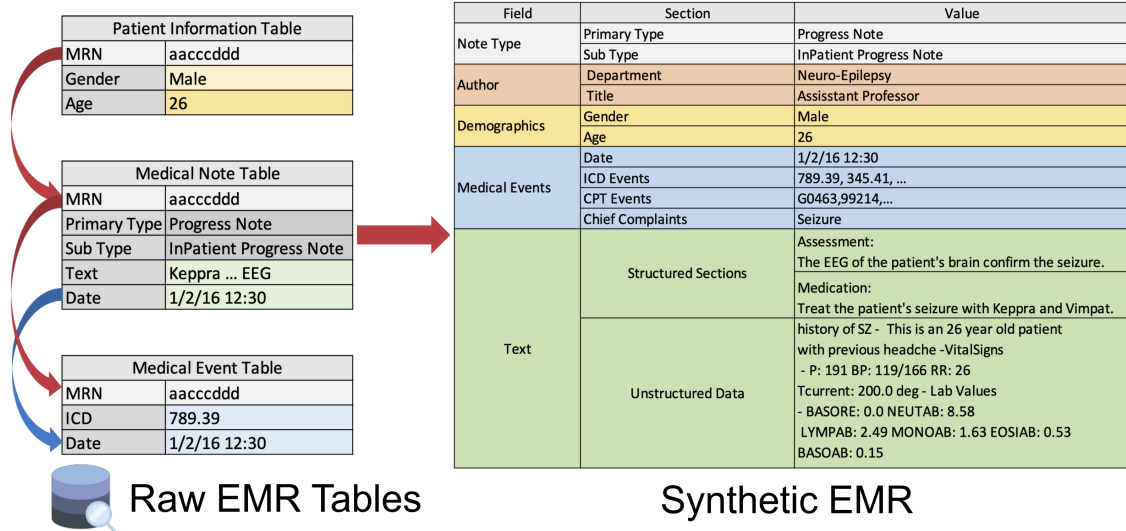


Figure 2.1: Example structure of an EMR system.

Before the concept of EMRs emerged around the late 1960s, all medical records, such as diagnoses, prescriptions, clinical visits, were in paper format. Since then, more and more healthcare providers developed and deployed electronic medical record systems to make the transmission (e.g., from one hospital to another) and retrieval (e.g., identify all historical records of patients with diabetes) of medical records easier and faster than using paper records. As the applications of EMR systems become more and more popular, in 2004, the Office of the National Coordinator of Health Information Technology was created to manage and improve the health IT efforts nationwide, including the EMR systems. After that, in the Health Information Technology for Economic and Clinical Health Act (HITECH Act [67]) of 2009, EMRs were required to provide “higher payments to health care providers that meet ‘meaningful use’ criteria, which involve using EHR for relevant purposes and meeting certain technological requirements.” In addition, the Health Insurance Portability and Accountability Act of 1996 (HIPAA act) was adjusted, according to the HITECH Act, to provide security and privacy rules for EMR systems.

After years of development and improvement, EMRs systems are widely used in healthcare providers, make the maintenance and transmission of patients health data easy and secure. In addition, as EMR systems record the details of healthcare services, healthcare



providers can leverage EMR systems to measure and improve the performance of healthcare [68, 69, 70]. Most importantly, as EMR systems record the detailed healthcare data of different types of patient cohorts, EMRs become one of the most important data sources to support clinical research [71, 10, 61, 72, 73, 74, 75, 8, 7].

## 2.2 Secondary Utilization of EMRs in Medical Research

Researches have shown that the secondary utilization of EMRs have great potential in enhancing the quality of healthcare. By labeling the raw EMRs and identifying important snippets in the raw EMRs, medical researchers can further extend EMRs to support medical research other than recording patients' medical conditions. **Labeled EMRs**, all called **labeled medical data sets**, contain labeled medical notes, labeled patients and snippets in medical notes that support the labels. A label could be Boolean values (e.g., whether the patient had a diabetes history or not), categories (e.g., the types of cancer) or continuous values (e.g., a patient's length-of-stay after being admitted).

Labeled medical data sets are utilized in different medical research, such as helping the diagnosis of specific diseases (e.g., skin cancer [76, 12, 10], diabetes [76, 77]), predicting the length of stay of patients from a specific cohort [78, 79, 80, 81, 82], predicting the readmission of specific patient cohorts [83, 84, 85, 86, 87], and predicting the diagnosis codes(e.g., ICD-9 codes) [88, 89, 90, 63]. In addition, as the size of EMRs increases significantly in recent years, applying big data analytic to EMRs was introduced to further enhance the research in healthcare [91, 92, 93, 94, 95, 96].

Automatic scripts and manually review are two methods to produce labeled medical data sets. An automatic labeling script [5, 97, 98] identifies specific patterns from medical notes, such as phrases, values, and then determines the labels with a given threshold (e.g., labeling a document as important if the number of "diabetes" in the document is more than 10). Automatic labeling scripts are the cheapest and fastest ways to label a medical data set.

However, as the unstructured medical data (e.g., progress notes in free text format, scanned paper documents) dominates the EMR systems [73, 99], it is difficult to generate reliable and high-quality labels for medical data sets with automatic labeling scripts. In this case, manual review methods (e.g., chart reviews [71, 100]) are more reliable to produce labels for medical data sets. In a chart review task, people with professional knowledge (e.g., the knowledge about the diagnosis of all kinds of cancers) review the medical notes of a patient (i.e. the patient chart) to label patients or notes (e.g., labeling if a patient had lung cancer or labeling if a note mentions about lung cancer). Although manual review methods provide high-quality labels for medical data sets, it is expensive (e.g., range from hundreds of dollars per hours to thousands of dollars per patient chart) and time-consuming (e.g., range from days to months).

### 2.3 Labeling Datasets through Crowdsourcing

Many research and engineering works have been done to improve the methods used for providing high-quality labels for training supervised machine learning models. For example, researchers have developed semi-auto methods to generate high-quality labels, such as human-in-the-loop [101] and interactive labeling [102], to help non-experts in computer science to transform professional knowledge into automatic scripts to generate labels. Another example is the idea of crowdsourcing [103, 104, 105], which asks a ground of workers to create labels for a data set. Since most of the medical notes are unstructured and are filled with misspellings, medical acronyms, and abbreviations, the applications of semi-auto methods are limited. In contrast, crowdsourcing research data sets have been proved to be cheap and fast. However, due to the security and privacy rules of HIPAA act, it is impossible to crowdsource medical data sets in public commercial crowdsourcing platform, such as the Amazon Mechanical Turk [106, 107, 108, 17, 18, 19, 20]. Although there exist open-sourced crowdsourcing platforms (e.g., the PyBossa [109]), little research has been done to build a crowdsourcing platform to fulfill the security and privacy require-

ments when crowdsourcing medical data sets. Most importantly, crowdsourcing medical data sets not only requires a platform to protect patients' privacy but also requires professional crowd workers and efficient tools to help retrieving information from patients' chart fast, such as keyword search engine [110, 30, 45, 46, 44, 111] with advanced features(e.g., query expansion and text highlighting) and data visualization [112, 113, 114].

Although crowdsourcing data sets have been proved to be cheap and fast, the reliability of crowdsourcing data sets depends on multiple facts, such as the design of questions, the abilities and knowledge levels of workers [28, 105, 20]. The Amazon's Mechanical Turk (AMT) has already proved that crowdsourcing provides reliable labels as other traditional methods in psychology and other social sciences [20]. In healthcare, researchers have done research about the outcomes from crowdsourcing, such as the reliability of medical diagnosis from crowdsourcing [105]. In general, a crowdsourced data set has a standard expectation (e.g., >80% accuracy) for the quality of the results from crowd workers. If the result is below the expectation, we may introduce redundant workers, since previous research showed that the more redundant worker we recruited, the better accuracy we can expect from the workers [28]. For some complex datasets, such as medical datasets, researchers may conduct a training session to train the workers and verify the result with some golden standards. In this case, we may better guarantee the reliability of the results of crowd workers [21, 115].

## 2.4 Information Retrieval Systems

An Information Retrieval System [116, 117, 118] is a collection of computer programs that runs on the top of electronic files, such as plain text files [119], digital images (e.g., JPG files or camera data [120, 121]), databases (e.g., SQL [47] and NoSQL [122] databases), to get inputs (e.g., keywords) from the user and extract and return relevant information in the electronic files. In some application fields, the information retrieval system and search engine may refer to the same type of computer programs [123]. However, the informa-

tion retrieval system is the superset of search engine (as shown in Figure 2.2), which may also be used when referring to data extraction programs [124, 125, 126, 127, 128, 129, 130, 131, 49] that identifies meaningful data points from complex and abstract datasets, such as determining phase shifts in multiple types of circadian time-course data [132] or data visualization programs [113, 85, 133, 114, 112] to help users better understand the medical data. While search engines are used more frequently when referring to the computer program that identifies the important information the users preferred from structured, unstructured readable data.

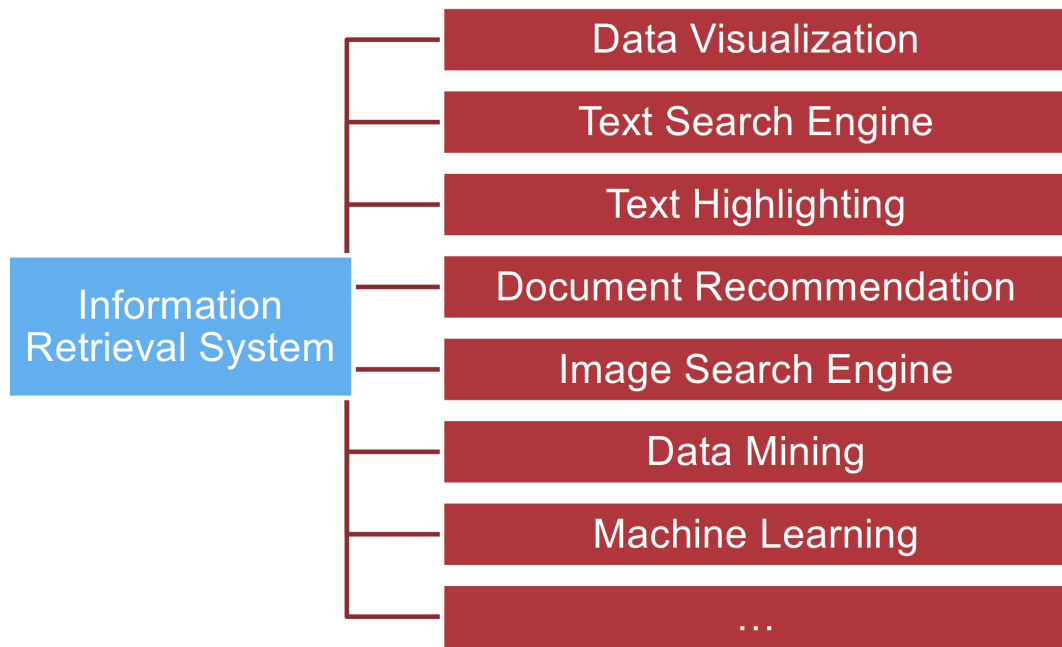


Figure 2.2: Tools to support clinical chart reviews

Researchers have developed different types of information retrieval systems to support chart reviews, including search engine, text highlighting and document ranking.

A keyword search engine [30, 134, 135] returns documents that contain the search term and rank documents by a specific metric value, such as the frequency of the search term in the document. Commercial keyword search engines, such as Google, embed supportive features, such as query expansion [136, 32, 137, 41, 39] and learning-to-rank [138, 139, 140, 141, 142, 143, 94], to refine the quality of search results. Query expansion is a method

that expands the search to include terms that are semantically similar to the term (e.g., “insulin” is semantic similar to “diabetes,” “boy” is semantic similar to “man”) and returns documents with any of the similar terms. In addition, some search engine associates with text highlighting feature, which highlights the similar terms in the returned documents to support users reading the documents.

## 2.5 Search Engine

### 2.5.1 Overview of Search Engine

The first search engine is called Archie, was created in 1990 by Alan Emtage, a student from the McGill University in Montreal [144]. Archie searched FTP sites and created indexes of downloadable files to speed up the searching for specific files. Since then, people developed many search engines for different application fields, such as Yahoo! [145], Google [146] and Bing [147] for web page search, EMERSE [30] for medical note search, and Pinterest [120] for visual search.

Typically, a search engine retrieves information in three steps:

1. Get input from the users:
  - takes explicit input from users, such as keywords [119], example text [119], example images [120];
  - takes implicit input from users, such as click-through data [142], time spent in each clicked items in web pages [148], saved items in an online retail website (e.g., the Amazon) [48];
2. Retrieves relevant data, such as text, website, web pages, from candidate locations, such as a web page index database, using specific metrics (e.g., from simple number-of-keywords to complex metric such as Text Rank [149]).

3. Ranks relevant data by some ranking metrics, such as the number of keywords in the web pages or the page rank values [150] of the pages;

Google [146] (Figure 2.3) and Bing [147] (Figure 2.4) are two typical commercial text-based search engines that identifies useful web pages back to the user given a list of keywords. In medical research, the target data points are complex and abstract, which require specific data extraction methods.

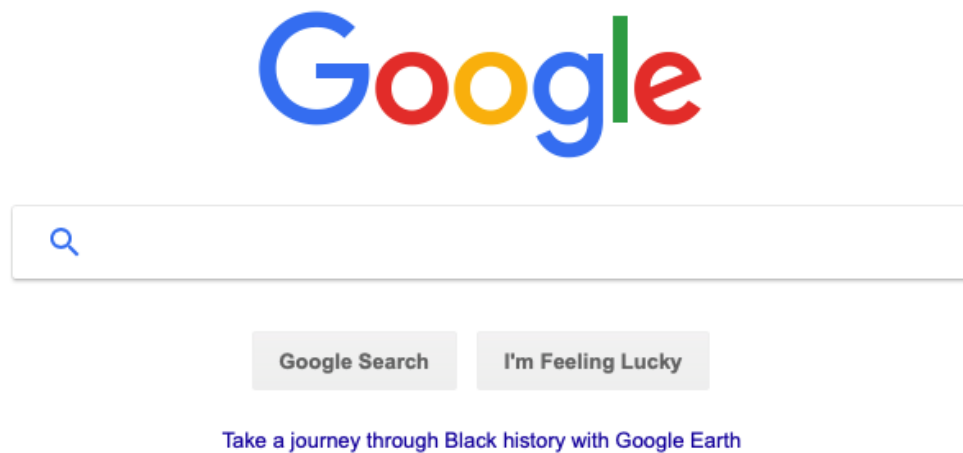


Figure 2.3: Screenshot of the Google search engine, <https://www.google.com/>. Retrieved February,13, 2019.

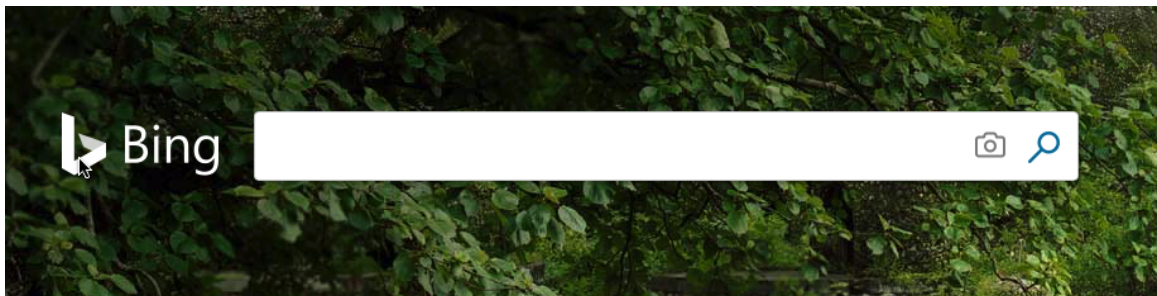


Figure 2.4: Screenshot of the Bing search engine, <https://www.bing.com/>. Retrieved February,13, 2019.

Figure 2.5 shows an example search engine used for retrieving relevant medical notes

from a medical data set. The medical search engine helps the user to retrieve relevant medical notes in three steps:


1. The user first enters a keyword “epilepsy” as the input;
2. The search engine executes a database command in the background to identify all medical notes that contain the keyword from the database;
3. The medical search engine ranks the retrieved medical notes by the number of keyword “epilepsy” in each medical note;
4. The medical search engine also provides visualization efforts, such as text highlighting and frequency timeline, to support the user review the retrieved notes;

By Keywords ▼ epilepsy

🔍 Search
🔗 Expansion
💾 Save
📄 Load

📄 Documents
📅 Timeline

## Search Results



show search terms
✕ close

Filter Documents:

Document ID	Document Type	Date	Rank Value	Length	Snapshot
1544	epilepsy follow-up clinic visit	2011-01-28	3.00	764	...line 5 : ...< **Institution: <span style="background-color: #28a745; color: white;">epilepsy</span> ...line 79 : ...this is a <b>age</b> [in 40s]-year-old man with history of <span style="background-color: #28a745; color: white;">epilepsy</span> ...line 95 : ...instructor of clinical neurology and <span style="background-color: #28a745; color: white;">epilepsy</span> ...
1428	epilepsy clinic follow-up visit	2015-05-15	2.00	585	...line 66 : ...with the interictal expression of a partial <span style="background-color: #28a745; color: white;">epilepsy</span> with a ...line 88 : ...1. continue current <span style="background-color: #28a745; color: white;">epilepsy</span> medications at the same dose ...
1821	epilepsy clinic follow-up visit	2013-11-29	2.00	587	...line 64 : ...with the interictal expression of a partial <span style="background-color: #28a745; color: white;">epilepsy</span> with a ...line 86 : ...1. continue current <span style="background-color: #28a745; color: white;">epilepsy</span> medications at the same dose ...
1614	stallworth admission history and physical	2012-06-17	2.00	1339	...line 19 : ...**age[in 40s]yo aam with history of <span style="background-color: #28a745; color: white;">epilepsy</span> and cognitive i ...line 142 : ...**age[in 40s]yo aam with history of <span style="background-color: #28a745; color: white;">epilepsy</span> and cognitive ...
1401	neurology follow-up visit	2010-09-10	2.00	731	...line 40 : ...we discussed a number of things related to <span style="background-color: #28a745; color: white;">epilepsy</span> and hi ...line 71 : ...with the interictal expression of a partial <span style="background-color: #28a745; color: white;">epilepsy</span> with a ...

Showing 1 to 10 of 27 entries

Figure 2.5: Example search engine for clinical chart reviews.

## 2.5.2 Advanced Features of Search Engines

### 2.5.2.1 Query Expansion and Recommendation

Query expansion [151, 152, 153, 33, 39, 136, 154, 155, 156, 32, 157] is an information retrieval technique that automatically adds relevant keywords to the original keywords provided by the user to expand the search scope and therefore get documents that are relevant to the user's search goal. Query recommendation is another information retrieval technique to help the user search more efficiently. Similar to query expansion, a query recommendation method [158, 159] takes the user's original keywords and recommends relevant keywords. However, a query recommendation method may provide the user with more options to select or refine the recommended expanded keywords. Moreover, many query recommendation methods applied advanced techniques, such as the users' previous search log [160], and the users' behavior models [160], to better recommend keywords to the users.

Both the query expansion and query recommendation features of current web search engines [161, 142, 94] and EMR search engines [43, 44, 46] highly rely on the usage log, such as the click-through data or search log. However, it is hard to get enough click-through data or search log from new or small chart review projects. Formally, providing query expansion or query recommendation with limited information about the users is called as the **“Cold Start Problem”** [162] faced by many search engines.

Therefore, pre-identified clinically similar terms are essential to enhance EMR search engines [30, 163] to support chart reviews. There are two popular ways to produce clinically similar terms: (i) ontologies, such as SNOMED-CT [38], UMLS [39], and (ii) EMR-based semantic embeddings. While clinical ontologies are hard to construct and update, EMR-based semantic embeddings are trained using unsupervised machine learning methods (e.g., GloVe [164], word2vec [40]) on EMR text and identify similar terms based on the EMRs semantics. For example, Pakhomov et al. [165] found that word embeddings cap-



ture semantic relatedness between medical terms. Moreover, Zhu et al. [41] and Hanauer et al. [46] showed that semantically-based query recommendation systems could effectively expand search queries.

### **2.5.2.2 Document Ranking**

When using the keyword search or expanded-keyword search, a search engine may return thousands or even billions of results (e.g., web pages that contain “search”). Therefore, specific methods are needed to rank the search results to show users the best results. Researchers have developed different types of ranking methods, such as ranking by the number of keywords [46, 37], the number of expanded keywords [166, 46], the relationships between documents (e.g., the page rank [149, 150]), and information-theory-based models (e.g., cross entropy [155], TF-IDF [167]).

To further enhance the quality of document ranking, a technique named “learning-to-rank” was introduced. Learning-to-rank [138, 142] re-ranks search results by learning from labels provided by the users of search engines. In general, a learning-to-rank system represents each document by a set of features, such as bag-of-words. It then trains a classification model, such as support vector machine or logistic regression, with user-provided labels to re-rank the search result [142, 143]. Learning-to-rank has been widely used in web search engines, such as Google [141], and recommendation systems in online retail, such as Amazon [168]. In medical research, researchers also applied learning-to-rank approaches to identify important terms in a clinical document [169, 170] or to re-rank clinical documents to support medical research [171].

### **2.5.3 Data Preparation in Information Retrieval**

Data preparation is the process to i) clear “dirty and useless” data points from the database, which may reduce the quality of search result, ii) fix errors, which may bring incorrect search results and iii) create features for the data points in the database to support

advanced information retrieval features, such as learning-to-rank. The following are four typical types of data points that a data preparation process needs to deal with:

1. Missing data [172, 73]. Missing data points are the data points with invalid values. For example, a patient with “None” age value in an EMR system.
2. Incorrect data [172]. Incorrect data points are the data points with sub-values that are not in a valid format. For example, the negative age values of patients are incorrect data.
3. Outliers [172, 173, 174, 175]. Outliers are those data points that have valid format but do not meet some metrics based on common sense or domain knowledge. For example, a data point that records a user spent one hour in reviewing a patient’s chart during a chart review task, which may be an outlier since most of the users in the same task spent ten minutes in reviewing a patient’s chart.
4. Unstructured data. Unstructured data are those data points with varying formats. For example, “DOB” and “Date of Birth” are unstructured data points, since there are used interchangeably in the database. Unstructured data may cause an information retrieval tool to miss important data points.

It is challenging to handle missing data, incorrect data, outliers and unstructured data in the data preparation process. Simple solutions include removing incorrect data and outliers [176, 177], filling missing data with the average value the same type of data points (e.g., filling missing pixels with the average value of its neighbors in damaged images) [178]. Depends on the tasks, researchers developed different types of data preparation solutions, such as using a Naive Bayes model to fill missing data [179] and using Kernel-based methods to filling missing data [180].

Constructing features[181, 47] is the last step in preparing data before providing information retrieval service to users. Features are a list of continuous or categorized values that

represent the data points in the databases. For example, the length of a document could be a feature of the document. Constructing features not only reduces the time for identifying relevant documents in an information retrieval process but also supports advanced information retrieval techniques, such as learning-to-rank [141], which could learn from the users' behavior to improve the search result.

#### 2.5.4 Performance Metrics for Information Retrieval Evaluation

In this section, we introduce the metrics for measuring the performance of information retrieval tools.

Given an information retrieval tool and an evaluation database, in which the data points have golden standard labels to certain inputs, we test the performance of the information retrieval tool using the following steps.

1. Provide input to the information retrieval tool, such as keyword(s), example(s).
2. The information retrieval tool goes through the database and identifies the data points, such as documents, that are relevant to the input.
3. The information retrieval tool then ranks the returned data points by their relevance to the input.
4. We measure the quality of the ranked result with a certain metric (e.g., the accuracy of the top 10 returned data points). In the rest of this section, we introduce five typical metrics for measuring the performance of information retrieval tools: **precision, recall, precision-at-K (P@K), F1 score and ROC AUC** [182].
5. To make sure the evaluation results are reliable, we may repeat the evaluation in a certain way, such as **cross validation** [183, 184, 185, 186, 187], which is introduced in the rest of this section, and compute the average performance.

First of all, we introduce the definitions of basic concepts of measuring the performance of information retrieval tools.

Given a database  $D$ , in which each data point  $d_i$  has a relevance label to an input  $I_j$ :  $l(d_i, I_j)$ . For example, if document  $d_i$  is relevant to keyword “diabetes”, the relevance label is 1, otherwise the label is 0. We define the relevance subset of the input  $I_j$  in  $D$  as  $D(I_j) = \{d_1, d_2, \dots, d_l\}$ , where  $l(d_i, I_j) = 1$

Each information retrieval tool  $T$  contains a relevance measurement method  $R_T = \{R_1, R_2, \dots, R_n\}$ , in which  $R_T(d_i, I_j)$ , to compute the relevance of a data point to an input  $I_j$ . For example, given keyword “diabetes” as the input, one possible relevance metric could be the number of “diabetes” in each documents.

Given an input  $I_j$ , an information retrieval tool  $T$  goes through the database  $D$ , computes the relevance values of all data points, selects data points with relevance values no less than a cutoff value  $C$ , and returns the ranked search result:

$$S_T(D, I_j) = \{r_1, r_2, \dots, r_p\}, \text{ where } R_T(r_q, I_j) \geq R_T(r_m, I_j) \text{ and } q > m \text{ and } R_T(r_1, I_j) > C \quad (2.1)$$

We define the precision of the search result of the information tool  $T$  as the ratio of relevant data points in the search result. Precision is useful when we focus on the accuracy of an information retrieval tool in identifying relevant data points.

$$\text{precision}_T(I_j) = \frac{|r_j | r_j \in D(I_j)|}{|S_T(D, I_j)|} \quad (2.2)$$

We define the precision-at-K (P@K) of the search result of the information tool  $T$  as the ratio of relevant data points in the top K (e.g., top 5) search result. P@K is useful when we focus on how well an information retrieval tool identifies the most relevant data points.

$$\text{precision-at-K}_T(I_j) = \frac{|r_j | r_j \in D(I_j), j \leq K|}{|S_T(D, I_j)|} \quad (2.3)$$

We define the recall of the search result of the information tool  $T$  as the ratio of relevant data points in the search result to all relevant data points in  $D$ . Recall is useful when we focus on how well an information retrieval tool in identifying all relevant data points.

$$\text{recall}_T(I_j) = \frac{|r_j|r_j \in D(I_j)|}{|D(I_j)|} \quad (2.4)$$

We define the F1 score of the search result of the information tool  $T$  as the harmonic mean of the precision and recall. The F1 score measures the balance of precision and recall value of an information retrieval tool. The higher the F1 score is, the better the information retrieval in balancing precision and recall.

$$\text{F1}_T(I_j) = 2 \times \frac{\text{precision}_T(I_j) * \text{recall}_T(I_j)}{\text{precision}_T(I_j) + \text{recall}_T(I_j)} \quad (2.5)$$

Before introducing the ROC AUC score, we first introduce the concepts of true positive ratio (TPR) and false positive ratio (FPR). TPR is the same concept of recall with a different name. The FPR is the ratio of the number of irrelevant data points in the search result to the irrelevant data points in the database  $D$ .

$$\text{TPR}_T(I_j) = \frac{|r_j|r_j \in D(I_j)|}{|D(I_j)|} \quad (2.6)$$

$$\text{FPR}_T(I_j) = \frac{|r_j|r_j \notin D(I_j)|}{|D| - |D(I_j)|} \quad (2.7)$$

Given different relevance cutoff value  $C$ , the search result changes. With a serial of cutoff value  $C$ , we can get a list of TPR and FPR values. When using the FPR as the x-axis and the TPR as the y-axis, we draw a receiver operating characteristic curve, also called the ROC curve. Figure 2.6 shows examples of ROC curves. The Area under the ROC curve is called the Area Under the ROC Curve (AUC). Since the maximum values of TPR and FPR are 1.0, the maximum value of AUC is 1.0. AUC measures the trade-off between TPR and

FPR. AUC is useful in information retrieval experiments, in which we want to balance the quality of search result. High AUC value means the information retrieval tool returns as many true positive results as possible while introducing as the less false positive result as possible. The higher the AUC value is, the better the information retrieval tool is.

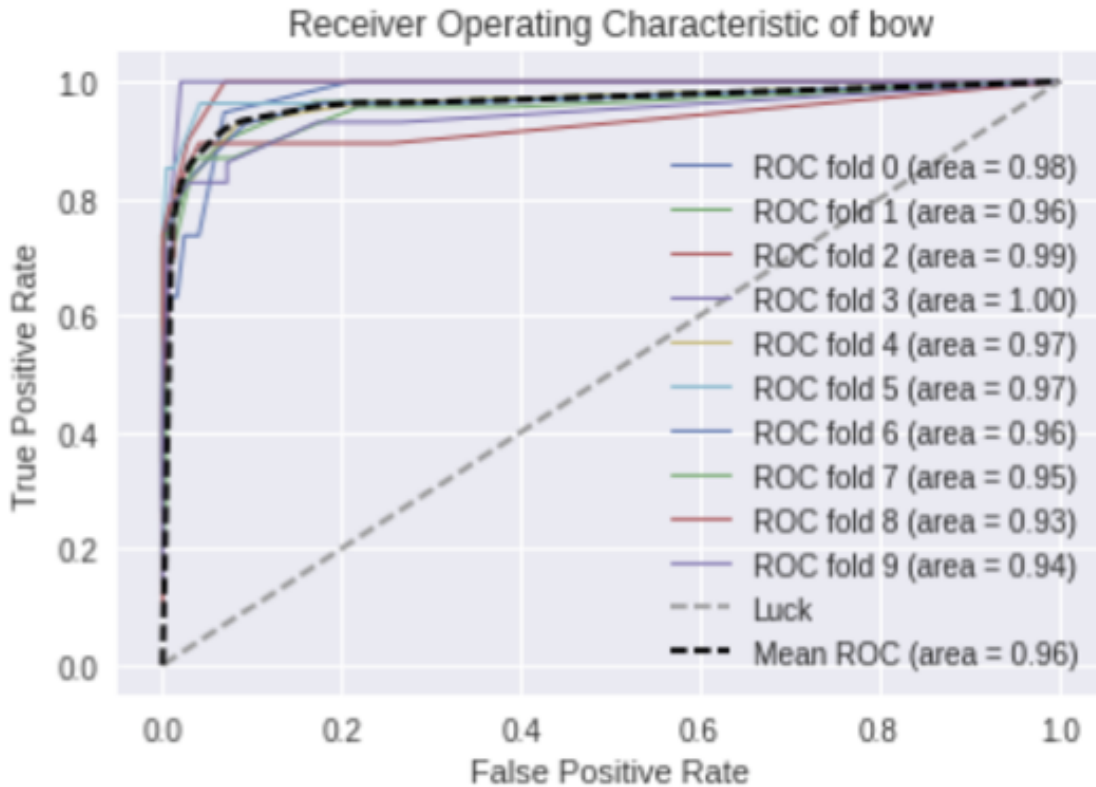


Figure 2.6: Example receiver operating characteristic (ROC) curves.

To better test the stability and reliability of information retrieval tools, especially the ones with the learning-to-rank feature [188, 167, 143, 189, 190, 138], researchers have introduced the idea of cross-validation. The idea of cross-validation is as following.

1. Divide the evaluation database  $D$  into two parts, the training set  $D_{train}$  and test set  $D_{test}$ .
2. Train a learning-to-rank model using the training set and test the model with the test set.

3. Re-Train the learning-to-rank model using the test set and test the model with the training set.

K-fold cross-validation [191, 183] is an enhanced version of the cross-validation. Given a  $K$  where  $K \geq 2$  and  $K \leq |D|$ , we divide the database  $D$  into  $K = \{1, 2, \dots, k\}$  equal subsets, we choose the  $i$  subset as the test set and train the learning-to-rank model with the combination of other subsets. Then, we measure the performance of the learning-to-rank model  $K$  times and compute the average performance. Using K-fold cross-validation, we can measure how well an information retrieval tool performs when given unseen data points while guaranteeing the reliability of the result.

Besides the K-fold cross-validation, there are other versions of cross-validation, such as Leave-p-out cross validation [192, 155], which are designed for the needs to evaluate different types of information retrieval tools.

## Chapter 3

### THE VBOSSA PLATFORM FOR CROWDSOURCING MEDICAL DATA SETS

#### 3.1 Introduction

Crowdsourcing has gained notoriety as services like Amazon Mechanical Turk (AMT) have enabled researchers to ask questions to crowds of workers and quickly receive labeled responses. These human labeled data sets are increasingly important for training supervised machine models, as labels do not exist for many important research questions and cannot be produced with automated methods. Unfortunately, crowds composed of individuals from the general public are inappropriate for numerous types of data sets that require crowdsourcing, such as clinical data, due to legislation (e.g., the Health Insurance Portability and Accountability Act of 1996) and organizational policies. In particular, privacy concerns prevent arbitrary users from accessing these data. Moreover, the subject matter being analyzed requires highly specialized training and expertise to accurately produce a label, which is often not available in a public crowd.

This chapter outlines the crowdsourcing framework we developed for medical data sets and one current deployment of the system. There are many components necessary for building such an environment to allow for scalable human computation on medical data sets. Broadly, the main components of the system include: (i) a crowdsourcing system that can be deployed within an organization that has the ability to specify workers' attributes, roles and access controls, (ii) de-identification routines to perturb identifiers and meet ethical and legal requirements, (iii) graphical user interfaces to display sensitive data, (iv) and machine learning tools to assist workers to produce labels quickly. Moreover, beyond the technical components, this chapter describes organizational processes that are needed to train researchers about crowdsourcing so they can construct well-defined questions for the crowd, and approaches to recruit skilled workers.



In the rest of this chapter, we present the crowdsourcing framework for medical data sets. An effective crowdsourcing system for medical data sets can change how medical research is done and allow researchers to solve important problems. In our experience, the chart review process is often a key rate limiting step for modern studies; crowdsourcing has the ability to substantially lower the time to complete clinical studies. Additionally, the resulting labels are invaluable resources for supervised machine learning researchers that otherwise would be limited by smaller training data sets. As shown in Figure 3.1, the crowdsourcing framework consists of an internal crowdsourcing platform, a pool of professional crowd workers, a professional workshop and the tools to support chart reviews.

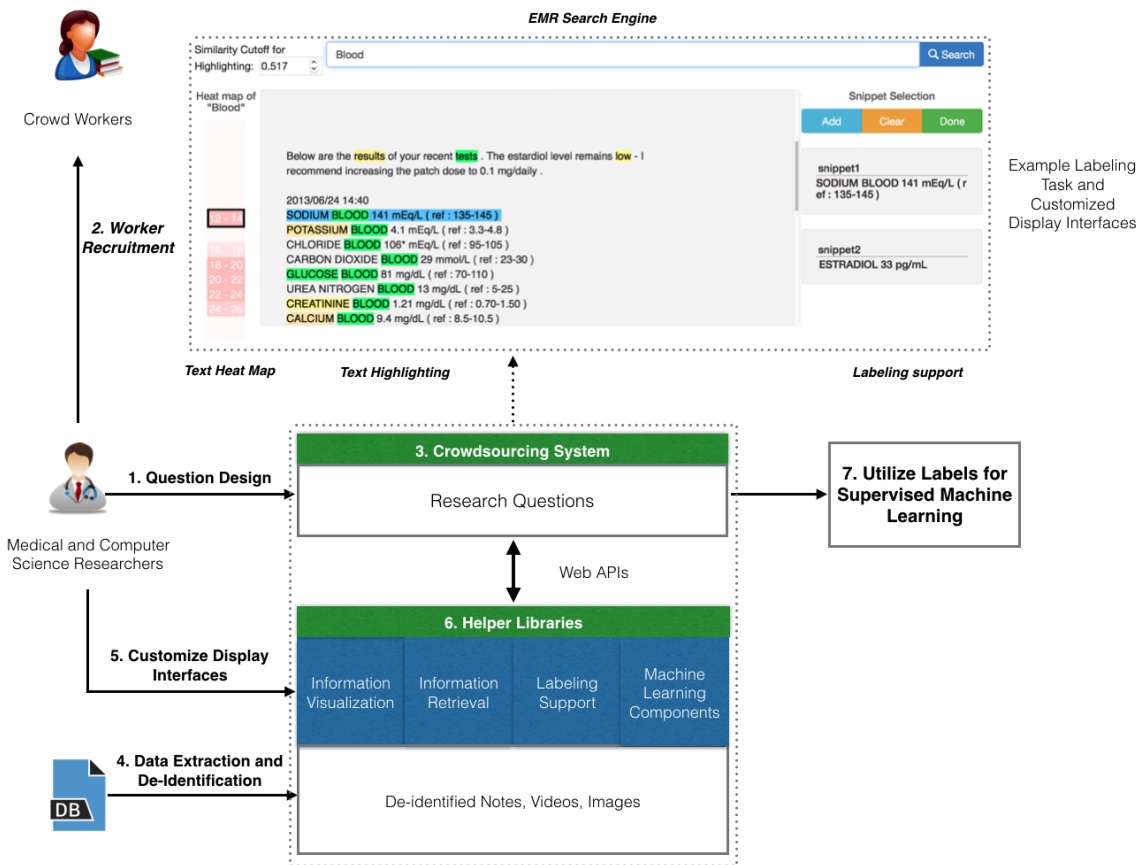
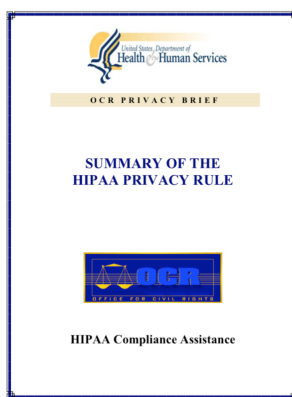


Figure 3.1: Structure of the framework for crowdsourcing medical data sets.

## 3.2 Challenges in Building a Crowdsourcing Platform for Clinical Chart Reviews

### 3.2.1 Security and Privacy

First, as there are specific **security and privacy requirements** for crowdsourcing medical data sets. Crowds composed of individuals from the general public are inappropriate for numerous types of data sets that require crowdsourcing, such as clinical data, due to legislation (e.g., the Health Insurance Portability and Accountability Act of 1996 [193, 194], as shown in Figure 3.2) and organizational policies. Privacy concerns prevent arbitrary users from accessing these data. Moreover, allowing full access to trusted people may still expose too much sensitive information of patients.



- Health Insurance Portability and Accountability Act (HIPAA).
- **Privacy** Rule protects the privacy of individually identifiable health information.
- **Security** Rule protects all individually identifiable health information a covered entity creates, receives, maintains or transmits in electronic form.

Figure 3.2: Privacy and security rules of the HIPAA act.

### 3.2.2 Professional Clinical Crowds

Second, the **complex medical data sets** being analyzed requires highly specialized training and expertise to accurately produce a label, which is often not available in a public crowd. For example, as the research question shown in Figure 3.3, researchers need to answer a complex medical question “Was a patient with Crohn’s clinically responsive to anti-TNF medication?” by reviewing hundreds of medical notes and identify text snippets in medical notes as evidence. A professional workshop including medical researchers and

computer scientists is needed to design the research question, customize the user interfaces and tools. In addition, a pool of professorial crowd workers with specific medical knowledge is needed for analyzing complex medical data sets.

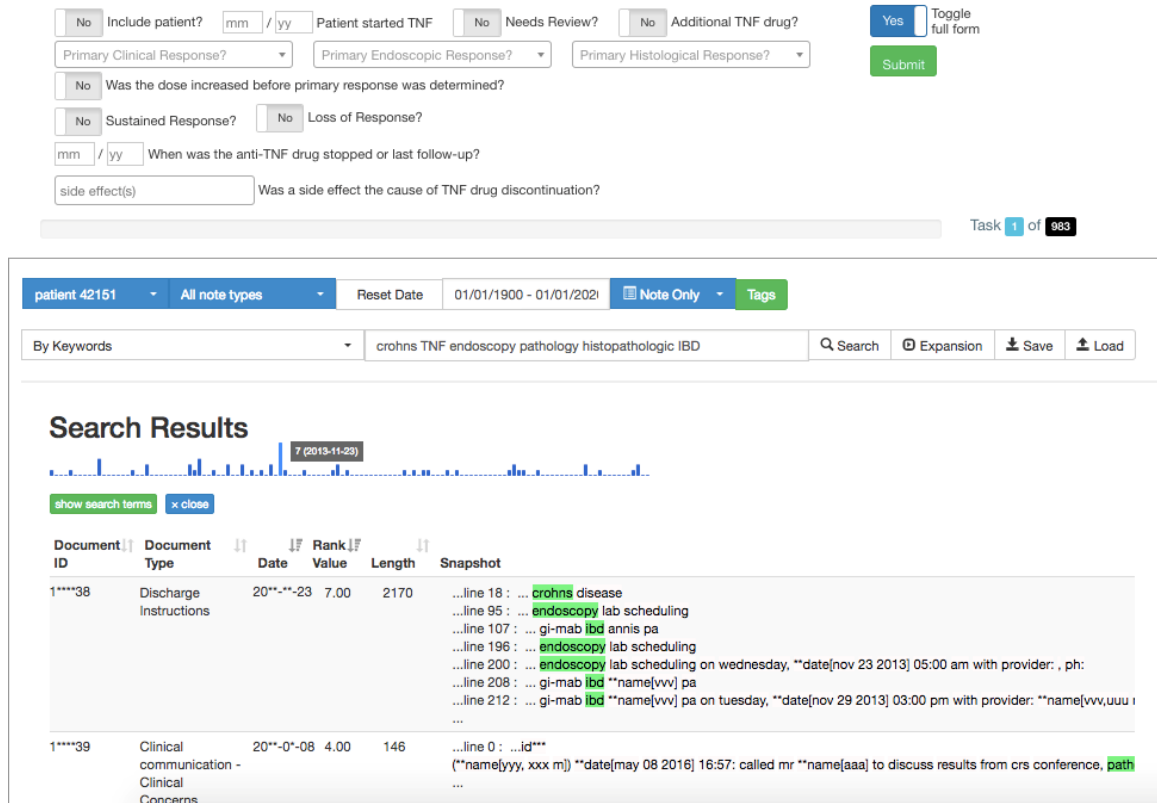


Figure 3.3: Example research question for analyzing if a patient with Crohn’s was clinically responsive to anti-TNF medication.

The crowdsourcing platform should provides medical researchers a lightweight, customizable pipeline that significantly reduces the cost and time to complete medical research while increasing reproducibility and accuracy and maintaining privacy and security standards.

### 3.2.3 Customizable Tools to Support Clinical Chart Reviews

One major challenge for crowdsourcing workers is uncovering relevant information quickly from complex data sets. For example, in healthcare, patient charts are managed as

a collection hundreds, if not thousands, of clinical documents, each of which may include tens of pages of information. Finding the specific paragraph related to a patient’s diabetes care history or cancer medication adherence is nontrivial. While keyword search can help find some content, variations in terminology and other clinical semantics make finding all relevant data challenging [29]. Moreover, identifying all relevant text in a single note related to, say, seizures remains time-consuming and requires extensive skimming. For these reasons, the crowdsourcing framework requires additional tools to assist workers in finding relevant content quickly, such as text highlighting and data visualization for summarization.

### 3.3 VBOSSA Crowdsourcing System

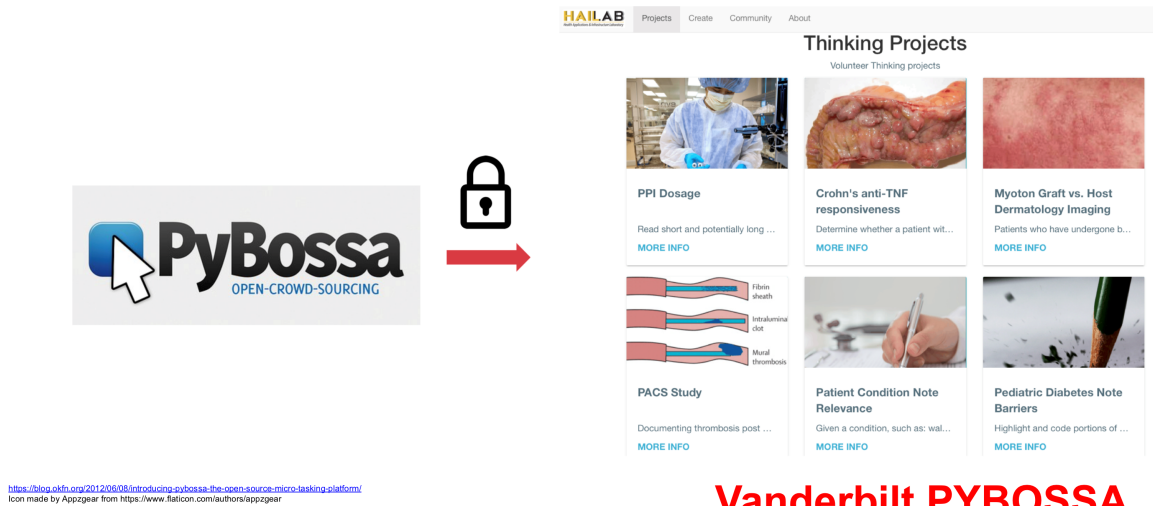


Figure 3.4: Vanderbilt PyBossa(VBOSSA) crowdsourcing platform.

The open source crowdsourcing framework, PyBossa[25] provides many basic crowd-sourcing features, such as loading and styling questions (known as a presenter), registering workers, assigning workers to tasks, collecting answers, timing tasks and extracting aggregate statistics and labels. Unfortunately, the default version of PyBossa lacks many of the privacy controls that are needed to manage sensitive data. Therefore, fine-grained access controls with two-factor authentication were added to limit access for each worker.

Moreover, we added worker attributes (or properties) to the underlying worker data models so we can categorize each crowd worker by his or her skill level and specialty. These attributes allow for fine-grained question assignment and weighting. The VBOSSA system (Figure 3.4) allows researchers to customize how questions are ‘presented’ to workers via basic HTML and JavaScript coding. These presenters are simple template HTML forms that read from an API and populate question text and candidate answers.

The VBOSSA crowdsourcing system provides multiple layers of access control for securing the sensitive medical data. First of all, the VBOSSA system was deployed on an internal server within the Vanderbilt University Medical Center firewall. The site was not open to the public. All worker registration, task assignment, question answering, and data extraction were managed through a web interface over HTTPS, and the activity is logged. Second, as shown in Figure 3.5, crowd workers need to manually authorize the login in their smartphone using the DUO application [124]. The DUO application is an online security service, which provides two-factor authentication adds a second layer of security for the user account. The DUO application prevents anyone but the crowd workers from logging in, even if other people know crowd workers’ passwords.

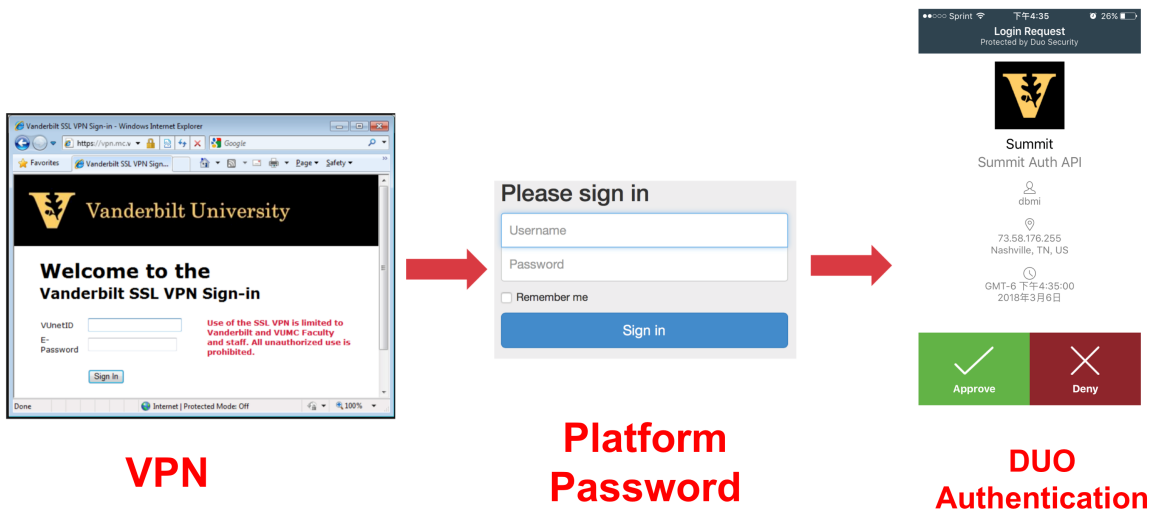


Figure 3.5: Multiple-layer access control mechanism of the VBOSSA system.

As shown in Figure 3.6, to protect the sensitive information, we first ask the researcher

to provide the Institutional Review Board (IRB) of the targeted patient cohort and the medical record number (MRN) of each patient. The crowd workers could only access documents associated with the assigned IRB. Moreover, upon querying the charts, the APIs apply open-source, de-identification tools (e.g., the MITRE Identification Scrubber Toolkit [195]), to remove or scrub HIPAA-designated identifiers, such as patient name and residential addresses.

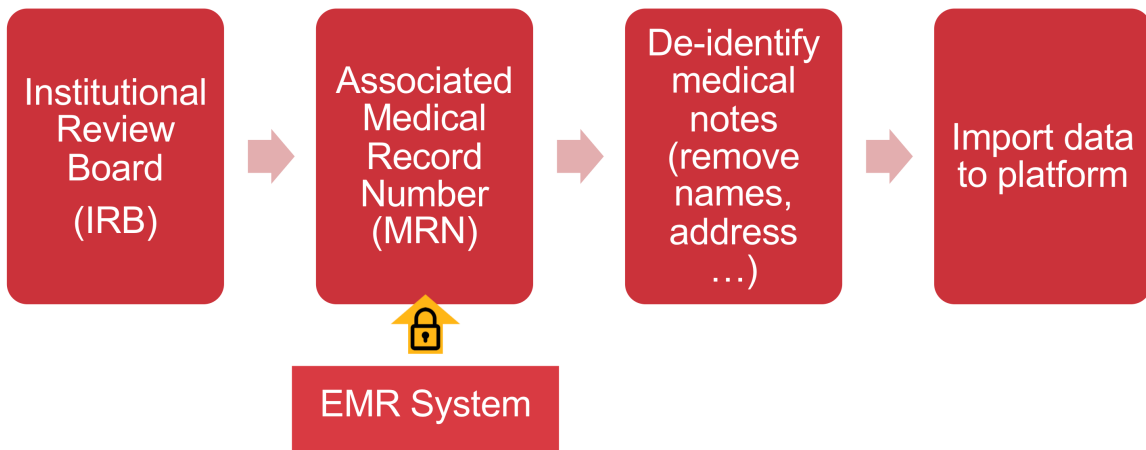


Figure 3.6: Sensitive data de-identification mechanism of the VBOSSA system.

### 3.4 Professional Workshop and Crowd Worker Pool

Before deploying a crowdsourcing project, we conduct a design workshop. The workshop includes the researcher, medical personnel, computer science researchers and anthropologists. The workshop begins by introducing the researcher to crowdsourcing preliminaries and non-healthcare crowdsourcing examples. Next, the team works to clarify and decompose the research objective into atomic questions by refining the structure of the crowdsourcing project as in Figure3.7. The workshop discusses data needs (e.g., all notes or specific note types), question format (e.g., boolean, multiple choice or text snipping), scope of tasks (e.g., multiple questions per patient or a single one), and worker skills requirements for the tasks .Crowdsourcing questions are constructed as narrowly as possible. For example:

- (a) Does a clinical note document patient conversations regarding diabetic diet alternatives? (Y/N)
- (b) Which of the following dietary alternatives were discussed with the patient? (Healthy oil choices, Sugar-free sweets, Unsweetened tea, None)
- (c) Of the patient's current diet choices listed in this note, rank them in terms of most problematic to their long-term health: (Soda, French fries, Dark chocolate, Broccoli)

In addition to true/false questions, multiple choice questions and ranking questions, researchers may also ask that workers snip (or extract) text from notes that support the answer. These snippets are extremely helpful when experts need to adjudicate disagreements.

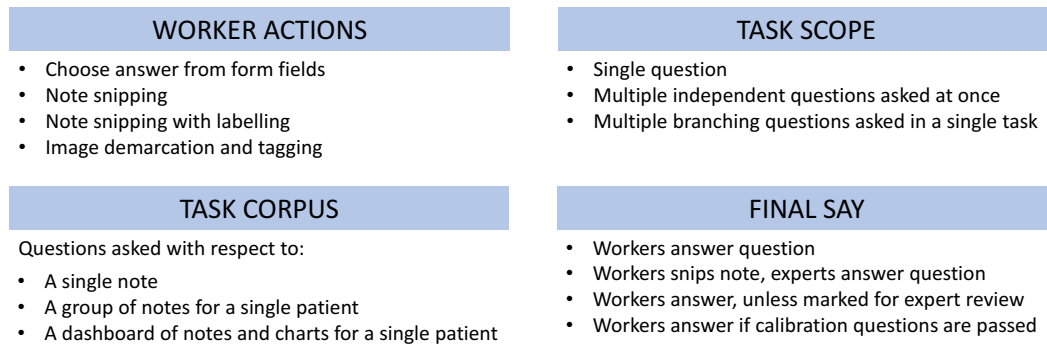


Figure 3.7: Agenda for crowdsourcing workshop with researchers.

After designing the crowdsourcing project, the researchers need to recruit workers with enough background knowledge and necessary credentials to access the data. For instance, in healthcare, only hospital employees (which includes faculty, staff, and trainees) can access medical records. While the pool of workers is limited (in contrast to the aforementioned public crowd on Amazon), there are often groups of highly motivated workers, such as medical students, who are willing to work given incentives. As shown in Figure 3.8, our worker pool consists of mostly medical students and nursing students, with a small number of faculty. These workers were recruited through Grand Round presentations and

IRB-approved email communications. For a worker to participate, he or she signed a data use agreement and, in some cases, was added as key personnel to the researcher’s IRB. We also recorded skill-level of each worker (e.g., medical student, intern, resident, fellow, attending, and nurse) and their specialization (if any), as these answers can impact which questions they are qualified to answer.

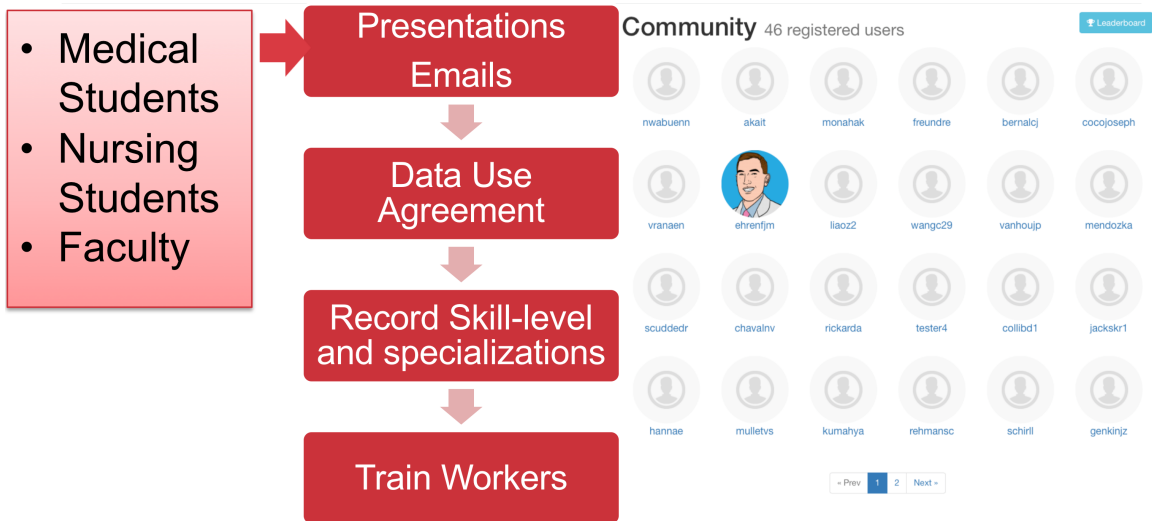


Figure 3.8: Professional crowd worker pool of the VBOSSA crowdsourcing platform.

Moreover, the medical researcher can further control the access of crowd workers in a more fine-grained way. As shown in Figure 3.9, the medical researcher can control the crowd workers to access a specific sub-study of a chart review project by setting constraints, such as ICD constraints and demographic constraints (e.g., genders and ages). The medical researcher can easily assign projects to specific workers or remove specific workers from a project.



**Administration** Back To Overview Logout

User Table **Create User** Projects

Basic Information ICD constraints Demographic constraints Lab constraints CPT constraints MRN constraints Create

RUID constraints

---

**User Name** Username

**Password**  
default: 123456

**VUID**

**Projects**

SD enable  RD enable  Single Patient Search enable  Administrator

**SD Database** FabbriSDDatab...

**SD Database** FabbriRDDatab...

**Date range constraints**  
1990-01-01 - 2015-01-01

Figure 3.9: Access control of the professional crowd worker pool of the VBOSSA crowdsourcing platform.

### 3.5 Tools to Support Crowd Workers

As shown in Figure 3.10, we design tools to support different types of crowdsourcing projects, including text search engine, text highlighting and document ranking.

Chart Review Task Type	Tool
Label Patients	Text Search Engine
Label Notes	Text Highlighting
Mark Supportive Snippets for Labels	Text Highlighting
Label Pixels in Medical Images	Pixel Brush

Figure 3.10: Customizable tools for different types of Crowdsourced chart reviews

### 3.5.1 Text Search Engine and Document Ranking

Figure 3.11 shows the search engine we developed to support chart reviews. The search engine provides two search modes:

- **Keyword Search:** Return documents that contain the search term. Rank documents by the frequency of the search term in the document;
- **Expanded Keyword Search:** Given a search term, expand the search to include terms that are semantically similar to the term, and return documents with any of the similar terms. Rank documents by a specific metric, such as the number and extent of similar terms in the document;

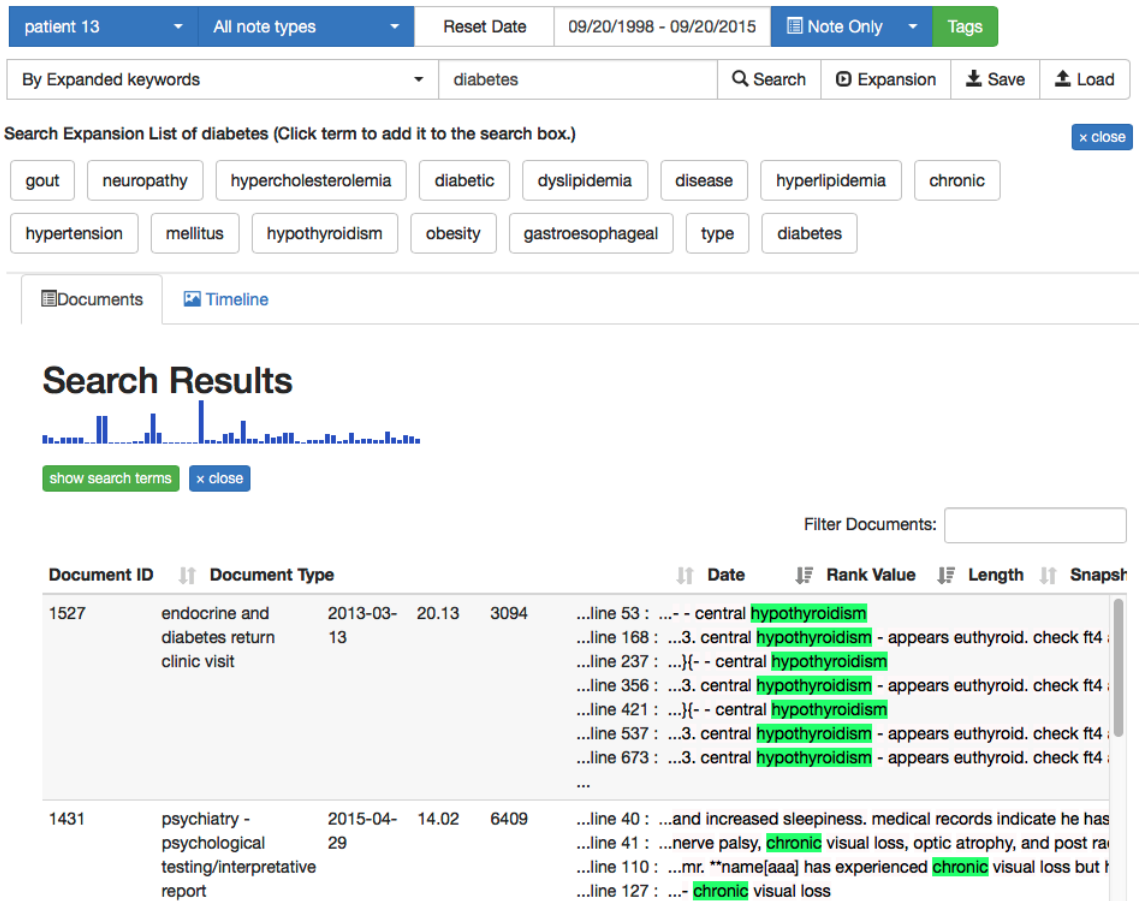
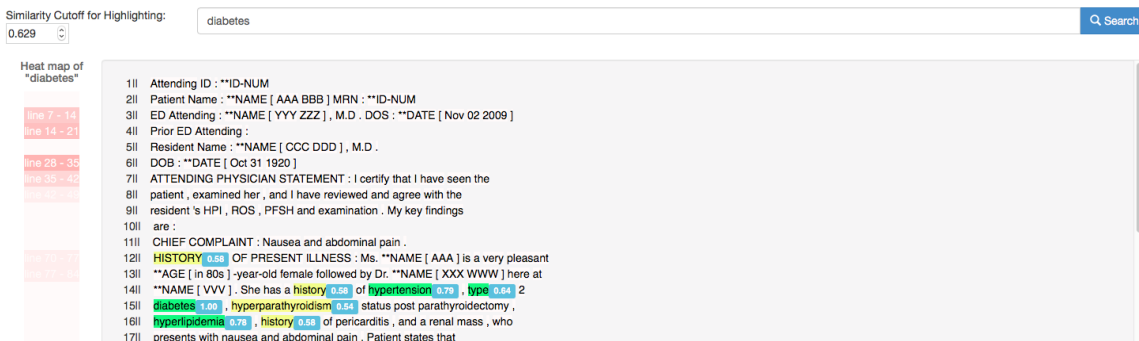


Figure 3.11: Search engine to support crowd workers

### 3.5.2 Text Highlighting



the text-highlighting tool we developed to support chart reviews. As shown in Figure 3.12, the text highlighting tool highlights “diabetes” and similar terms of “diabetes,” such as “hyperlipidemia” and “obesity.” On the left side of the note, a heat map displays the number of terms related to diabetes in each section of the note. The crowd workers can quickly navigate to the important snippets by clicking the highlighted bars in the heat map.

### 3.5.3 Result Review&Comparison Tool

Worker  Comparison

Identify the following sections: **Medications, Physical examination, Assessment/Plan** in the two medical notes.

**GENERAL HINTS:**  
 --> For Allergies: Patient summary: often called 'Adverse and Allergic Drug Reactions'  
 --> For Medical history: Patient summary: this may be a combination of 'Structured Problems' and 'Significant Procedures'  
 --> For Physical examination: Omit the vital signs  
 --> For 24-hour events: sometimes called 'interval events' - usually at the beginning of the note. This can also be called 'Subjective' in the progress notes.  
 --> For Assessment/Plan: sometimes called Impression/Plan

Identified sections in the first document: **Physical examination, Assessment/Plan, Medications,**  
 Identified sections in the right document: **Physical examination, Medications, Assessment/Plan,**

Search Document 1  Search Document 2

Medications | Review

heparin subcutaneous injection 5000 un subcut 1st stat q8h - methylprednisolone 0.9% 125 ml/hr iv infusion - ciprofloxacin inj: cipro 400 mg iv q12h - x3days - ondansetron injection 4 mg iv q6h prn - esomeprazole sodium inj : 2 po 1st now qid LABS AND DATA: PCVs stable. WBC on the low side - will cor Hgb: 13.8 PCV: 41 MPV: 8.4 Pit-

- heparin subcutaneous injection 5000 un subcut 1st stat q8h  
 - methylprednisolone inj: 30 mg iv 1st now q12h  
 - ns: sodium chloride 0.9% 125 ml/hr iv infusion  
 - ciprofloxacin inj: cipro 400 mg iv q12h  
 - metronidazole inj: flagyl 500 mg iv q8h

Figure 3.13: Result review&comparison tool to support medical researchers

The result review&comparison tool is used to support medical researchers to access and compare the chart review results provided by the crowd workers. As shown in Figure 3.13, the medical researcher can select the worker, the task, and review the decision and evidence provided by the worker. The medical researcher can update, modify or reject the result provided by the worker.

### 3.5.4 Pixel Selection Tool

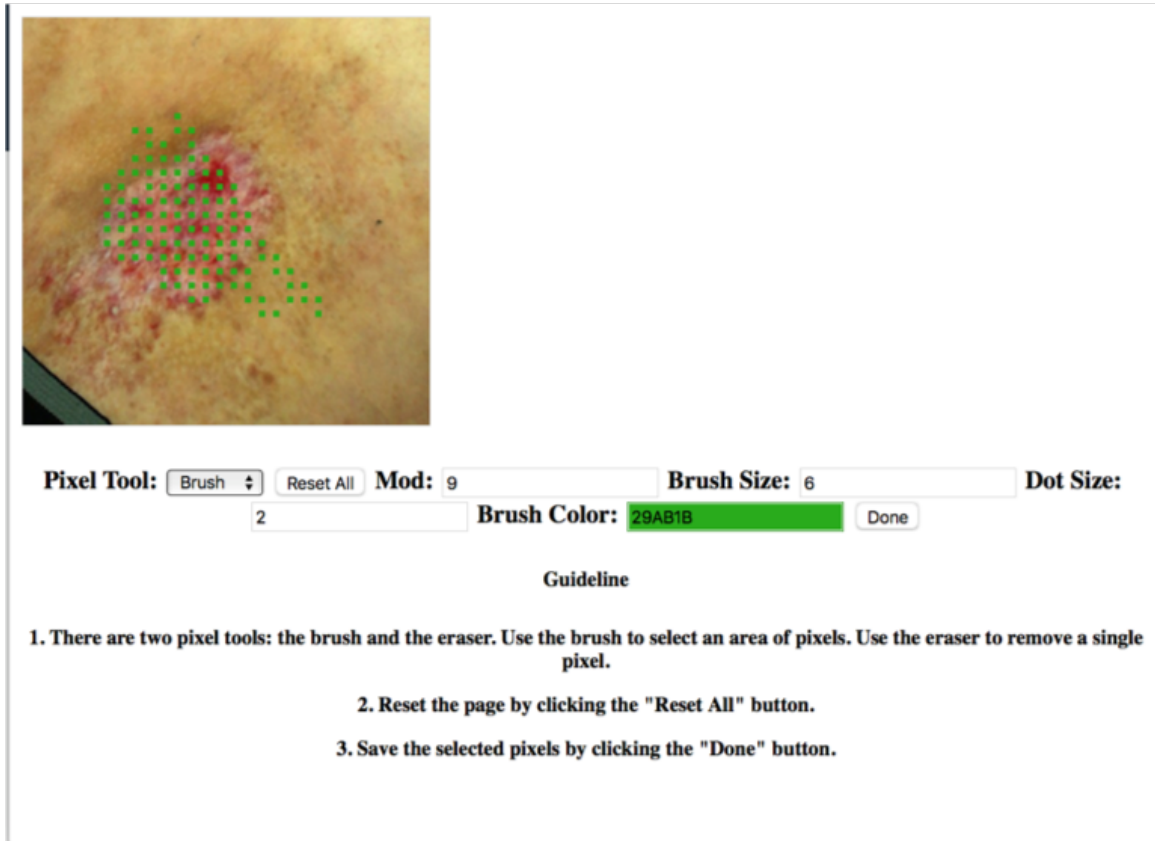


Figure 3.14: Medical image review&comparison tool to support medical researchers

The medical image review&comparison tool is used to support medical researchers to access and compare the medical image chart review results, such as important pixels in medical images that are relevant to a specific disease (e.g., GVHD), provided by the crowd workers. As shown in Figure 3.14, the medical researcher can select the worker, the task, and review the decision and evidence in a medical image provided by the worker. The medical researcher can update, modify or reject the highlighted pixels provided by the worker.

### 3.6 An Example Use Case of the Framework

In this section, we present an example use case of the framework. As shown in Figure 3.15, The framework simplifies the deploying of a chart review tasks into five steps:

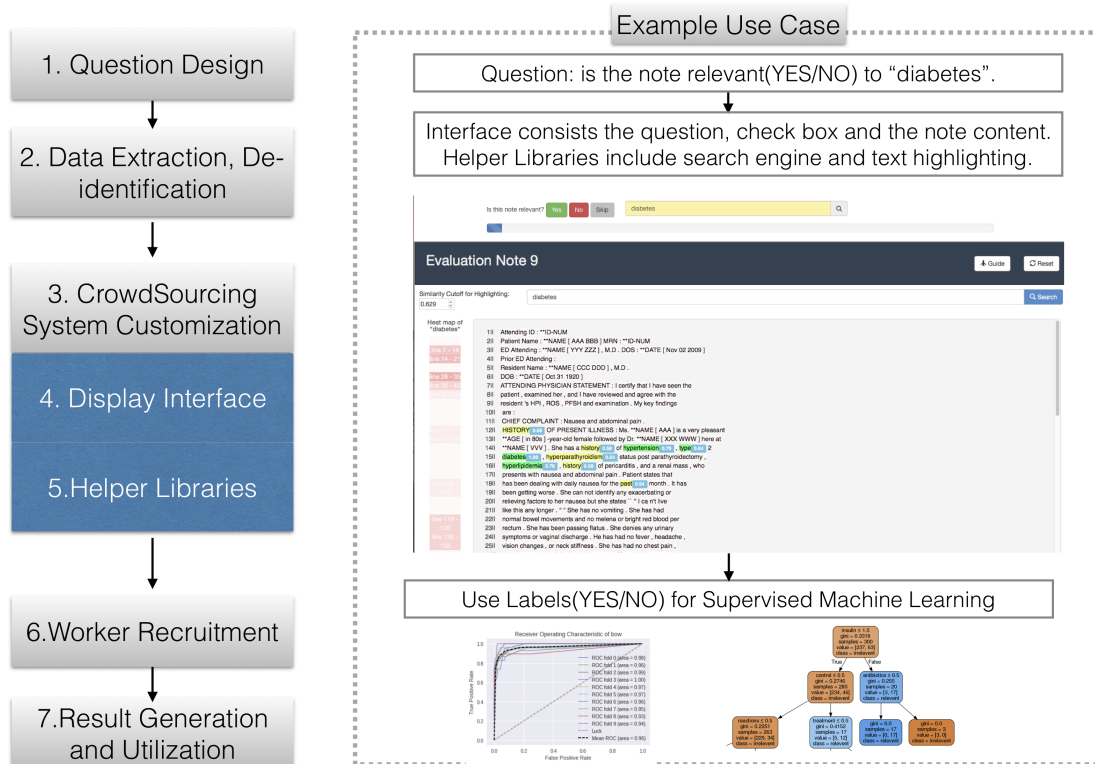


Figure 3.15: Workflow of crowdsourcing medical data sets using the VBOSSA system.

- **Design Research Question:** Suppose a researcher needs to label notes from a diabetes cohort (e.g., patients with ICD-9 code 250. \*). For each note, a worker selects one of the following labels: not relevant, relevant, or partially relevant to diabetes care. Moreover, for a note with a relevant or partially relevant label, the researcher also wants to extract supportive snippets from the note.
- **Customize Tools and De-identify Data:** After clarifying the research question, task scope, task corpus, and worker action, a presenter with a text search engine (Figure

3.16) is customized. Next, notes are extracted from the internal EMR system, de-identified and loaded into the VBOSSA system.

- **Recruit and Assign Crowd Workers:** Workers are recruited and assigned to the project. A pre-test determines if each candidate worker has sufficient knowledge about diabetes to participate. Only candidates who pass the test are admitted into the worker pool and assigned tasks.
- **Deploy Chart Review Project:** Admitted workers then begin reading notes assigned to them and producing labels. One note is shown to each worker at a time. The worker reads the content of the note, chooses a label, and selects relevant snippets from the note. This process continues until all notes are labeled. Depending on the coverage requirements, multiple workers might answer the same question.
- **Leverage Labels:** After all the tasks are completed, the researcher receives the labels and snippets. The researcher then utilizes the data in a supervised machine learning task, such as document classification.

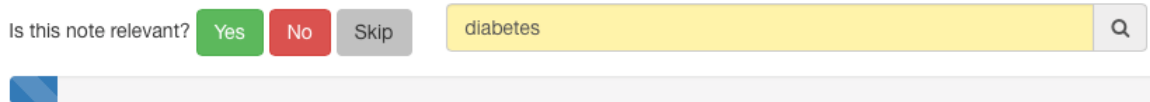


Figure 3.16: An example VBOSSA presenter with a text search engine.

### 3.7 Overview of the Finished Crowdsourced Medical Research

In this section, we briefly introduce the medical research projects supported by the crowdsourcing-based information retrieval system. The result and activity log of these medical research projects are used in the rest of this section as:

1. the evaluation datasets of advanced machine learning tools to support crowd workers;
2. the sources of user behavior analysis to identify metrics for measuring the performance of next generation of information retrieval systems for medical research;

Figure 3.17 shows the user interface for the Diabetes/Seizure Note Relevance and Patient Condition Note Relevance Chart Review project. The interface includes the search engine and text highlighting tools. The crowd workers reviewed each note and determine if notes are related to a specific medical condition, such as diabetes and seizure. In the rest of this dissertation, we construct evaluation datasets from the results of these projects for evaluating the information retrieval performance of the advanced features of search engine, such as semantic search and semantic ranking.

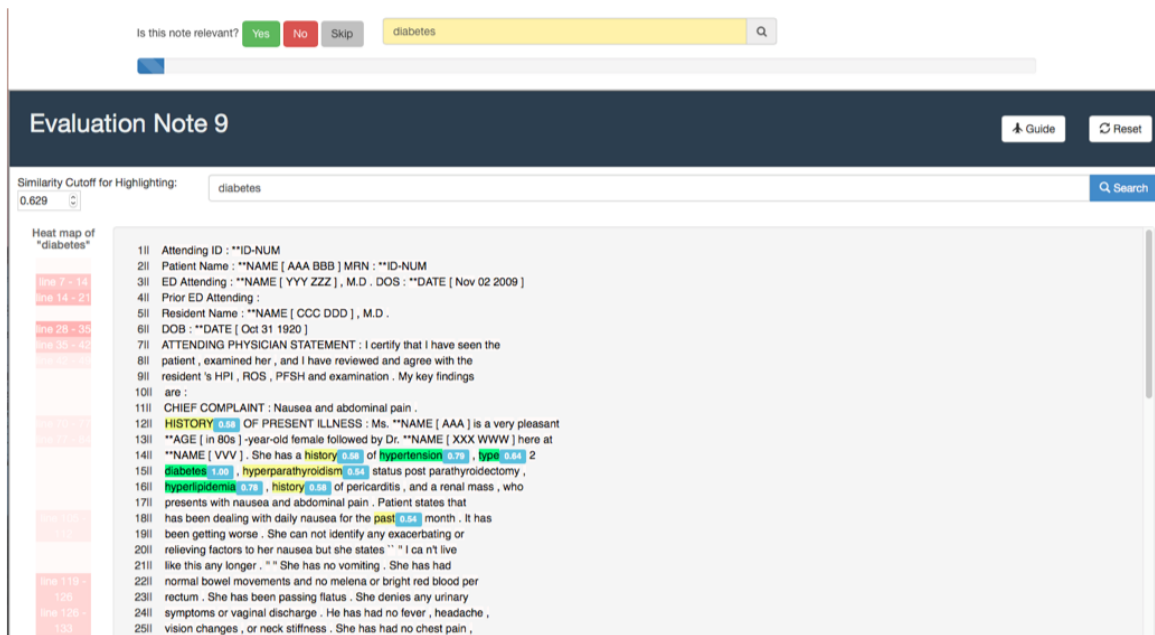


Figure 3.17: User interface of the diabetes/seizure note relevance and patient condition note relevance chart review project.

Figure 3.18 shows the user interface for the Acute Myocardial Infarction Chart Review project. In this project, the crowd workers reviewed documents and snip any portion of a note which contains references to diagnostic, medication, procedures or symptoms to AMI. In the rest of this dissertation, we construct evaluation datasets from the results of this project for evaluating the advanced features of search engine, such as text highlighting and query recommendation.



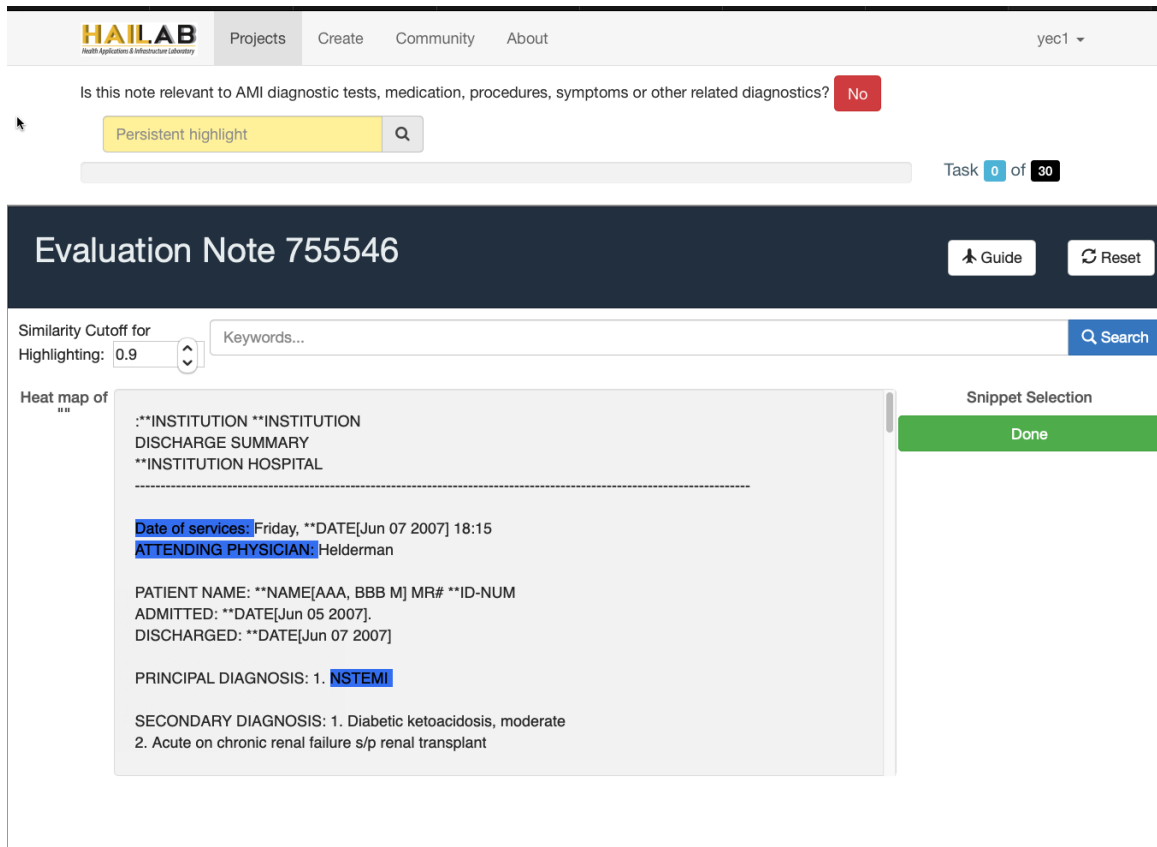


Figure 3.18: User interface of the Acute Myocardial Infarction Chart Review project.

Figure 3.19 shows the user interface for the Crohn’s Anti-TNF Responsiveness Chart Review project. In this project, the crowd workers reviewed documents and determined whether a patient with Crohn’s was clinically responsive to anti-TNF medication. In the rest of this dissertation, we construct evaluation datasets from the results of this project for evaluating the advanced features of search engine, such as text highlighting and query recommendation.

No Include patient?  mm /  yy Patient started TNF  No Needs Review?  No Additional TNF drug?  Toggle full form

Primary Clinical Response?  Primary Endoscopic Response?  Primary Histological Response?

No Was the dose increased before primary response was determined?

No Sustained Response?  No Loss of Response?

mm /  yy When was the anti-TNF drug stopped or last follow-up?

side effect(s) Was a side effect the cause of TNF drug discontinuation?

Task 1 of 983

---

patient 42151   01/01/1900 - 01/01/2021

By Keywords

### Search Results

7 (2018-11-23)

Document ID	Document Type	Date	Rank Value	Length	Snapshot
1****38	Discharge Instructions	20**-**-23	7.00	2170	...line 18 : ... <b>crohns disease</b> ...line 95 : ... <b>endoscopy lab scheduling</b> ...line 107 : ... gi-mab <b>ibd</b> annis pa ...line 196 : ... <b>endoscopy lab scheduling</b> ...line 200 : ... <b>endoscopy lab scheduling</b> on wednesday, **date[nov 23 2013] 05:00 am with provider: , ph: ...line 208 : ... gi-mab <b>ibd</b> **name[vwv] pa ...line 212 : ... gi-mab <b>ibd</b> **name[vwv] pa on tuesday, **date[nov 29 2013] 03:00 pm with provider: **name[vwv,uuu]
1****39	Clinical communication - Clinical Concerns	20**-0*-08	4.00	146	...line 0 : ...id*** (**name[yyy, xxx m]) **date[may 08 2016] 16:57: called mr **name[aaa] to discuss results from crs conference, <b>path</b>

Figure 3.19: User interface of the Crohn’s Anti-TNF responsiveness chart review project.

Figure 3.20 shows the user interface for the Anesthesiology Patients on Dialysis Chart Review project. In this project, the crowd workers reviewed documents and determined a patient has undergone dialysis between 2 weeks prior to their surgery. In the rest of this dissertation, we construct evaluation datasets from the results of this project for evaluating the advanced features of search engine, such as semantic search and semantic ranking.

Did this patient have dialysis within 2 weeks of surgery? Yes No Task 6 of 318

**Project "anes\_dialysis\_1000"** Overview

patient 1 | All note types | Reset Date: 01/01/1900 - 01/01/2020 | Note Only | Tags

By Expanded keywords: HD dialysis CRRT | Search | Expansion | Save | Load

Documents | Timeline | Single Document of patient (ID 1) | Highlight Model | close

ID: 7493 Date: 2010-04-28 Type: Nephrology Fellow Format:text

Similarity Cutoff: 0.40 | OR | HD dialysis CRRT | Highlight

Pt received large amount of PRBC during liver **transplant**, still on pressors, called to start **CRRT** for further volume removal. Range of 0 to \* mL/hr placed if further removal required please contact. Will instruct patient's nurse to call with lab results so that if adjustment of **CRRT** prescription is needed we can do so.

Search Results

show search terms | close

Filter Documents:

Document ID	Document Type	Date	Rank Value	Length
7493	Nephrology Fellow	2010-04-28	2.00	67
7494	OPERATIVE REPORT	2010-04-28	1.00	1956

Figure 3.20: User interface of the anesthesiology patients on dialysis chart review project.

Figure 3.21 shows the user interface for the Anesthesiology Patients on Student Patient Interaction Note Comparison Chart Review project. In this project, the crowd workers reviewed documents and compared analogous sections of notes. In the rest of this section, we construct evaluation datasets from the results of this project for evaluating the accuracy, agreements of the result of the crowdsourcing information retrieval system.

Worker:  Comparison:

Finished Comparison:

---

\*\*\*

Similarity Metric	sd_embedding	jaccard_auc	jaccard_hmean	edit_distance
Medications	0.53	0.03	0.0	0.24
Physical examination	0.96	0.24	0.16	0.5
Assessment/Plan	0.87	0.07	0.0	0.22

Identified sections in the first document: **Medications, Physical examination, Assessment/Plan,**  
 Identified sections in the right document: **Assessment/Plan, Physical examination, Medications,**

Search Document 1  Search Document 2

Physical examination

General: in bed; well appearing; pleasant man;

Skin: normal turgor; warm and dry; no rash; no jaundice;

HEENT: mucous membranes moist; no oropharyngeal inflammation; sclerae anicteric;

Chest: CTA; moving air well; breath sounds symmetric

CV: RRR; normal S1 & S2; 2/6 systolic murmur

Abdomen: soft, non-tender, non-distended, normal active bowel sounds, liver palpated at 2 cm below the costal margin, 1cmx3cm mobile mass palpable below liver margin

Ext: no edema;

hematuria, no urgency, no nocturia

Neurologic: +dizziness, no weakness, no syncope PHYSICAL EXAMINATION:

General:  
 pleasant man sitting in bed, no acute distress, comfortable Vital  
 Signs: P: 60 BP: 109/64 RR: 18 Tcurrent: 98.3 deg F  
 T Max: 98.3 deg F Wt: 138.03 lb Ht: 69.1 in

Physical examination

Skin: warm and dry, no rash, no jaundice Eyes: PERRL, EOMI, conjunctiva clear, sclera anicteric ENT: no oral lesions, mucous membranes moist Neck: nontender, CTA bilaterally, moving air well, breath sounds symmetric C2/6 systolic murmur, JVP 4cm Abd: soft, non-tender, non-distended, normal active bowel sounds, liver palpated at 2 cm below the costal margin, 1cmx3cm mobile mass palpable below liver margin L supraclavicular, inguinal nodes Extremities: no edema, pulses 2+ radial and 2+

results: 01/22/13 08:13: CBC:WBC: 10.3 Hgb: 13.2 PCV: 39 MPV: 11.4 Ptt-Ct: 168 RBC: 4.49 MCV: 87 MCH: 29.4 MCHC: 33.7 RDWSD: 42.7 RDW: 13.4 Differentl:NRBC: 0 NRBC#: 0.00 Basic Metab: Na: 142 K: 3.4 Cl: 100 CO2: 32 BUN: 44 Creat: 1.71 Gluc:

Figure 3.21: User interface of the student patient interaction note comparison chart review project.

The other chart review projects, such as the PACS Project (Thrombus) has the similar project structures, user interfaces and therefore, are also used for the evaluation of advanced features of search engine, such as semantic search and semantic ranking, in the rest of this dissertation.

### 3.8 Evaluation

#### 3.8.1 Overall Result

VBOSSA has been deployed within Vanderbilt University since November of 2016. Most workers are medical students. The average hours spent per chart review project is 74.2 hours with 4.8 workers. Half of the projects require a 30-60 minute training session.

The average cost per project (at \$20 per worker hour) is \$1,459, for an average of 2,066 tasks per project. Workers completed, on average, 433 tasks per project and spent 3.4 minutes per task. Despite striving to make tasks as simple as possible, task decomposition is challenging for chart-review tasks. As a result, workers often answered multiple questions at a time about each patient.

VBOSSA has been used by 18 workers to assist 10 researchers from a variety of clinical specialties answer 22,726 unique questions of varying degrees of difficulty. These workers have saved experts over 700 hours of manual chart review. Projects for which a gold standard were established had an average accuracy of 86%, while projects which had coverage greater than one worker had an average agreement between workers of 78%.

As shown in Table 3.1, crowdsourcing medical data sets have significantly reduce the time and cost for conducting a clinical chart review.

	Chart Review Task	Workers	Patients	Notes	Cost	Time (hours)
1	Acute Myocardial Infarction	3	152	200	\$810	40
2	Crohn’s Anti-TNF Responsiveness	6	983	437,993	\$3520	179
3	Pediatric Diabetes Note Barriers	6	76	210	\$1620	81
4	Anesthesiology Patients on Dialysis	2	670	49476	Free	4
5	PACS Project (Thrombus)	5	1002	7020	\$1400	70
6	Diabetes/Seizure Note Relevance	4	1000	600	\$600	30
7	Patient Condition Note Relevance	3	465	540	\$1080	54

Table 3.1: Chart review projects support by the crowdsourcing-based information retrieval system.

### 3.8.2 Impact of Instructions and Training Sessions

As we described in the previous section the projects for which a gold standard were established had an average accuracy of 86%, while projects which had coverage greater than one worker had an average agreement between workers of 78%. All these projects provide rich instructions to the workers or have 30-60 minute training sessions. To measure the importance of instructions and training sessions, we conducted a baseline project, in

which we gave the crowd worker unclear goal, limited instruction, without any training session. As shown in Table 3.2, the average accuracy is 0.58 and the average agreement is 0.50, which are much lower than the projects with rich instructions and enough training sessions. Moreover, in the Student Patient Interaction Note Comparison project, we did two training sessions and updated the instructions multiple time, and the accuracy is nearly 100% based on the review result of the medical researchers, and the agreement is around 97.66%.

<b>Task Goal</b>	<b>Average Accuracy</b>	<b>Average Agreement</b>
<b>walking</b>	0.69	0.66
<b>respiration</b>	0.43	0.47
<b>pruritus</b>	0.47	0.49
<b>epilepsy</b>	0.48	0.35
<b>fracture</b>	0.48	0.48
<b>rhinorrhea</b>	0.53	0.41
<b>breast cancer</b>	0.58	0.63
<b>Kidney</b>	0.70	0.59
<b>headache</b>	0.84	0.44
<b>Average</b>	0.58	0.50

Table 3.2: Average accuracy and average agreement in nine baseline crowdsourcing chart review projects.

### 3.9 Discussion

We have completed more than six crowdsourcing projects with the VBOSSA crowdsourcing platform. For each project, we conducted workshops, and recruited medical students and nursing students to participate in the crowd (over a dozen have participated). We paid the workers a flat fee to complete each project, which was determined by multiplying

an hourly rate times the expected number of hours of work.

For many projects, researchers have asked that workers snip the text used to make their decision. These snippets are then provided to an expert for validation. Even though this process requires an expert to review all answers, we find it is useful as the workers complete the time consuming task of scanning the entire document, while the expert simply reviews and approves snippets. If an expert's time is limited and much more costly than workers, then this design can be effective.

As the next step, we plan to design an optimization function to better design and conduct crowdsourced clinical chart reviews. First of all, we introduce the basic concepts before introducing the optimization function.

1.  $l_e$  represents the average number of labels produced by an expert per hour.
2.  $c_e$  represents the average cost of an expert per hour.
3.  $l_{e\&t}$  represents the average number of labels produced by an expert supported by tools per hour.
4.  $c_e$  represents the average cost of an expert per hour.
5.  $l_w$  represents the average number of labels produced by a worker per hour.
6.  $c_w$  represents the average cost of a worker per hour.
7.  $l_{w\&t}$  represents the average number of labels produced by a worker supported by tools per hour.

In general, we have  $c_e \geq c_w$ ,  $c_{e\&t} \geq c_{w\&t}$ ,  $l_e \geq l_w$ , and  $l_{e\&t} \geq l_{w\&t}$ .

Next, we introduce the verification time functions, which define the time to verify the labels produced by the experts ( $V_e$ ) and crowd workers ( $V_w$ ). In general, the labels and evidence produced by the expert have fewer noises compared to the labels and evidence

produced by crowd workers. Therefore, it may take less time to verify the labels produced by experts than to verify the labels produced by crowd workers (i.e.,  $v_e(l) < v_w(l)$ ).

$$V_e(L_e) = \sum_{l \in L_e} v_e(l) \quad (3.1)$$

$$V_w(L_w) = \sum_{l \in L_w} v_w(l) \quad (3.2)$$

The budget of a clinical chart reviews task  $p$  could be represented as  $(C, T)$ , the up limitation of cost and time.  $L_w$  are the labels assigned to experts and  $L_e$  are the labels assigned to workers.  $E$  is the number of experts, and  $W$  is the number of workers. We define the cost to produce and verify the labels as following.

$$C_p = C(E, W, L_w, L_e) = L_e \times \frac{l_{e\&t}}{c_{e\&t}} + L_w \times \frac{l_{w\&t}}{c_{w\&t}} \quad (3.3)$$

We define the time to produce and verify the labels as following.

$$T_p = T(E, W, L_w, L_e) = \frac{L_e}{E \times l_{e\&t}} + \frac{L_w}{W \times l_{w\&t}} + V_e(L_e) + V_w(L_w) \quad (3.4)$$

The optimization function is:

$$ArgMax_{\{E, W, L_w, L_e\}} = (C - C_p) \times (T - T_p), C \geq C_p, T \geq T_p \quad (3.5)$$

The trade-off behinds the optimization functions are:

1. The more experts we recruited, the more money we may spend since  $c_e > c_w$ . However, the time spent in producing and verifying may reduce since  $v_e < v_w$ .
2. The more workers we recruited, the less money we may spend. However, the time spent in producing and verifying may increase.
3. Given the budget  $C$  and time limitation  $T$ , we can identify the number of experts and



the number of workers we need to recruit and the number of labels we need to assign to recruited experts and workers.

### 3.10 Conclusion

In this chapter, we presented a crowdsourcing framework for sensitive medical data sets, such as electronic medical records. We developed a crowdsourcing platform that protects patient privacy and a set of helper libraries to assist workers complete tasks efficiently. Also, the user experience analysis shows up clear directions for building and evaluating supportive tools for chart reviews, such as medical search engine and text highlighting. Future extensions of the framework may include level-of-expertise weighted answers, quorum-detection, and machine learning prediction label assistance.

## Chapter 4

### CLINICALLY SIMILAR TERMS EXTRACTION

#### 4.1 Introduction

Word embeddings project semantically similar terms into nearby points in a vector space. When trained on clinical text, these embeddings can be leveraged to improve keyword search and text highlighting. In this chapter, we present methods to refine the selection process of similar terms from multiple EMR-based word embeddings, and evaluate their performance quantitatively and qualitatively across multiple chart review tasks.

To evaluate the identified similar terms across quantitative and qualitative dimensions, we conduct multiple experiments including an information retrieval evaluation, a user preference study and a timed chart review task. The results show that the identified similar terms achieved better IR performance than the baseline methods, were preferred by most users, and reduced the time to answer a question significantly. Moreover, the selection method is able to identify an optimal number of similar terms.

This work differs from previous work in two critical ways: (1) the EMR-subsets method extracts similar terms by combining multiple EMR-based word embeddings; and (2) is evaluated across multiple dimensions including information retrieval performance, user preference and time to answer a question from a chart.

#### 4.2 EMR-based word2vec embedding

A word2vec embedding projects words into a vector space by training a neural network with text [196, 197]. Word2vec embeddings can be trained with two different methods, the Continuous Bag-of-Words (CBOW) method and the skip-gram method (using a set of words vs. the position of words, respectively). Researchers have already applied the word2vec embeddings to support clinical chart reviews, such as with query expansion [31]

and search [171].

In this study, we use the CBOW method for training the EMR-based word2vec embedding, which is the default training algorithm of a word2vec model in Gensim[129]. The positions of words in the learned embedded vector space are used to estimate their similarity. Specifically, we measure the similarity of two words  $word_i$  and  $word_j$  using the cosine similarity of their embedded vectors  $v_i$  and  $v_j$ . The range of similarity is from zero to one:

$$S(word_i, word_j) = \frac{v_i \cdot v_j}{\|v_i\| \times \|v_j\|} \quad (4.1)$$

Table 4.1 lists the documents we used to train word2vec embeddings. The “Complete EMR” data set refers to all clinical notes from the Vanderbilt University Medical Center Synthetic Derivative [198], a de-identified mirror of the EMR, which contained approximately 100 million clinical notes at the time of this study. The other data sets are the largest 14 subsets of the EMR, each containing at least 1 million notes. For each dataset, we trained a word2vec model with the default parameters using the implementation provided by Gensim [199], a Python library for semantic analysis. We name each embedding with the name of its training data set, and we call any embedding trained with a subset of the EMR system an “EMR-subset embedding.”

In addition to the EMR-based embeddings, we downloaded the pre-trained word2vec model from Google News (which we refer to as the **News embedding**) [196], which contains 3 million word vectors in a 300-dimension vector space, as one of the baseline word embeddings. The News embedding has been used in prior work for query expansion [200] and identifying similar terms [18].

The preprocessing transformations applied before training the Complete EMR embedding and EMR-subset embedding include:

- (1) Parsing XML and HTML data formats to plain text using BeautifulSoup [201].
- (2) Excluding stop words, words with a length less than two characters, and words with

Training Data Set	Note Count	Vocabulary Size
Complete EMR	100m	277k
Clinical Communication	19.2m	67.0k
HP	8.0m	24.1k
Outpatient rx Order Summary	5.0m	16.2k
Prescription	4.0m	17.1k
Problem List	3.1m	6.4k
Provider Communications	2.6m	12.2k
Clinic Note	2.4m	33.2k
Respiratory Care	2.2m	3.4k
Clinic Summary	2.2m	14.7k
Clinic Summary 2	2.1m	28.2k
Rehab	1.7m	31.5k
Nurse’s Note	1.4m	16.9k
Emergency Department Nurse’s Triage	1.0m	19.3k
Letter	1.0m	26.2k

Table 4.1: Data sets used for training word2vec embeddings. Vocabulary size is the number of distinct words in the data set appearing at least 50 times

a frequency less than ten in the training data set.

- (3) Tokenizing the words using the Gensim [199] word tokenizer and lowercasing all words.

### 4.3 EMR-subsets Similar Terms Extraction Method

In this section, we describe the EMR-subsets method to extract and merge similar terms from multiple EMR-subset embeddings. The approach is motivated by the observation that embeddings created from the entire EMR can be distorted by frequently occurring text. Instead, terms should be similar to the keyword throughout subsets of the EMR. For example, the “Rehab” EMR-subset embedding identifies “ca” as a top-10 similar term for “cancer.” Similarly, the “Clinical Summary 2” EMR-subset embedding identifies “grandfather” as a similar word to “cancer”, likely because physicians document family history (We queried the complete EMR and found that 27% of the documents that contain “cancer” also contain “grandfather”, and that many of these were within five words of each other). However, the

term “cancer” is not similar to “ca” and “grandfather” in other subsets, indicating these similar terms might be biased by the text in the subset, and therefore may not be ideal for searching or highlighting clinical documents.

Similar Terms	Intra-subset Similarity	Inter-subsets Similarity	Harmonic Similarity	For Expansion
caner	0.84	0.26	0.35	Yes
metastasized	0.79	0.22	0.28	No
mesothelioma	0.77	0.29	0.35	Yes
colon	0.73	0.68	0.59	Yes
pancreatic	0.71	0.6	0.53	Yes
survivor	0.71	0.3	0.33	No
sister	0.69	0.45	0.43	Yes
hodgkins	0.69	0.39	0.39	Yes
grandfathers	0.69	0.21	0.24	No
deceased	0.68	0.48	0.44	Yes
nonhodgkin	0.68	0.19	0.22	No
greatgrandfather	0.68	0.14	0.17	No
noncancer	0.68	0	0	No
melanoma	0.67	0.86	0.62	Yes
malignant	0.67	0.76	0.58	Yes
metastasis	0.67	0.54	0.47	Yes
died	0.67	0.46	0.42	Yes
aunt	0.67	0.44	0.41	Yes
alzheimers	0.67	0.38	0.37	Yes
grandfather	0.67	0.35	0.35	Yes

Figure 4.1: Similar terms of “cancer” from the “Clinic Note” EMR-subset embedding broken down by intra-subset similarity, inter-subsets similarity, and harmonic similarity. The harmonic similarity is used for ranking terms.

The EMR-subsets method identifies similar terms of a given keyword  $w$  that have consistent similarity values across EMR subsets. As shown in Figure 4.1, three metrics are calculated to determine a similarity score for the EMR-subsets method. The **intra-subset similarity** is a term’s similarity to the keyword  $w$  using a specific subset’s embedding. The **inter-subsets similarity** is a term’s average similarity to the keyword  $w$  in all other subsets’ embeddings. The **harmonic similarity** is the harmonic mean between the intra-subset and inter-subsets similarities, which is maximized when the two similarities are equal and is zero if a term exists in a single subset.

Extracting similar terms from the EMR-subset embeddings requires multiple steps:

- (1) **Candidate Term Generation and Intra-Subset Similarity:** For a given keyword  $w$  and an EMR-subset embedding (e.g., the “Clinic Note” embedding), we generate the top- $K$  similar terms of the keyword  $w$ . The similarities of these terms define the

intra-subset similarities. The first column in Figure 4.1 lists the similar terms from the “Clinic Note” EMR-subset embedding for “cancer” including family history terms (e.g., grandfather), misspellings (e.g., caner) and organs (e.g., colon).

- (2) **Inter-Subsets Similarity:** For each candidate term  $t$  in each subset, we compute its average similarity to the keyword  $w$  (i.e., inter-subsets similarity) based on other EMR-subset embeddings (i.e., excluding the “Clinic Note” EMR-subset embedding). A candidate term that does not exist in some embeddings has a similarity of zero and lowers the inter-subsets similarity. If a candidate term only exists in the current EMR-subset embedding, we set its inter-subsets similarity to a minimum value (e.g., 0.001). The second column in Figure 4.1 lists those terms’ similarities to cancer across the other subsets - we observe that “grandfathers” has a lower similarity in other subsets, while melanoma is more similar.
- (3) **Harmonic Similarity:** For each candidate term  $t$  in each subset, we compute the harmonic mean of its intra-subset similarity and inter-subsets similarity. As shown in Figure 4.1, the inter-subsets similarity of “cancer” and “grandfathers” is 0.21, which is much lower than its intra-subset similarity. Therefore, “cancer” and “grandfathers” is only similar to each other in “clinic note” embedding, meaning it is unlikely to be included the similar term list.
- (4) **Term Cutoff:** For each subset, we apply the similarity-based cutoff method (described in detail below) to remove candidate terms with low harmonic similarities. As shown in Figure 4.1 in red, we remove some of the family terms, such as “grandfathers” and “great-grandfather,” using the similarity cutoff 0.33.
- (5) **Merge Similar Terms:** Repeat step (1)-(4) in each subset embedding and merge the similar terms by merging the similar terms extracted from each EMR-subset word embedding.

Formally, we present the process of extracting similar terms from a list of EMR-subset embeddings  $M = \{M_1, M_2, \dots, M_m\}$  for a keyword  $\mathbf{w}$  (i.e., there are  $m$  embeddings in the list, one for each note type). Given an EMR-subset embedding  $M_j$ , we define the intra-subset similarity of two words as  $S_j(w_1, w_2)$ , and the inter-subsets similarity of two words as  $I_j(w_1, w_2)$ . For each EMR-subset embedding  $M_j$ , we generate the top- $\mathbf{K}$  similar terms of the keyword  $\mathbf{w}$  as the candidate terms. We then compute the inter-subsets similarity of each candidate term:

$$I_j(t, w) = \sum_{k=1, k \neq j}^m \frac{S_k(t, w)}{m-1} \quad (4.2)$$

We then compute the harmonic similarity of each candidate term  $\mathbf{t}$ :

$$E_j(t, w) = 2 \times \frac{S_j(t, w) \times I_j(t, w)}{S_j(t, w) + I_j(t, w)} \quad (4.3)$$

Next, we remove low similarity terms provided by each EMR-subset embedding  $M_j$ , since the number of similar terms impacts the quality of search and highlighting. For example, Figure 4.2 shows that as the list of search terms is expanded from [epilepsy] to include additional terms, the relevance of the retrieved documents increase initially but then decreases as the list grows (here, relevance is defined as the percentage of highly similar terms from documents in the expanded search result).

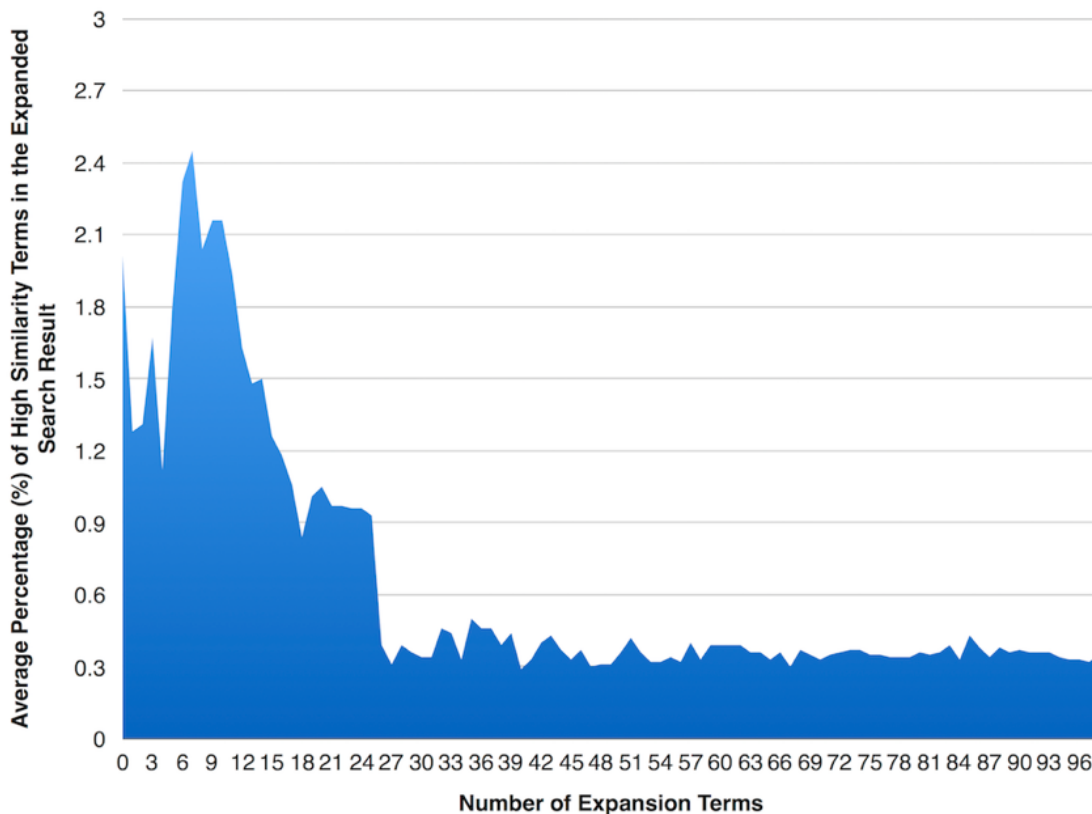


Figure 4.2: Example of expanded document quality analysis for “epilepsy.” The proportion of high similarity terms (i.e., terms that have similarities larger than 0.60 while 1.0 is the maximum value) decreases with similar term expansion.

**Cutoff Method:** The method to determine the similarity cutoff is outlined as follows. We represent the similar terms of a keyword as a two-dimension curve  $L$  (Figure 4.3), with the similar terms along the  $x$ -axis (represented by their indexes) sorted by the harmonic similarity in descending order, and their similarity values along the  $y$ -axis. We define the cutoff point as the “elbow” of the curve  $L$  because the benefit of adding more terms after this point is lower than the average benefit of choosing all terms. Formally, a cutoff point has a smoothed derivative equal to the slope of the line  $\ell$  joining the endpoints of  $L$ . Because there are irregularities in the curve  $L$  that produce multiple points with a derivative that matches the slope of  $\ell$ , we use an approximate method to identify a unique cutoff point in the curve  $L$ :



- (1) Draw a line  $\ell$  between the endpoints of L.
- (2) Calculate the minimum distance from each point in the curve L to the line  $\ell$ .
- (3) Choose the point that has the maximum distance to the line  $\ell$  as the cutoff point. The derivative of L at this point equals the slope of  $\ell$ , by the fundamental theorem of calculus.

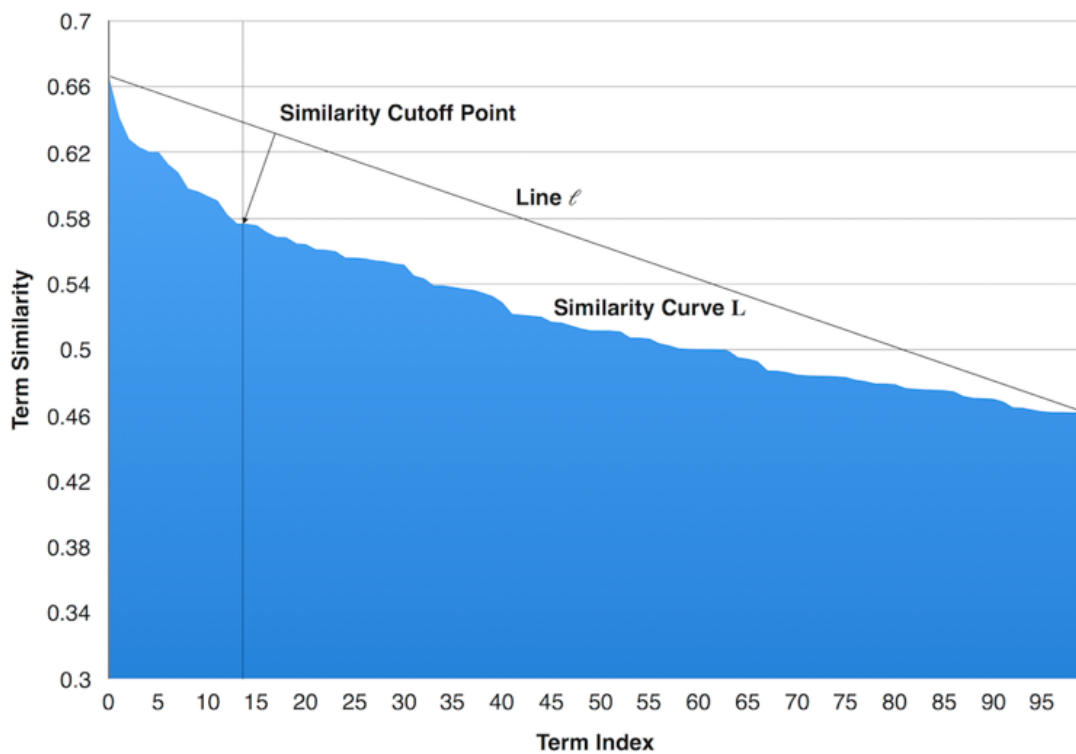


Figure 4.3: Example of similarity cutoff computation. Since all terms have similarities larger than 0.40, the y-axis starts from 0.3. Similarity cutoff is at the “elbow” of the similarity curve (arrow).

Finally, we merge the similar terms extracted from the EMR-subset embeddings as the final similar term list for the keyword  $w$ .

The results of the “elbow method” are dependent on the number of terms (i.e., the  $\mathbf{K}$ ) chosen. It is true that different  $\mathbf{K}$  values impact the curve and result in different cutoff

values. In fact, the elbow method can be used to choose the best  $\mathbf{K}$  value. Given the search terms “diabetes” and “seizure,” we tested different values for  $\mathbf{K}$  (from 1 to 1000) and the elbow method identified  $\mathbf{K} = 100$ . Larger values of  $\mathbf{K}$  did not improve results. The  $\mathbf{K}$  value may vary for different search terms.

## 4.4 Evaluation

### 4.4.1 User Preference Study

<b>(a) User types</b>		
<b>Name</b>	<b>Knowledge Level</b>	<b>Size</b>
MD	Medical Doctor Level	11
Non-MD	No Verified Level	20
<b>(b) Medical note review tasks</b>		
<b>Type</b>	<b>Keyword</b>	
General	Advil	
	Cancer	
	Fracture	
	Headache	
	Kidney	
	Ventilator	
Clinical	Walking	
	Cefuroxime	
	EEG	
	Epilepsy	
	Irrigate	
	Keppra	
	Pruritis	
Rhinorrhea		
<b>(c) Similar Term</b>		
<b>Source</b>	<b>Abbreviation</b>	
EMR word embedding	EMR	
News word embedding	News	
EMR and News word embeddings	EMR-News	
EMR-subset word embeddings	EMR-subsets	

Table 4.2: Framework of the user preference study.

We designed a user preference study to evaluate whether the extracted similar terms are preferred by users with different medical knowledge levels in various chart review scenar-

ios. We compared the selections of similar terms provided by the EMR-subsets method and the three baseline methods. As shown in Table 4.2, we recruited 11 Vanderbilt University Medical Center medical doctors (MDs) at the level of residency training or above, and 20 Non-MD Amazon Mechanical Turk [20] workers in the United States. Only the MDs have verified clinical knowledge. We chose fourteen keywords (each of which was categorized as a general or clinical term) and asked users to choose the best list of similar terms for each keyword.

Figure 4.4 shows the web page for the user preference study, which contains 14 questions asking participants to choose their preferred similar term list in a chart review task.

Instructions

- Imagine you are searching through medical records to complete a chart review. You choose an initial word to search through the medical notes (using exact match). Our tool will expand the search term to similar words to help you find relevant content. From the lists shown, which list do you prefer?

**1. Which list you prefer to expand when searching "advil" ?**

List 1  
 List 2  
 List 3  
 List 4

Similar Word List 1	Similar Word List 2	Similar Word List 3	Similar Word List 4
aleve	benadryl	acetaminofeno	colds
motrin	asprin	ibuprofeno	tylenol
tylenol	zanaflex	gargling	motrin
ibuprofen	strattera	ibuprofen	nuprin
naproxen	prevacid	acetaminiphen	sudafed
claritin	biaxin	tylenol	ibuprofen

Figure 4.4: Screenshot of the preference survey. An introduction is provided, followed by 14 questions that ask the participant to choose the best list to expand a keyword. List orders were randomized to hide source methods.

We applied multinomial logistic regression [202, 203] to analyze users’ preferences of similar terms across the extraction methods. As shown below, each logistic model takes user type (0-MD, 1-Non-MD) and task type (0-Clinical, 1-General) as the input, and outputs the log-odd ratio of choosing one method over the reference method. The null hypotheses are: (1) The user type and task type have no effect on the selection of similar terms; (2)

There is no significant preference among the similar terms provided by the EMR-subsets method and the baseline methods.

$$\ln\left(\frac{P(\text{Method})}{P(\text{BaselineMethod})}\right) = \text{Intercept} + \text{Coefficient}_u * \text{UserType} + \text{Coefficient}_t * \text{TaskType} \quad (4.4)$$

#### 4.4.2 Information Retrieval Performance

To evaluate the information retrieval (IR) performance of the EMR-subsets method, as shown in Table 4.2, we selected nine search terms from Table 4.2, including eight single-word search terms and one multiple-words search term. For each search term, we randomly selected 60 documents from patient cohorts defined by a specific ICD-9 code (Table 4.3) in which some documents contain the search term (referred as the exact-match subset), and others do not (referred as the non-exact match subset). Then we asked three medical researchers (referred to as users 1, 2, 3) to label each note’s relevance to the search term (1-relevant, 2-partially relevant or 3-irrelevant).

Search Term	ICD-9 code	Type	Number of Exact-Match Documents	Number of Non-Exact Match Documents
Breast Cancer	174.9	General	40	20
Epilepsy	345.9	Clinical	37	23
Fracture	829.0	General	38	22
Headache	784.0	General	30	30
Kidney	593.9	General	34	26
Pruritus	698.9	Clinical	26	34
Respiration	786.52	Clinical	36	24
Rhinorrhea	478.19	Clinical	31	29
Walking	719.7	General	36	24

Table 4.3: Information retrieval performance evaluation data sets

Next, for each search term and extraction method, we evaluated the P@5, P@10 and AUC scores for the various methods in different types of evaluation datasets.

1. We define the datasets extracted from patients with specific ICD-9 codes as the evaluation subsets (e.g., the evaluation subset defined by ICD-9 code 250.\*).
2. We define the combination of all evaluation subsets as the whole evaluation dataset. We define the Exact Match subset of the whole evaluation dataset as the combination of all Exact Match subsets from the evaluation subsets. We define the Non-Exact Match subset of the whole evaluation dataset as the combination of all Non-Exact Match subsets from the evaluation subsets.

For each evaluation subset and the whole evaluation dataset, we measure the P@5, P@10, and AUC score when searching and ranking by (1) keyword(s), (2) similar terms provided by the EMR-subsets method, (3) similar terms provided by the complete EMR word2vec embedding, (4) similar terms provided by the Google News word2vec embedding, and (5) similar terms provided by the EMR-News method.

The precision-at-K (P@K) is defined as the number of relevant or partially relevant notes in the top-K ranked notes. The definition of the AUC score is presented in section 2.5.4.

Notes are ranked proportionally to the number and weight of similar terms in a note. The formal equation is as follows for a keyword  $w$  and terms in a note.

$$Rank(Note) = \sum_{t \in Note} S(w, t) \quad (4.5)$$

Table 4.4 shows the average percentage of positive labels (i.e., relevant or partially relevant labels) in the exact match and non-exact match subsets of each evaluation data set. As we can see from Table 4.4, the non-exact match subsets contain non-negligible amounts of positive documents as the exact match subsets. Therefore, it is important that we develop efficient methods to identify useful documents in the non-exact match subsets.

Search Term	Average percentage of positive labeled	Average percentage of positive labeled
	Exact Match Documents	Non-Exact Match Documents
Breast Cancer	68.5%	73.6%
Epilepsy	47.7%	59.4%
Fracture	48.2%	54.5%
Headache	83.3%	65.6%
Kidney	69.6%	71.8%
Pruritus	47.4%	68.6%
Respiration	43.5%	51.4%
Rhinorrhea	52.7%	33.3%
Walking	68.5%	73.6%

Table 4.4: Distribution of positive labels in the evaluation data sets.

#### 4.4.3 Elbow Method

To evaluate the elbow method, we randomly identified 300 notes from patients in the EMR system that have an ICD-9 code for “seizure” (780.39), and another 300 notes from patients with an ICD-9 code for “diabetes” (250.\*). As a result, some notes are relevant to diabetes or seizure care, and some are not. Then we asked four medical researchers to label each note’s relevance to a disease (1-relevant, 2-partially relevant or 3-irrelevant), which produced four labeled document sets for the ‘diabetes’ cohort and four labeled document sets for the ‘seizure.’

Next, for each document set, we used “diabetes” and “seizure” as the initial queries for the respective document sets, expanded the search with the similar terms from the EMR-subsets method and evaluated the impact of the cutoff method by comparing its IR performance to three manually selected cutoff values.

#### 4.4.4 Time Efficiency Analysis

Two medical researchers, who were not investigators of this study, analyzed a cohort of 100 patients (with an average of 75 notes per patient) to determine if a patient had dialysis within 2 weeks of surgery. For each patient, the researchers answered the question YES or

NO. For half of the patients, we provided exact keyword search and highlighting to support chart review, in which notes were ranked higher proportionally to the keyword’s frequency in a note. For the other half of the patients, similar terms were used to expand the search and highlighting feature. In this case, We recorded and compared the time needed to identify the answer for the two methods. Moreover, we compared the results of medical researchers by measuring label accuracy.

## 4.5 Result

### 4.5.1 User Preference Study

<b>(a) Preference of Similar Terms</b>					
<b>Source</b>	<b>EMR</b>	<b>News</b>	<b>EMR-News</b>	<b>EMR-subsets</b>	<b>Total</b>
Total	44 (9.9%)	129 (29.0%)	39 (8.8%)	229 (52.0%)	441
<b>(b) Similar Terms Selections by User type</b>					
<b>Source</b>	<b>EMR</b>	<b>News</b>	<b>EMR-News</b>	<b>EMR-subsets</b>	<b>Total</b>
MD	15 (9.7%)	31 (20.0%)	15 (9.7%)	93 (60.0%)	154(100%)
Non-MD	29 (10.0%)	98 (34.0%)	24 (8.0%)	136 (47.0%)	287(100%)
<b>(c) Similar Term Selections by Task type</b>					
<b>Source</b>	<b>EMR</b>	<b>News</b>	<b>EMR-News</b>	<b>EMR-subsets</b>	<b>Total</b>
Clinical	26 (12.0%)	56 (25.0%)	21 (9.0%)	120 (54.0%)	223(100%)
General	18 (8.0%)	73 (33.0%)	18 (8.0%)	109 (50.0%)	218(100%)

Table 4.5: Form (a) records the overall preferences of similar terms extracted from different sources. Form (b) records the count and the percentage of selections of similar terms by User type and task type. Form (c) records the selections of each similar term extraction method.

We received 11 MDs’ and 20 Non-MDs’ response (the response rate is 100%) for a total 441 preferences (i.e.,  $31 \times 14 = 434 + 7$  multiple choices). As shown in Table 4.5, the EMR-subsets method received 52% of the selections, which is more than the other similar term extraction methods. Moreover, the selection of EMR-subsets method varies with the user type and task type.

We applied multinomial logistic regression models to analyze the result of the user preference study. As shown in Table 4.6, both the user type and task type have a significant

Index	Logistic Regression Model	Intercept	User type	Task type
1	<b>EMR vs. News</b>	-0.40	-0.51	-0.64
2	<b>EMR-News vs. News</b>	-0.50	-0.69	-0.43
3	<b>EMR-subsets vs. News</b>	1.30**	-0.78*	-0.38
4	<b>EMR-subsets vs. EMR</b>	1.70**	-0.27	0.26
5	<b>EMR-subsets vs. EMR-News</b>	1.80**	-0.09	0.06
6	<b>EMR vs. EMR-News</b>	0.09	0.18	-0.21

Table 4.6: Analysis of the impact of user type and task type on the preference of similar terms. User type (MD=0, Non-MD=1) and task type (Clinical=0, General=1) are the inputs of the multinomial logistic regression models. The significance levels are: \*\*: p-Value < 0.001, \*: p-Value < 0.05, one-tailed.

effect on user preference. Based on the intercepts and coefficients of models with indexes 3, 4, 5 in Table 4.6, we concluded that both the MD and Non-MD users prefer the similar terms provided by the EMR-subsets method compared to other baseline methods, in both the clinical and general tasks.

#### 4.5.2 Information Retrieval Performance

Data Sets	EMR-subsets	EMR	News	EMR-News	Keywords	Random
Exact & Non-Exact Match	<b>0.60</b>	0.48	0.59	0.55	0.48	0.56
Exact Match	0.57	0.60	<b>0.59</b>	0.56	0.48	0.58
Non-Exact Match	0.59	0.39**	0.37**	0.41**	0.00	0.62

Table 4.7: Average P@5 scores of each similar word extraction methods in the evaluation subsets. One-sided Mann-Whitney U test was applied to compare the P@5 scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with \*\* (p-Value < 0.001).



Data Sets	EMR-subsets	EMR	News	EMR-News	Keywords	Random
Exact & Non-Exact Match	0.56	0.46	0.50	0.55	0.50	0.60
Exact Match	0.53	0.46	0.47	0.57	0.50	0.58
Non-Exact Match	0.59	0.32**	0.19**	0.39**	0.00**	0.62

Table 4.8: Average P@10 Scores of each similar word extraction methods in the evaluation subsets. One-sided Mann-Whitney U test was applied to compare the P@10 scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with \*\* (p-Value < 0.001).

The average P@5 performances for all search terms in different evaluation datasets are shown in Table 4.7; The average P@10 performances for all search terms in different evaluation datasets are shown in Table 4.8. As we can see from Tables 4.7 and 4.8, adding similar words provided by the EMR-subsets method improves the average P@5 and P@10 results in all evaluation data sets compared to keyword-only search. Moreover, the EMR-subsets method outperforms the other extraction methods. Particularly, the EMR-subsets method significantly outperformance other methods in non-exact match subsets, which means the EMR-subsets method provides better similar words.

As shown in Table 4.9, the EMR-subsets achieved much higher AUC scores than other methods in the Exact & Non-Exact Match datasets and Non-Exact Match datasets.

Data Sets	EMR-subsets	EMR	News	EMR-News	Keywords	Random
Exact & Non-Exact Match	<b>0.65</b>	0.60**	0.60**	0.60**	0.53**	0.50**
Exact Match	0.60	0.60	0.58	0.59	0.59	0.50**
Non-Exact Match	<b>0.70</b>	0.61**	0.60**	0.61**	0.50**	0.50**

Table 4.9: Average AUC Scores of each similar word extraction methods in the evaluation subsets. One-sided Mann-Whitney U test was applied to compare the AUC scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with \*\* (p-Value < 0.001).

The average P@5 performances for all search terms in the whole evaluation dataset are

Data Sets	EMR-subsets	EMR	News	EMR-News	Keywords	Random
Exact & Non-Exact Match	0.30	0.21	0.16	0.11	0.38	0.06
Exact Match	0.39	0.23**	0.21**	0.24**	0.40	0.06**
Non-Exact Match	<b>0.19</b>	0.10**	0.02**	0.09**	0.00**	0.06**

Table 4.10: Average P@5 scores of each similar word extraction methods in the whole evaluation dataset. One-sided Mann-Whitney U test was applied to compare the P@5 scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with \*\* (p-Value < 0.001).

Data Sets	EMR-subsets	EMR	News	EMR-News	Keywords	Random
Exact & Non-Exact Match	0.30	0.18	0.17	0.18	0.34	0.07
Exact Match	0.34	0.25**	0.23**	0.23**	0.40	0.06**
Non-Exact Match	<b>0.17</b>	0.07**	0.05**	0.05**	0.00**	0.06**

Table 4.11: Average P@10 Scores of each similar word extraction methods in the whole dataset. One-sided Mann-Whitney U test was applied to compare the P@10 scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with \*\* (p-Value < 0.001).

shown in Table 4.10; The average P@10 performances for all search terms in the whole evaluation dataset are shown in Table 4.11. As we can see from Tables 4.10 and 4.11, adding similar words provided by the EMR-subsets method improves the average P@5 and P@10 results in the whole evaluation dataset compared to keyword-only search. Moreover, the EMR-subsets method outperforms other extraction methods. Particularly, the EMR-subsets method significantly outperformance other methods in non-exact match subsets, which means the EMR-subsets method provides better similar words.

As shown in Table 4.12, when being evaluated with the whole evaluation dataset, the EMR-subsets achieved much higher AUC scores compared to other methods in the Exact & Non-Exact Match datasets, Exact-Match datasets, and Non-Exact Match datasets.

Data Sets	EMR-subsets	EMR	News	EMR-News	Keywords	Random
Exact & Non-Exact Match	<b>0.72</b>	0.60**	0.59**	0.64**	0.67**	0.50**
Exact Match	<b>0.80</b>	0.65**	0.67**	0.70**	0.65**	0.50**
Non-Exact Match	<b>0.64</b>	0.53**	0.49**	0.56**	0.48**	0.50**

Table 4.12: Average AUC Scores of each similar word extraction methods in the whole evaluation dataset. One-sided Mann-Whitney U test was applied to compare the AUC scores of EMR-subsets and other methods. Methods that the EMR-subsets method significantly outperformed are marked with \*\* (p-Value < 0.001).

#### 4.5.3 Elbow Method

As shown in Table 4.13, the similarity cutoff method is able to identify an optimal similarity cutoff, which provides a better P@20 score than the manually selected similar cutoffs when using the EMR-subsets method.

Similarity Cutoff	Average P@20 when searching “diabetes.”	Average P@20 when searching “seizure.”
1.0	0.64	0.80
0.8	0.64	0.89
0.4	0.61	0.90
0.2	0.54	0.61
<b>Elbow method</b>	<b>0.68</b>	<b>0.94</b>

Table 4.13: Average P@20 scores of searching “diabetes” and “seizure” with similar words defined by different similarity cutoff.

#### 4.5.4 Time Efficiency Analysis

For the note review task, we measured the time to complete each task and the quality of labels produced by the two researchers. Ideally, the researchers would maintain their label accuracy while completing tasks faster. The result showed that the labels provided by the researchers were highly consistent. The researchers agreed on all documents except one. Table 4.14 shows the median time and the Interquartile Range (IQR) of time that each

researcher spent reviewing notes with or without highlighting similar words of the search query. We used a one-sided Mann-Whitney U test to analyze the difference in average times with and without highlighting similar words. All Mann-Whitney U test provided p-values less than 0.05, which showed that searching and highlighting similar words reduced task time.

Researcher	Median time in seconds when reviewing one patient’s notes (25th and 75th percentile time) with highlighted similar words	Median time in seconds when reviewing one patient’s notes (25th and 75th percentile time) with highlighted exact words
	1	<b>9.0 (8.0 11.0)**</b>
2	<b>76.5 (57.0 112.0)*</b>	91.5 (73.5 135.0)*

Table 4.14: The median time (25th and 75th percentile time) medical researchers spent on reviewing one patient’s notes. One-sided Mann-Whitney U test was applied for the analysis. The significance levels are: \*\*: p-Value < 0.001, \*: p-Value < 0.05 one-tailed.

## 4.6 Discussion

This chapter reports the development and evaluation of a novel similar term extraction method, the EMR-subsets method. The EMR-subsets method utilizes the subsets of an EMR system to extract similar terms that are applicable to support efficient search and consumption of clinical documents. The EMR-subsets method (i) utilized less training data, (ii) received more selections in the user preference study, (iii) achieved higher IR performance than to the baseline methods, and (iv) reduced the time needed to answer questions in a timed chart review task.

Previous research demonstrated that ensemble semantic embeddings provide better similar terms (for example, summing the similarities from multiple semantic spaces [16] or combining vectors from multiple semantic embeddings [40]). However, these methods combined embeddings trained with different data sources or attempted to learn a global embedding instead of merging the most similar terms from each subset. In this chapter, the EMR-subsets method utilized the subsets of a single data set and was preferred by users,

while the combination of the EMR and News embeddings was less preferred in the user study.

Interestingly, as shown in Table 4.15, highly similar terms for “cancer” in the Complete EMR embedding are related to family history, while the similar terms from the News embedding describe types of cancer. In contrast, the EMR-subsets method listed more clinical terms as being similar to “cancer.” One possible reason for this difference is physicians commonly document a patient’s family history of cancer in specific note types. The EMR-subsets method reduces the impact of co-occurring words from a popular note type. Therefore, the community should be careful about incorporating increasingly large data sets when training semantic embeddings for clinical applications.

<b>EMR-subsets</b>	<b>EMR-News</b>	<b>EMR</b>	<b>News</b>
melanoma	leukemia	cancern	lung cancer
breast	hashimoto	cnacer	colon cancer
prostate	malignancies	endocrinopathies	leukemia
carcinoma	nonpolyposis	at age	cancers
metastatic	diabetes	cousins	liver cancer
colon	cancer	gf	brain tumor
malignant	alzheimer	social history	brain tumors
tumor	hpth	grandfather	bladder cancer
radiation	sitosterolemia	meopausal	prostrate cancer
ca	masectomy	cance	colorectal cancer

Table 4.15: The similar terms for “cancer” provided by the EMR-subsets, EMR-News, EMR , and News similar term extraction methods.

There are several limitations and possible future work of this study. First, we limited the EMR-subsets method to the largest clinical note types in an EMR system. Future work can consider all note types or subsets constructed in alternative methods such as by common phenotypes [204]. Second, while the study attempted to discern the scenarios in which the News embedding would perform best (i.e., general note review tasks), additional analysis is needed to understand why some users preferred the similar terms provided by the News embedding in some tasks. In addition, a fine-grained information retrieval analysis is needed to determine if positive search preferences correlates with information retrieval

performance across many search scenarios (i.e., the preferred similar terms provide better information retrieval performance). Third, the constructions of user types and task types can be formalized and made more fine-grained, for example, by categorizing MD users by discipline or skill. Fourth, we utilized semantic embeddings to identify similar words, while other methods could be used find related terms like graphical models [205]. Fifth, we only included unigrams when training EMR-based embeddings in our current study. We did try word embeddings based on bi-grams or trigrams. However, bi- and trigram embeddings needed much more training data and computational resources due to the larger vocabulary space. Moreover, some bigrams have no clinical meaning, such as “table\_also.” One possible future work is extending the vocabulary with bigrams or trigrams using a clinical dictionary, such as SNOMED CT or RxNorm. Sixth, we ranked notes by the sum of term similarities. Possible future work includes normalizing the similarities before ranking and introducing other ranking methods. Moreover, as shown in Table 4.4, many notes contain the search term but were not marked as relevant, which confounds recall evaluations. Therefore, in the information retrieval experiments, we only presented the P@K scores.

## 4.7 Conclusion

This chapter presents the EMR-subsets method, which extracts similar terms from multiple semantic embeddings trained from subsets of the EMR. We systematically evaluated the similar terms extracted by the approach using qualitative and quantitative methods. Compared to the other baseline methods, the similar terms provided by the EMR-subsets method were preferred in a user preference study, achieved higher P@5 and P@10 scores across multiple search terms, and reduced the time spent searching and consuming clinical information for two researchers in a small pilot study.

## Chapter 5

### CLINICALLY SIMILAR TERM RECOMMENDATION

#### 5.1 Introduction

In Chapter 4, we present the method that extracts high-quality similar terms from a list of EMR-based word2vec embeddings to support chart reviews. EMR-based word2vec embeddings capture the semantic relationships of clinical terms by learning to predict a word from its text context (e.g., as shown in Figure 1.6 (a), “EEG” is in the text context of “epilepsy”). However, depending on the task, users might require different similar terms for a search term. For example, when searching for “epilepsy” in a chart review task that focuses on the diagnosis of “epilepsy,” users may prefer “EEG” and “brain.” However, when searching for “epilepsy” in a task that focuses on treatment, users may prefer medications, such as “Keppra” or “Vimpat.” However, methods based on word2vec embeddings may recommend all *EEG, brain, Keppra, and Vimpat* given epilepsy, which is not suitable for varying tasks.

The main observation of the requirement for clinically similar terms in different tasks can be quantified by how whom, and where those terms are used in EMRs. This is exemplified by Figure 1.6 (a) and (b), which show how context differs for “epilepsy” with respect to “EEG” and “Keppra.” Figure 5.1 shows the possibilities of five terms (“Keppra,” “EEG,” “seizures,” “epilepsy” and “Vimpat”) across some medical note types. It can be seen that “EEG” is frequently used in the Diagnosis section while “Keppra” is frequently used in the Medication section. Therefore, specific methods are needed to capture and leverage such context information to better fulfill the requirement for clinically similar terms in chart reviews.

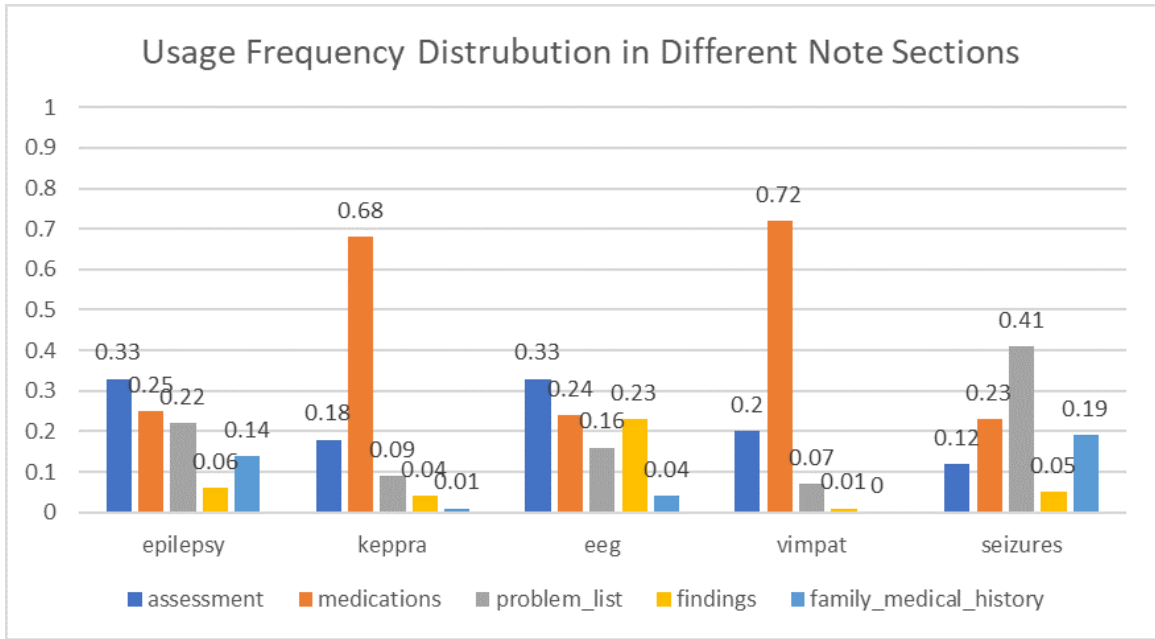


Figure 5.1: Usage frequencies of the similar Terms of “epilepsy” in different note sections.

In this chapter, we present a usage vector space, which corresponds to a collection of the usage frequencies of clinical terms in various medical usage contexts, to identify task-appropriate similar terms. We evaluate the usage vector space to predict the preferred similar terms of users across chart review tasks for acute myocardial infarction (AMI), Crohn’s disease, and diabetes, all of which have complex requirements for clinically similar terms, including terms for relevant diagnosis, medications, findings, history and so on. The results show that the usage vector space significantly boosts the performance of predicting users’ preferred similar terms, compared to baseline methods.

## 5.2 Usage Vector Space

Figure 5.2 shows the medical contexts associated with an example medical note. The note was created for a 26-year-old male patient by a physician in the Neuro-Epilepsy Department at 12:30pm on January 1, 2016. Our objective is to capture the usage information regarding how terms are used in different medical contexts.



Usage Scenario	Sub-dimension	Attribute
Medical Note Type	Note Type	Progress Note
Hospital Organizational Structure	Author Department	Neuro-Epilepsy
	Author Staff Title	Assistant Professor
Demographics	Gender	Male
	Age	26
Medical Events	Date	1/2/16 12:30
	ICD Events within 48 hours	789.39, 345.41, ...
	CPT Events within 48 hours	G0463,99214,...
	Chief Complaints within 48 hours	Seizure
Note Sections	Note Section-Header	..., Vanderbilt ... Progress ...
	Note Section-Assessment	... EEG ... Brain ...
	Note Section-Medications	... Keppra .... Vimpat ...

Figure 5.2: Usage context information of an example medical note.

We identified four types of contexts of clinical terms from the EMR system [75] of Vanderbilt University Medical Center(VUMC):

1. **Hospital Organizational Structure.** The roles of the note writers (i.e., job titles and departments) based on the hospital organization.
2. **Medical Events.** Documented diagnoses and procedures of a patient include ICD-9, ICD-10, CPT codes, and Emergency Department chief complaints.
3. **Demographics.** Patient gender (male, female and unknown) and age (quantized into ten-year bins).
4. **Medical Note Structure.** Clinical note types and sections.

We build the usage vector space from an EMR system through a series of steps:

- (1) **Preprocessing:** (1) Preprocessing: First, we extract a list of medical notes from the EMR system (e.g., all medical notes created in the year 2016). For each medical note, we extract its usage contexts (as shown in Figure 4) and filter out stop words, and words with a length less than two characters from the note.
- (2) **Initialization:** We define ten usage contexts  $C = \{C_1, C, \dots, C_{10}\}$  as shown in Table 5.1. For each clinical term  $w$ , we initialize its usage vector as **zero vector**, of which

the value in each dimension of each usage context is zero:

$$u_w = \{u_{c1}(w), u_{c2}(w), \dots, u_{c10}(w)\} = \{\vec{0}_{c1}, \vec{0}_{c2}, \dots, \vec{0}_{c10}\} \quad (5.1)$$

- (3) **Accumulation:** For each word in an extracted medical note, we update its usage vector using the usage contexts of the extracted notes. For example, given the example medical note shown in Figure 5.3, since the total count of “EEG” is 4, and the note was created by an EMR user from the “Neuro-Epilepsy” department, we add 4 to the “Neuro-Epilepsy” dimension of the “department” usage context in the usage vector of “EEG.”
- (4) **Normalization:** We repeat step (3) in each extracted medical note. We then normalize the usage counts of clinical terms in each usage context into usage frequencies (0.0 ~ 1.0) (i.e., the usage vectors). At the end of this process, each clinical term is represented as a usage vector that consists of its usage frequencies in each usage context.

Usage Context	Dimension	Value
Note Type	Progress Note	4
	Communications	0
	Clinical Summary	0
	...	0
Author Department	Neuro-Epilepsy	4
	Pediatric Neurology	0
	Internal Medicine	0
	...	0
Author Staff Title	Professor	0
	Associate Professor	0
	Assistant Professor	4
	...	0
Gender	Male	4
	Female	0
	Unknown	0
Age	0~9	0
	10~19	0
	20~29	4
	...	0
ICD Event	789	4
	345	4
	250	0
	...	0
CPT Event	G0463	4
	99214	4
	93010	0
	...	0
Chief Complaint Event	Epilepsy	0
	Seizure	4
	Fever	0
	...	0
Note Sections	Assessment	3
	Medications	1
	History	0
	...	0

Figure 5.3: Usage counts of “EEG” distributed by the medical usage contexts of the example note shown in Figure 5.2.

Usage Context	Top 3 dimensions	usage counts by usage context	usage frequencies by usage context
Note Type	history & physical - anesthesiology (pediatric anes)	4661	0.19
	progress note - neurology	3398	0.14
	administrative - prior authorization	2170	0.09
Author Department	Neuro-Epilepsy Division	18663	0.21
	In-Patient Nurse Practitioners	10812	0.12
	VUH Nurse Practitioners	6901	0.08
Author Staff Title	Clinical Fellow	20163	0.22
	Asst Professor	14512	0.16
	Assistant In, NP Acute	11885	0.13
Gender	F	12645	0.5
	M	12430	0.5
	U	3	0
Age	0~9	6660	0.27
	10~19	3636	0.14
	60~69	2777	0.11
ICD Event	780	4735	0.13
	345	4265	0.12
	V72	1792	0.05
CPT Event	80048	6282	0.04
	G0463	4678	0.03
	85027	4543	0.03
Chief Complaint Event	OBSTRUCTIVE SLEEP APNEA (ADULT) (PEDIATRIC)	248	0.07
	EPILEPSY	224	0.07
	DENTAL CARIES	192	0.06
Note Section	history_present_illness	3315	0.16
	assessment	2814	0.14
	findings	2726	0.13
Top Note Section	assessment	4838	0.33
	medications	3573	0.24
	findings	3418	0.23

Figure 5.4: Top 3 dimensions in each medical usage context of the usage vector of the clinical term “EEG.”

We define the usage frequencies of a clinical term in a context as its usage vector in that context. We define the set of usage vectors in a usage context as the usage vector space. We define the usage similarity of two clinical terms  $w_i$  and  $w_j$  in an usage context  $C_k$  as the cosine similarity of the usage vectors of  $w_i$  and  $w_j$  in the usage context  $C_k$ :

$$S_{c_k}(w_i, w_j) = \frac{u_{c_k}(w_i) \cdot u_{c_k}(w_j)}{\|u_{c_k}(w_i)\| \times \|u_{c_k}(w_j)\|} \quad (5.2)$$

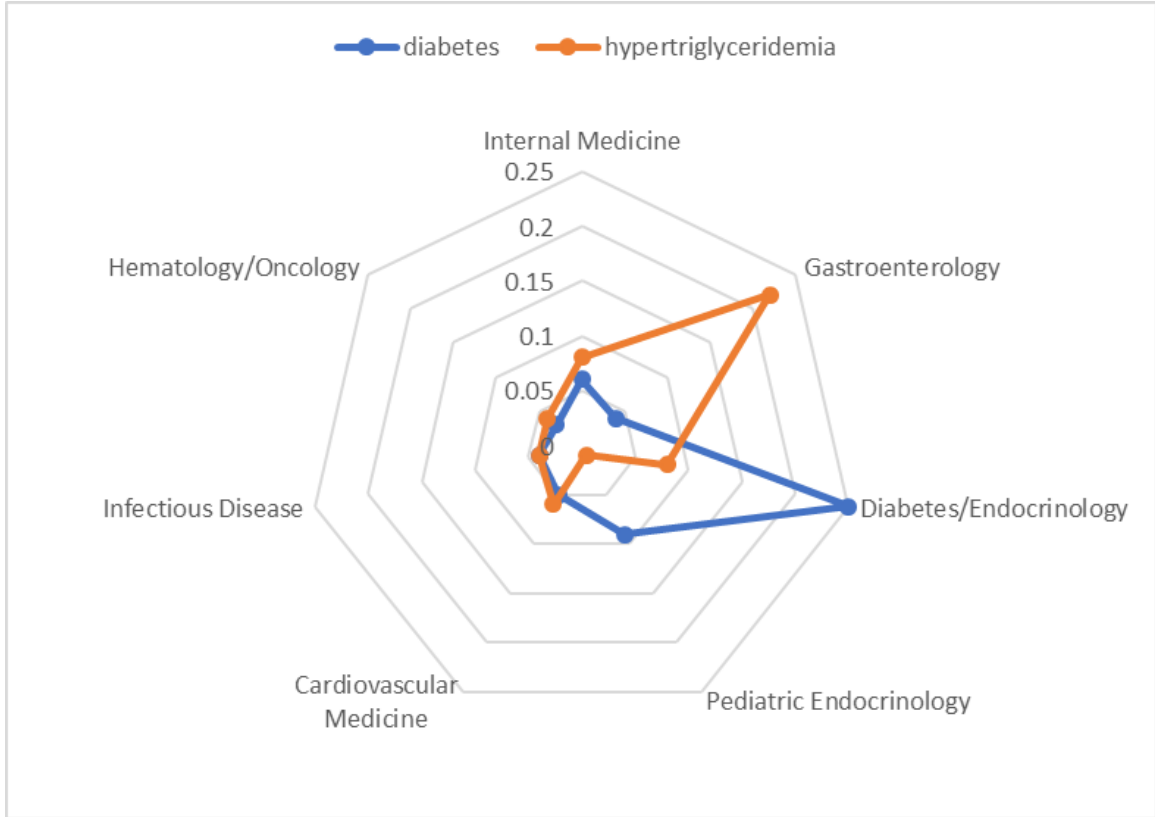


Figure 5.5: uUsage vectors of “diabetes” and “hypertriglyceridemia” in the “Department” usage context. Only the non-zero dimensions are displayed.

The usage similarity of two clinically similar terms in a medical usage context provides intuition into their semantic relationships. For example, as shown in Figure 5.5, in the department usage context, the cosine of the usage vectors of two clinically similar terms “Diabetes” and “hypertriglyceridemia” is 0.56, which suggests that they are not very similar in the department usage context since the top similarity in that context is 1.0. Therefore, if a user prefers similar terms that are invoked frequently by the “Diabetes/Endocrinology” department, then the user may not prefer “hypertriglyceridemia.”

We define the usage similarity vector of two clinical terms  $w_i$  and  $w_j$  as a vector of their usage similarities in all usage contexts:

$$S(w_i, w_j) = \{S_{c1}(w_i, w_j), S_{c2}(w_i, w_j), \dots, S_{c10}(w_i, w_j)\} \quad (5.3)$$

The usage similarity vector of two clinical terms represents their relationships in all usage contexts. For example, Figure 5.6 shows the usage similarity vector of “Diabetes” and “hypertriglyceridemia.” The usage similarities of “diabetes” and “hypertriglyceridemia” in the “Note Type,” “Department” and “Chief Complaints” contexts are much lower than in other contexts. Therefore, if a user prefers terms that have the similar distribution of usage frequencies as “diabetes” in the “Note Type,” Department and “Chief Complaints” contexts, then the user may not prefer “hypertriglyceridemia.”

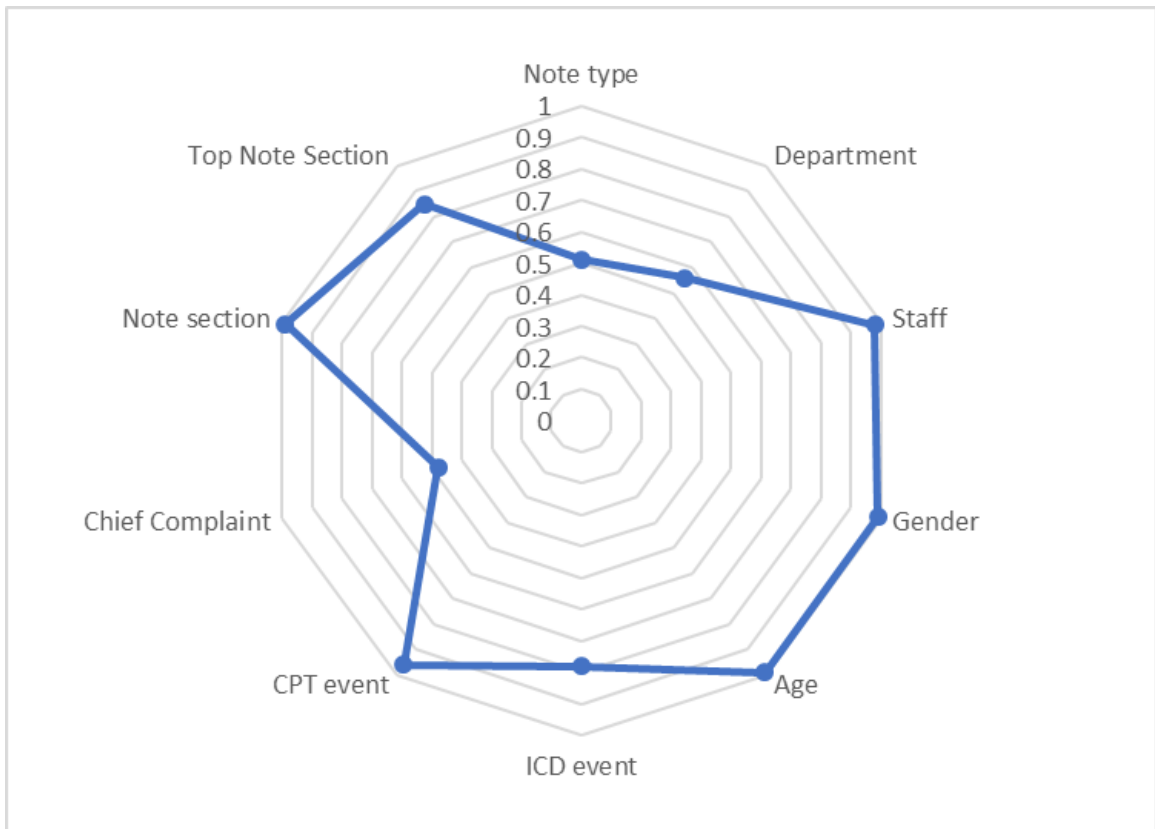


Figure 5.6: Usage similarities of “diabetes” and “hypertriglyceridemia” in all usage contexts.

## 5.3 Evaluation

### 5.3.1 Usage Vector Space

We extracted the medical records between January 1, 2016, and January 2, 2016, from the EMR system of VUMC. The usage contexts were each distributed across a set of dimensions as follows (Table 5.1):

1. **Hospital Organizational Structure.** 258 departments and 158 types of staff;
2. **Medical Events.** 957 ICD-9 codes, 6537 CPT codes and 11595 chief complaints in free-text format;
3. **Demographics.** Three patient genders (male, female and unknown) and ten age ranges (quantized into ten-year bins);
4. **Medical Note Structure.** 1514 note types; 61 note sections (defined by the headers as determined by the SecTag method [206]); five sections (“assessment,” “findings,” “family medical history,” “medications” and “problem list”) that contain the most important information in a chart review task based on our discussions with the medical researchers.

Context Type	Usage Context	Number of Dimensions
<b>Hospital Organizational Structure</b>	Departments	258
	Staffs	158
<b>Medical Events</b>	CPT events	6537
	ICD events	957
	Chief Complaint events	11595
<b>Demographics</b>	Patient Ages	10
	Patient Genders	3
<b>Medical Note Structure</b>	Note Types	1514
	Note Sections	61
	Top Five Note Sections	5

Table 5.1: Usage dimensions of clinical terms in each usage context.

### 5.3.2 Datasets

We created three evaluation datasets associated with chart review tasks (Table 5.2):

- (1) Acute Myocardial Infarction Note Relevance (referred to as the **AMI project**). This task requires researchers to highlight any portion of a note that contains references to diagnosis, medications, procedures or symptoms of AMI.
- (2) Crohn’s Anti-TNF Responsiveness (referred to as the **Crohn’s project**). This task requires researchers need to review and determine whether a patient with Crohn’s disease was clinically responsive to anti-TNF medication and highlight any portion of a note that supports the decision (i.e., Yes/No).
- (3) Pediatric Diabetes Note Barriers (referred to as the **Diabetes project**). This task requires researchers to review a list of medical notes, highlight and label portions of the notes that may be related to barriers in the documentation of diabetes plans.

Chart Review Tasks	Topic Word	Patients	Notes
<b>Acute Myocardial Infarction</b>	AMI	152	200
<b>Crohn’s Anti-TNF Responsiveness</b>	Crohn	983	437,993
<b>Pediatric Diabetes Note Barriers</b>	Diabetes	76	210

Table 5.2: Chart review tasks defined for the evaluation.

In each of the chart review tasks, the researchers searched and reviewed medical notes to identify and highlight important text snippets. Given the medical notes  $D_T$  of a chart review task T, we define the highlighted count of a clinical term  $w_{si} \in W_s$  as the sum of its highlighted counts in each document  $d_i$  from  $D_T$ :

$$H(w_{si}|D_T) = \sum_{d_i \in D_T} H(w_{si}|d_i) \quad (5.4)$$



### 5.3.3 Experiments

For each chart review task, a topic word  $K$  is chosen as the most important keyword of the research goal (e.g., “diabetes” is a topic word of the research task Pediatric Diabetes Note Barriers Problem) and serves as the basis for a clinically similar term generator. Table 5.2 shows the topic word of each chart review task.

We define the similar terms that might be preferred by the researchers of a chart review task as the candidate semantic set. A candidate semantic set could be provided by any existing similar term generator, such as EMR-based word2vec embeddings [30, 31, 39, 207] or the EMR-subsets method. We define the semantic preference of a chart review task as a subset of preferred similar terms and a subset of non-preferred similar terms from the candidate semantic set.

A supervised machine learning model learns the semantic preference from a small set of preferred similar terms and non-preferred similar terms (i.e., the training label set). The feature of a similar term  $w_t$  is its usage similarity vector based on the topic word  $K$ . The label of a similar term is based on its highlighted count and a given importance cutoff  $I$ . If the highlighted count of a similar term  $w_{si} \in W_s$  is no less than  $I$ , we label it as an important term (i.e., label = 1), otherwise, we label it as a non-important term (i.e., label = 0).

Figure 5.7 shows an example application of the usage vector space to predict the preferred similar terms of users in a chart review task. A logistic regression model is trained to weight each usage context and obtain the weights of usage context as  $W_c = W_{c1}, W_{c2}, \dots, W_{c10}$  and a threshold  $I$ . Given the usage similarity vector  $\{S_{c1}(K, t), S_{c2}(K, t), \dots, S_{c10}(K, t)\}$  of an unlabeled similar term  $t$ , if  $\sum_{k=1}^{10} W_{c_k} \cdot S_{ck}(K, t) \geq I$ , the logistic regression model then predict if an unlabeled similar term  $t$  as a preferred similar term of the user.

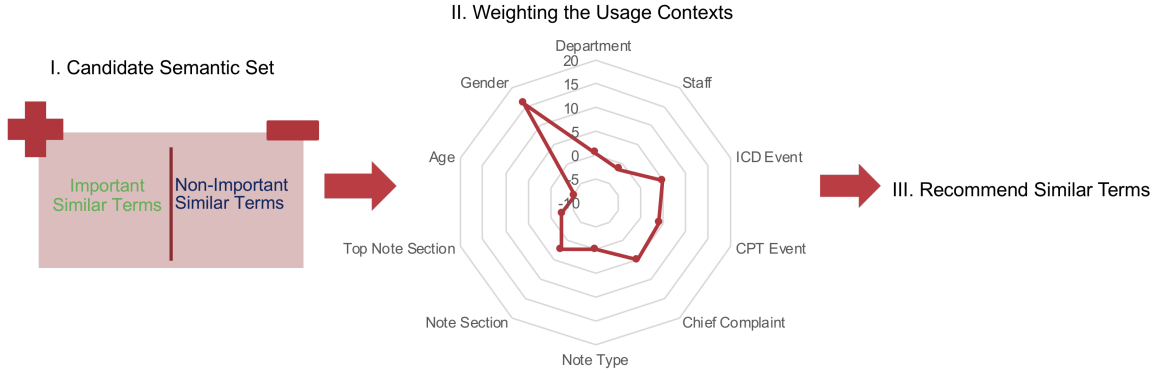


Figure 5.7: Workflow of learning and recommending clinically similar terms to users during a chart review by re-weighting the usage similarity vectors.

### 5.3.3.1 Semantic Preference Prediction Experiment

To evaluate the performance of usage vector space for predicting the semantic preference of chart reviews, we assessed how well it identified highlighted terms. This Semantic Preference Prediction Evaluation proceeds as follows:

- (1) Given an evaluation data set, we first generate a candidate semantic set  $W_s$  for its topic word using an existing similar term generator.
- (2) With the candidate semantic set  $W_s$ , we construct a label set with an importance cutoff  $\mathbf{I}=1$ . If the highlighted count of a similar term  $w_{si} \in W_s$  is greater than 1, we label it as an important term (i.e., label = 1); otherwise, we label it as a non-important term (i.e., label = 0). For each similar term  $w_{si}$  in the candidate semantic set  $W_s$ , we generate its usage similar vector  $S(w_{si}, K)$  as its feature and its word vector.
- (3) We train and evaluate a supervised machine learning model in the label set using ten-fold cross-validation. We tested three classifiers (Logistic Regression, Random Forest, and Support Vector Machine) by measuring the ROC (Receiver Operating Characteristic Curve) AUC (Area Under the Curve).
- (4) We increase the importance cutoff  $\mathbf{I}$  by 1 and repeat step (2) and (3) until the number of important terms is less than 10 in the resulting label set. Since we do ten-fold

cross-validation in the experiment, we defined 10 as the minimum number of positive labels to make sure each test fold has at least one positive label.

We repeated this process with different similar term generators. Specifically, we test three generators: 1) the EMR-subsets method [37], 2) the Completed EMR word2vec embedding [37] and 3) the Google News word2vec embedding [196]. We use the Complete EMR word2vec embedding as the baseline.

### 5.3.3.2 Learning Curve Experiment

We further assessed how the size of the training dataset influences the performance of the usage vector space. In a chart review task, the fewer labels required for learning the semantic preference, the earlier we can provide semantic support to users. To conduct this experiment, we apply the learning curve analysis [208].

The Learning Curve Analysis Evaluation proceeds as follows:

- (1) Given an evaluation data set, we first generate a candidate semantic set  $W_s$  using an existing similar term generator;
- (2) With the candidate semantic set  $W_s$ , we constructed a label set with an importance cutoff  $\mathbf{I}$ . When the highlighted count of a similar term  $w_{si} \in W_s$  is no less than  $\mathbf{I}$ , we label it as an important term (i.e., label = 1), otherwise, we label it as a non-important term (i.e., label = 0). For each similar term  $w_{si}$  in the candidate semantic set, we generate its usage similar vector  $S(w_{si}, K)$ .
- (3) Given the label set, we set  $x$  to 1% of the data points as the training set and the remaining 99% as the test set.
- (4) Train a supervised machine learning model with the training set and evaluate its AUC with the test set. Repeat step (3) and (4) 100 times and measure the AUC.
- (5) Increase  $x$  by 1% and repeat step (3) and (4) until  $x$  is greater than 90%.

- (6) Increase the importance cutoff  $I$  and repeat step (2) to (5) until the number of important terms is less than 10 in the resulting label set.

We repeat this process with the similar term generators used in the Semantic Preference Prediction Experiment.

## 5.4 Result

### 5.4.1 Semantic Preference Prediction

labelSection heading 531 The usage vector space outperformed the baseline EMR-based word2vec embedding in all evaluation datasets with all similar term generators. Table 3 shows the candidate semantic set provided by the EMR-subsets method. In the remainder of this chapter, we only show the results based on the candidate semantic sets provided by the EMR-subsets method because the usage vector space achieved a similar performance with the other similar term generators.

<b>Evaluation Dataset</b>	<b>Number of Candidate Similar Terms</b>	<b>Number of Highlighted Similar Terms</b>
<b>AMI</b>	1949	1414
<b>Crohn</b>	1204	438
<b>Diabetes</b>	1055	273

Table 5.3: Candidate semantic sets of the chart review tasks.

Figure 5.8 shows the result of the Semantic Preference Prediction evaluation using the Diabetes dataset and the candidate semantic set generated by the EMR-subsets method [15]. Table 5.4, 5.5 and 5.6 provide the detailed comparisons of the usage vector space and the baseline Complete EMR word2vec embeddings for three label sets. A one-sided Mann-Whitney U test indicated that the usage vector space statistically significantly outperformed the baseline Complete EMR word2vec embedding.

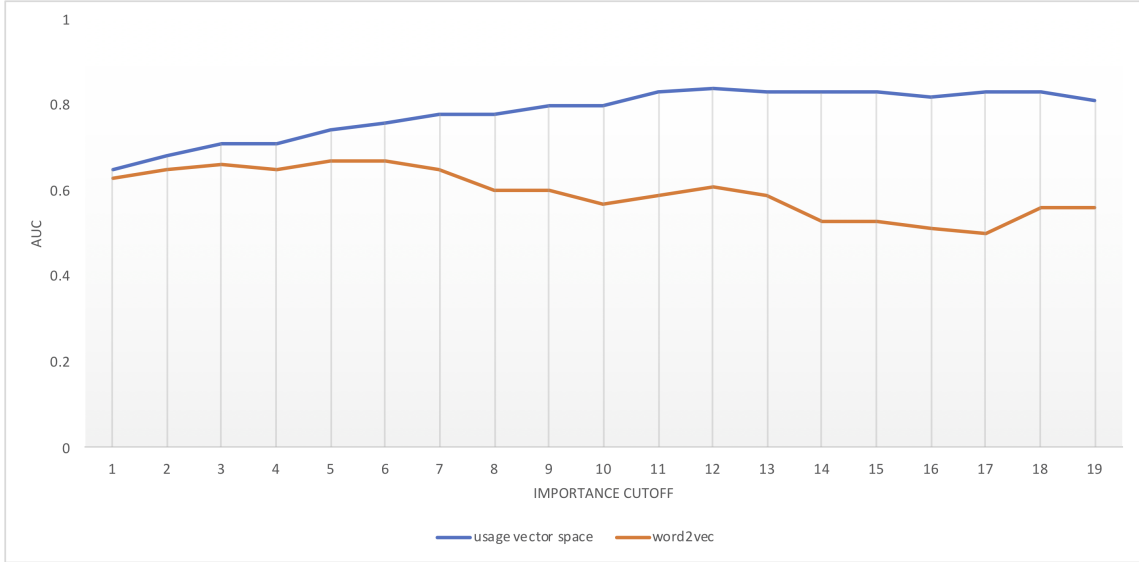


Figure 5.8: Average AUC scores (based on a 10-fold cross validation) achieved by logistic regression in the label sets constructed from the Diabetes Dataset.

Machine Learning Model	Usage vector space features	word2vec features
<b>Logistic Regression AUC</b>	<b>0.80*</b>	0.58
<b>Random Forest AUC</b>	<b>0.68*</b>	0.54
<b>Support Vector Machine AUC</b>	<b>0.78*</b>	0.57

Table 5.4: Average ROC AUC scores (10-fold cross validation) achieved by supervised machine learning models in the label set constructed from the Diabetes dataset with importance cutoff 10. One-sided Mann-Whitney U test was applied. The significance levels are: {\*\*\*: p-Value < 0.001, \*\*: p-Value < 0.01, \*: p-Value < 0.05 one-tailed}

Machine Learning Model	Usage vector space features	word2vec features
<b>Logistic Regression AUC</b>	<b>0.80**</b>	0.73
<b>Random Forest AUC</b>	<b>0.75***</b>	0.56
<b>Support Vector Machine AUC</b>	<b>0.75*</b>	0.71

Table 5.5: Average ROC AUC scores (10-fold cross validation) achieved by supervised machine learning models in the label set constructed from the AMI dataset with importance cutoff 40. One-sided Mann-Whitney U test was applied. The significance levels are: {\*\*\*: p-Value < 0.001, \*\*: p-Value < 0.01, \*: p-Value < 0.05 one-tailed}

<b>Machine Learning Model</b>	<b>Usage vector space features</b>	<b>word2vec features</b>
<b>Logistic Regression AUC</b>	<b>0.79*</b>	0.68
<b>Random Forest AUC</b>	<b>0.80***</b>	0.60
<b>Support Vector Machine AUC</b>	<b>0.79***</b>	0.68

Table 5.6: Average ROC AUC scores (10-fold cross validation) achieved by supervised machine learning models in the label set constructed from the Crohn dataset with importance cutoff 1. One-sided Mann-Whitney U test was applied. The significance levels are: {\*\*\*: p-Value < 0.001, \*\*: p-Value < 0.01, \*: p-Value < 0.05 one-tailed}

#### 5.4.2 Learning Curve Analysis

As shown in Figure 5.9, the usage vector space outperformed the EMR-based word2vec embedding in training data sets of all sizes. As shown in Figure 5.9, the usage vector space significantly reduces the number of required labels for learning the semantic preference. For example, as shown in Figure 11, with only 1% of the label set, the usage vector space reached an AUC of 0.7 while the baseline Complete EMR word2vec embedding only achieved 0.5.

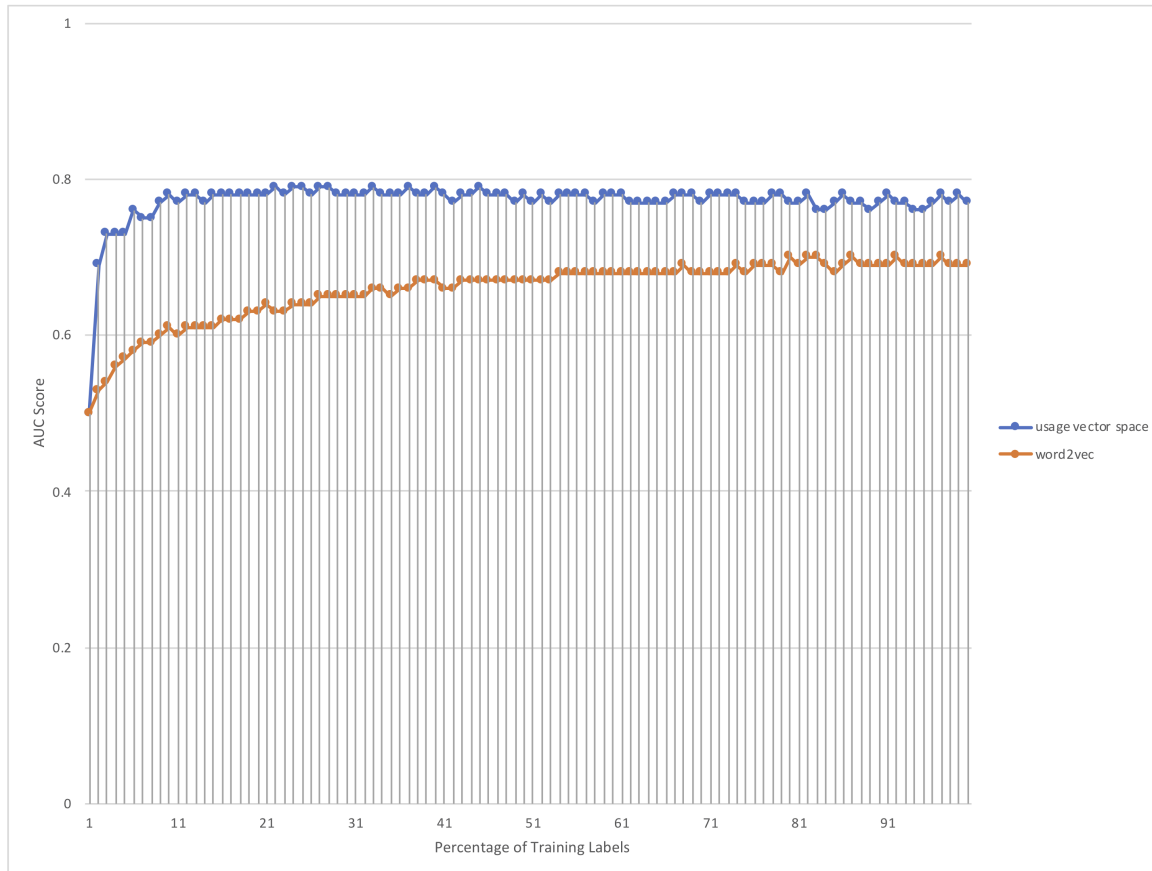


Figure 5.9: Average AUC score (based on 10-fold cross validation) achieved by the logistic regression as a function of training data set sizes in the label set constructed from the Crohn Dataset with an importance cutoff of 1.

## 5.5 Discussion

This chapter presents a novel vector space model, the usage vector space, to identify similar terms. The usage vector space is a collection of the usage frequencies of clinical terms in different medical usage contexts, which provide us with a better understanding of the relationships between clinical terms. We evaluated the usage vector space in predicting the preferred similar terms of users in three chart review tasks. The results show that the usage vector space efficiently learned the preferred similar terms of users and outperformed the baseline Complete EMR word2vec embedding. Our usage vector space outperformed baseline EMR-based word2vec embedding (e.g., AUC 0.80 vs. AUC 0.60) in all three

chart review tasks. Additionally, the usage vector space significantly reduced the number of labels (e.g., from thousands of labels to tens of labels) required to learn and predict the preferred similar terms of users.

Possible reasons that the usage vector space outperformed the baseline method are: (1) the feature space provided by the usage vector space is much smaller than the feature space provided by the word2vec embedding (i.e., 10 dimensions vs. 100 dimensions). In general, the smaller a feature space is, the less data required to train a machine learning model to achieve high prediction performance; and (2) the usage vector space pre-retrieve the semantic preference from the daily usages of an EMR system (e.g., the clinical terms used for the same CPT events). Therefore, the learning process is simplified to adjusting the weights of different usage contexts instead of learning the semantic preference from scratch.

Training data size has a significant impact on the quality of vector space models, such as the word2vec embeddings [209, 205]. To test the impact of the number of medical notes to the resulted usage vector space, we built a list of usage vector space models with different number of medical notes (experiment details not shown due to space limitations), and the results show that the structure of the usage vector space is stable (i.e., the vectors of the usage vector space have no significant change) with enough medical notes (e.g., one years' medical notes from an EMR system).

Previous research demonstrated that clinical natural language processing models (e.g., word sense disambiguation) could be trained by asking experts to provide labeled instances and contextual features [101, 210]. These methods focus on specific medical research tasks while the usage vector space demonstrates its potential in supporting multiple types of chart reviews. In addition to this benefit, the usage vector space requires the users to provide a small set of labels of clinical terms while other published methods require users to provide both labels and features.

Constructing interpretable feature space is essential for medical applications [211, 212],



especially the chart review tasks. In one pilot study, the usage vector space demonstrates its potential in providing interpretable feature space. We applied the binary logistic regression to analyze the impacts of usage contexts to users’ semantic preference of the three chart review tasks (Table 5.2):

$$\ln\left(\frac{P(Important)}{P(Less - important)}\right) = Intercept + \sum_{k=1}^{10} W_i \cdot US_{c_k}(w_s, w_t) \quad (5.5)$$

As shown in Table 5.7, 5.8, and 5.9, the “Chief Complaint” usage context has a significant positive impact on workers’ semantic preference, which means the clinical terms that are similar in describing the same chief complaint of a chart review task are preferred by the users. It is interesting that the “Gender” context (Table 5.7) had the highest significant positive impact on the semantic preference of the AMI chart review task. Since the topic word “AMI” is irrelevant to gender, terms highly relevant to gender were not preferred by the users.

<b>Index</b>	<b>Context</b>	<b>Coefficient</b>
1	<b>Intercept</b>	-16.16***
2	<b>Department</b>	0.58
3	<b>Staff</b>	1.90
4	<b>ICD Event</b>	-2.94**
5	<b>CPT Event</b>	0.37
6	<b>Chief Complaint</b>	5.75***
7	<b>Note Type</b>	5.70***
8	<b>Note Section</b>	2.00***
9	<b>Top Five Note Section</b>	1.43
10	<b>Age</b>	0.37
11	<b>Gender</b>	8.85***

Table 5.7: Binomial logistic regression analysis of the impact of usage similarity in different usage contexts on workers' preference of similar terms generated by the EMR-subsets method in the AMI project with importance cutoff 40. The significance levels are: {\*\*\*: p-Value < 0.001, \*\*: p-Value < 0.01, \*: p-Value < 0.05, one-tailed.}

<b>Index</b>	<b>Context</b>	<b>Coefficient</b>
1	<b>Intercept</b>	-13.82***
2	<b>Department</b>	0.42
3	<b>Staff</b>	0.64
4	<b>ICD Event</b>	2.18**
5	<b>CPT Event</b>	1.35*
6	<b>Chief Complaint</b>	7.24***
7	<b>Note Type</b>	-1.92*
8	<b>Note Section</b>	-0.23
9	<b>Top Five Note Section</b>	1.37*
10	<b>Age</b>	2.65***
11	<b>Gender</b>	8.03***

Table 5.8: Binomial logistic regression analysis of the impact of usage similarity in different usage contexts on workers' preference of similar terms generated by the EMR-subsets method in the Crohn project with importance cutoff 1. The significance levels are: {\*\*\*: p-Value < 0.001, \*\*: p-Value < 0.01, \*: p-Value < 0.05, one-tailed.}

<b>Index</b>	<b>Context</b>	<b>Coefficient</b>
1	<b>Intercept</b>	-22.04
2	<b>Department</b>	0.33
3	<b>Staff</b>	-1.49
4	<b>ICD Event</b>	4.89
5	<b>CPT Event</b>	4.10
6	<b>Chief Complaint</b>	4.93***
7	<b>Note Type</b>	-0.05
8	<b>Note Section</b>	2.37**
9	<b>Top Five Note Section</b>	-2.54
10	<b>Age</b>	-5.41***
11	<b>Gender</b>	15.78

Table 5.9: Binomial logistic regression analysis of the impact of usage similarity in different usage contexts on workers’ preference of similar terms generated by the EMR-subsets method in the Diabetes project with importance cutoff 10. The significance levels are: {\*\*\*: p-Value < 0.001, \*\*: p-Value < 0.01, \*: p-Value < 0.05, one-tailed.}

There are several limitations of this study that may inform future work. First, when building the usage vector space, we limited the time range for counting the medical event contexts of a medical note to 48 hours. Future work could consider a different time range for considering a medical event context or introducing Gaussian distribution to weigh the impacts of a medical event to a medical note. Second, in this study, we chose ten usage contexts when building the usage vector space. It may also be helpful to choose other types of medical usage contexts such as the medical events defined by ICD-10 codes. Third, this pilot study revealed that some usage contexts (e.g., the “CPT events” usage context) have a significant impact on users’ semantic preferences in a chart review task. Future work could incorporate a user survey to better understand why users preferred the clinical terms that are similar in specific usage contexts. Finally, as shown in Table A in the appendix,

we analyze the number of major dimensions of clinical terms (i.e., the dimensions that cover 50% of usage frequencies of a clinical term in a usage context). Table A shows that most of the clinical terms associate with a limited number of usage dimensions (e.g., most of the clinical terms are frequently used by 2-3 hospital departments). Future work can consider developing specific methods to learn fine-grained semantic preference within a usage context.

## 5.6 Conclusion

In this chapter, we present a novel vector space model, the usage vector space, to represent how clinical terms were used in different medical contexts in an EMR system. We evaluated the performance of the usage vector space in predicting the preferred similar terms of users in three chart reviews, and the result shows that the usage vector space achieves high performance (i.e.,  $AUC > 0.75$ ) in predicting users preferred similar terms and significantly outperforms baseline word2vec embedding. Most importantly, the usage vector space significantly reduced the number of labels required to learn and predict the preferred similar terms of users. Therefore, clinician's preferred similar terms could be learned faster and more accurate by introducing the usage vector space.

## Chapter 6

### DOCUMENT RANKING

#### 6.1 Introduction

In Chapter 4 and 5, we developed methods to i) extract high quality clinically similar terms from multiple EMR-based word2vec embeddings, and ii) learn the preferred clinically similar terms of users in chart reviews. Clinical similar terms are essential for an EMR search engine since they are critical to the advanced features, such as query expansion and query recommendation. As a keyword search or expanded keyword search may still return hundreds of medical notes in a chart review, ranking documents by their importance to the search goal is the essential step before returning the search result to users.

Ranking methods have been thoroughly researched and there exists many ranking methods in both research and commercial applications [213, 135, 150, 169, 214, 215, 216, 188, 190, 141, 168]. However, by analyzing the activity log of our previous chart review tasks, we noticed that specific document ranking methods are needed to better support information retrieval in chart reviews. For example, Table 6.1 shows the statistical results of four crowd workers in doing the same chart review task. In this chart review, the crowd workers may use an expanded keyword search to select and review documents. The search result is ranked by the number of similar terms of the keywords in the documents. As shown in Table 6.1, we noticed that not all crowd workers prefer to use search and ranking during chart reviews and crowd workers who do not use search and ranking may finish the tasks faster than crowd workers who used search and ranking. The analysis shows that a ranking method may interrupt workers' cognitive process (e.g., identifying the next document after reading the current one) and therefore, some workers stopped using searching and ranking in the middle of chart review tasks. Therefore, a specific metric is needed to measure the degree of a ranking method in interrupting the cognitive process of a crowd worker.

<b>worker</b>	<b>Average Search Per Patient</b>	<b>Average Number of Reviewed Documents</b>	<b>Average Time (Second)</b>
1	0.13**	24.59**	784.38**
2	0.09**	9.65**	157.82**
3	0.00**	8.64**	391.24*
4	1.10**	5.54**	1087.75**

Table 6.1: Statistical analysis of the behavior of different crowd workers in searching and reviewing documents in a chart review. Two-sided Mann-Whitney U test was applied to compare the activities of crowd workers. Results that are significantly different from worker 4 are marked with \*\* (p-Value < 0.001) and \*(p-Value < 0.05).

The other example is, as shown in Figure 6.1, there exist some critical documents that significantly impact the decision of crowd workers during chart reviews. As shown in Figure 6.1, the workers spent a significantly different time in reviewing some documents and therefore make different decisions. Therefore, if we can predict such documents in chart reviews, the system could remind the crowd workers to pay specific attention to such documents and produce more reliable labels. However, before we move on to training and applying supervised machine learning to predict such documents, we need to do more behavior analysis to better define such documents.



Figure 6.1: Time spent by two crowd workers in reviewing the same document list in a chart review task. The first worker made positive decision while the second worker made negative decision.

The above examples motivated us to answer the following question:

*“How we develop suitable ranking methods for chart reviews and what ranking metrics should we choose?”*

In this chapter, we first present the behavior analysis of users’ activities during chart reviews and then propose two novel ranking metrics, the negative guarantee ratio (NGR) and critical document. We then measure the NGR of different ranking methods. We also measure the performance of three learning-to-rank methods in predicting critical documents in chart reviews. The result shows that: i) The NGR and critical document metrics better reflect users’ need for ranking methods during chart reviews and ii) more research is needed to develop better ranking methods to support chart reviews.



## 6.2 Document Ranking Metrics

### 6.2.1 Negative Guarantee Ratio (NGR)

By analyzing the activity log of chart reviews, we noticed that in chart reviews that produce labels for patients, crowd workers don't review all the medical notes of a patient to make a decision. Instead, they first rank the documents by note types, dates, or keywords and then review the documents one by one to make a decision. After making a decision, some crowd workers keep reviewing additional medical notes to double check or confirm the decision. In this case, traditional IR performance metrics, such as precision-at-K (P@K), F1 score and AUC are not suitable to measure if a ranked document list is better than another. For example, Figure 6.2 shows two ranked document list of the same patient from a chart review task. In this example patient, to make the correct decision, a crowd worker needs to review at least 20 relevant documents. Ranking list 1 has high P@5 and P@10 score while ranking list 2 has low P@5 and P@10 scores. However, it is more time-consuming when we provide ranking list 1 to crowd workers. Given ranking list 1, the crowd worker needs to review at least 200 documents to review all 20 relevant documents to make a decision. In the contract, given ranking list 2, the crowd worker only needs to review 50 documents to make a decision. In this case, P@K metric or average precision can't reflect such a significant difference between two ranked document list in a chart review.

Index	Ranking 1	Ranking 2
0	1	1
1	1	0
2	1	0
3	1	0
4	0	1
5	1	0
6	1	0
7	1	0
8	0	1
9	1	0
...	...	...
50	0	1
...	...	...
100	1	0
...	...	...
200	1	0

Figure 6.2: Comparison of two ranked document list (1-relevant document, 0-irrelevant document). Ranking 1 list has high P@5 and P@10 scores while Ranking 2 list has low P@5 and P@10 scores. The last relevant document of ranking 1 list exists in position 200 while the last relevant document of ranking 2 list exists in position 50.

In this section, we present a novel ranking metric, called Negative Guarantee Ratio (NGR), to better measure the performance of ranking methods for chart reviews. Given a document set  $D$ , there is a small subset of documents  $d$ , which is sufficient to make a decision. Different chart review tasks may have different subsets of sufficient documents. Given a ranking method  $R$ , we rank the document set  $D$  as a list  $L_R$ , and identify a position  $C$  in the list such that all documents below the position  $C$  are not in the subset  $d$ . We define

the negative guarantee ratio as the division of the number of documents below cutoff  $C$  and the total number of documents in  $D$ .

$$NGR(L_R) = 1 - \frac{C}{|D|} \quad (6.1)$$

The higher the NGR of a ranking list is, the fewer documents a crowd worker need to review to make a decision. For example, as shown in Figure 6.2, the NGR of ranking list 1 is 0.0 while the NGR of ranking list 2 is 0.75. Therefore, ranking list 1 is better than ranking list 2, even ranking list 1 has much higher P@5 and P@10 scores compared to ranking list 2.

### 6.2.2 Critical Document

Each selected chart reviews (Table 6.3), had a small subset of patients that were reviewed by multiple crowd workers. Therefore, we can compare the time spent in the same documents and identify the documents that significantly impact workers' decision. Figure 6.1 shows the time spent in the same document list by two crowd workers. In this patient, both workers reviewed the document list by default order (i.e., without searching and ranking). Worker 1 reviewed 12 documents and made a negative decision while worker 2 review 15 document and made a positive decision. By comparing the time spent in the same document by the workers, we noticed that the workers spent significantly different time in document 0-7, 11, 12.

We assume that the more time is spent in a document by a worker, the more important the document is for making a decision. In this case, we identify document 0-7, 11, 12 as critical documents for making a decision. Another example is, Figure 6.3 shows the processes of two crowd workers in making a decision for labeling the same patient. In this case, we identify document 1, 5, 6, 7, 8 as the critical document since the workers spent similar time in other documents.

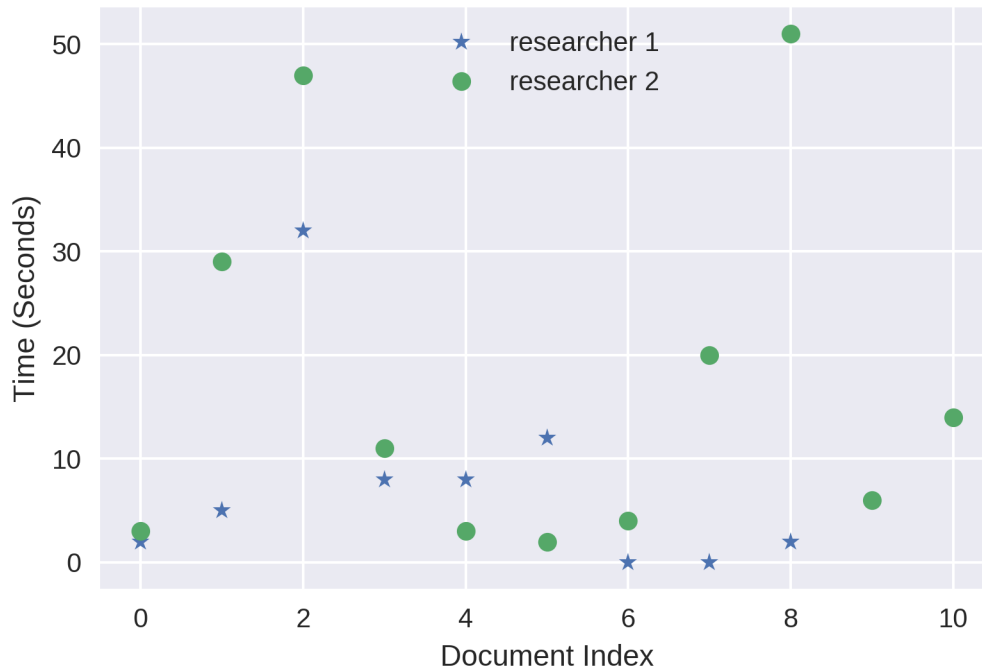


Figure 6.3: Examples of critical document in a crowdsourcing chart review project.

Formally, a critical document could be, but not be limited to:

- when two workers made the same decision but review a significantly different number of documents (e.g., 10 documents vs. 50 documents), we define the first document that two workers spent significantly different time as a critical document;
- when two workers made different decisions, the first one that two workers spent a significantly different time;

The criticality of documents can be used as labels to train supervised machine learning models and then by predicting if a document is critical or not, we may remind the crowd workers to pay specific attention to the document.

### 6.3 Document Ranking Methods

In this section, we present the ranking methods we selected to evaluate their NGR performance in chart review tasks. Table 6.2 shows the ranking methods we selected based on three intelligent sources: i) literature, ii) brainstormed ideas of the author, and iii) crowd-sourced ideas.

Index	Ranking Method	Source
1	Rank by the lengths of documents	crowds
2	Rank by the number of keywords	literature
3	Rank by the number of similar words	Brainstorm
4	Rank by the normalized number of similar words	Brainstorm
5	Rank by the usage similarity of note types to query	Brainstorm
6	Learning-to-rank based on bag-of-words (BOW)	literature
7	Learning-to-rank based on bag-of-similar-words (BOSW)	Brainstorm
8	Learning-to-rank based on word vectors (word2vec)	literature
9	Learning-to-rank based on word vectors (doc2vec)	literature
10	Learning-to-rank based on usage vectors (usage vector space)	Brainstorm

Table 6.2: Ranking and Learning-to-Rank methods defined for the evaluation.

We define the ranking methods shown in Table 6.2 as following:

1. For a document list  $D = \{d_1, d_2, \dots, d_n\}$ , we rank documents by their lengths:  $length(d_i)$ ;
2. For a document list  $D = \{d_1, d_2, \dots, d_n\}$ , we rank documents by the number of keyword  $w$  in the document  $d_i$ :  $Count(d_i, w)$ .
3. For a document list  $D = \{d_1, d_2, \dots, d_n\}$  and a keyword  $w$ , we first identify the similar words of  $w$ ,  $S(w) = \{s_1, s_2, \dots, s_m\}$  using a similar term generator such as the EMR-subsets method, and then rank document  $d_i$  using the number of similar terms in the document:  $S(d_i) = Count(d_i, w_{ij}), \forall w_{ij} \in S(w)$ .
4. For a document list  $D = \{d_1, d_2, \dots, d_n\}$  and a keyword  $w$ , we first identify the similar words of  $w$ ,  $S(w) = \{s_1, s_2, \dots, s_m\}$  using a similar term generator such as the EMR-subsets method [37], and then rank document  $d_i$  using the normalized number of similar terms in the document:  $S(d_i) = \frac{Count(d_i, w_{ij})}{|d_i|}, \forall w_{ij} \in S(w)$ .

5. For a document list  $D = \{d_1, d_2, \dots, d_n\}$  and a keyword  $w$ , we first identify the similar words of  $w$ ,  $US(w) = \{us_1, us_2, \dots, us_m\}$  using the usage similar term generator based on the usage vector space, and then rank document  $d_i$  using the number of similar terms in the document:  $S(d_i) = \text{Count}(d_i, w_{ij}), \forall w_{ij} \in US(w)$ .
6. For a document list  $D = \{d_1, d_2, \dots, d_n\}$ , we first represent each document  $d_i$  using the bag-of-words (BOW) features and then ask the user to provide some labelled documents and learn to re-rank the rest documents.
7. For a document list  $D = \{d_1, d_2, \dots, d_n\}$ , we first represent each document  $d_i$  using the bag-of-similar-words (BOSW) features and then ask the user to provide some labelled documents and learn to re-rank the rest documents.
8. For a document list  $D = \{d_1, d_2, \dots, d_n\}$ , we first represent each document  $d_i$  using the average word vectors [217] of all the words in  $d_i$  and then ask the user to provide some labelled documents and learn to re-rank the rest documents.
9. For a document list  $D = d_1, d_2, \dots, d_n$ , we first represent each document  $d_i$  using the document vectors [161] of  $d_i$  and then ask the user to provide some labelled documents and learn to re-rank the rest documents.
10. For a document list  $D = \{d_1, d_2, \dots, d_n\}$ , we first represent each document  $d_i$  using the average usage vectors of all the words in  $d_i$  and then ask the user to provide some labelled documents and learn to re-rank the rest documents.

## 6.4 Evaluation

### 6.4.1 Chart Reviews

Table 6.3 shows the chart review projects we selected for evaluating the ranking methods.

Index	Chart Review Task	Workers	Patients	Notes
1	Acute Myocardial Infarction	3	152	200
2	Crohn’s Anti-TNF Responsiveness	6	983	437,993
3	Pediatric Diabetes Note Barriers	6	76	210
4	Anesthesiology Patients on Dialysis	2	670	49476
5	PACS Project (Thrombus)	5	1002	7020

Table 6.3: Chart review tasks selected for evaluating the ranking methods.

- Acute Myocardial Infarction Note Relevance (referred to as the AMI project). This task requires workers to highlight any portion of a note that contains references to diagnosis, medications, procedures or symptoms of AMI.
- Crohn’s Anti-TNF Responsiveness (referred to as the Crohn’s project). This task requires workers need to review and determine whether a patient with Crohn’s disease was clinically responsive to anti-TNF medication.
- Pediatric Diabetes Note Barriers (referred to as the Diabetes project). This task requires workers to review a list of medical notes, highlight and label portions of the notes that may be related to barriers in the documentation of diabetes plans.
- Anesthesiology Patients on Dialysis (referred to as the Dialysis project). This task requires workers to determine whether a patient has undergone dialysis between 2 weeks before their surgery.
- PACS Project (Thrombus), (referred to as the PACS project). This task requires workers to review documents and identify thrombosis post pediatric surgery and later determine time range when central lines are inserted in patients.

#### 6.4.2 Evaluation datasets for NGR analysis

For each chart review project, we first built a document set  $D$  with size  $|D|$  (e.g., 200 documents). Then we extract  $X\%$  (e.g., 10%) positive samples and  $1-X\%$  negative samples

from the label sets of each crowd worker who participated in the projects. We then ranked the document set and then measured the NGR score of the ranked document list. We tested a serial of document size  $|D|$  and positive ratio  $X\%$  (from 1% to 90%).

### 6.4.3 Evaluation datasets for Critical Document Analysis

For each chart review project in Table 6.3, we first extracted documents that are reviewed by at least two crowd workers. Then we identified a critical document as the one that was reviewed by two crowd workers with a minimum time difference as 30, 60 and 90 seconds.

Table 6.4 shows the evaluation datasets we constructed from the Crohn chart review project for evaluating the critical document prediction.

<b>Index</b>	<b>Minimum Time Difference (Seconds)</b>	<b>Critical Document</b>	<b>Non-Critical Document</b>
<b>1</b>	<b>30</b>	647	2800
<b>2</b>	<b>60</b>	244	3203
3	90	119	3328

Table 6.4: Datasets for evaluating the critical document prediction of learning-to-rank methods

We selected five types of document features:

1. The bag-of-words (BOW) [213]. A document is represented as the bag (i.e., a set) of its words, without considering the word order in the document.
2. The bag-of-similar-words (BOSW). The BOSW feature is similar to the BOW feature. A document is represented as the bag of similar words in the document, provided by the EMR-subsets method.
3. The average word vector of a document. We first transform the words of a document into vectors using the word vectors provided by the Complete EMR word2vec



embedding (Chapter 4). Then we compute the average word vector to represent the document.

4. The average usage vector of a document. We first transform the words of a document into vectors, using the usage vectors provided by the usage vector space (Chapter 5). Then we compute the average usage vector to represent the document.
5. The document vector [218] based on the doc2vec embedding trained with all the medical notes from the SD EMR system.

We tested three classifiers (Logistic Regression, Random Forest, and Support Vector Machine) by measuring the ROC (Receiver Operating Characteristic Curve) AUC (Area Under the Curve) in predicting if a document is critical to the chart review or not.

To better understand the difficulty in predicting the critical documents, we selected the accessing of documents as the baseline labels (i.e., access=1, ignore=0).

## 6.5 Result

### 6.5.1 Negative Guarantee Ratio

The result (Table 6.5, 6.6, 6.7) shows that a ranking method with high traditional IR performance such as P@10 does not guarantee a high NGR score. A ranking method with low traditional IR performance such as P@10 may have a high NGR score. For example, as shown in Table 6.6, ranking by the number of similar words provides high P@10 score, 0.76, but has the lowest NGR score, 0.0. In this case, a crowd worker needs to review all 200 documents to make a decision. In the contract, the other methods, such as ranking by the length, and ranking by the normalized number of similar words, have relatively low P@10 score, 0.42, but have relatively high NGR score, 0.17, which means a crowd worker only need to review 166 documents to make a decision.

<b>Index</b>	<b>Ranking Method</b>	<b>NGR</b>	<b>P@10</b>
<b>1</b>	<b>Rank by the number of keywords</b>	0.15	0.20
<b>2</b>	<b>Rank by the lengths of documents</b>	0.15	0.71**
<b>3</b>	<b>Rank by the number of similar words</b>	0.24**	0.71**
<b>4</b>	<b>Rank by the normalized number of similar words</b>	0.24**	0.40*
<b>5</b>	<b>Rank by the usage similarity of note types to query</b>	0.20**	0.72**

Table 6.5: IR performances of different ranking methods in the Crohn project with document set size 200 and a positive ratio of 20%. One-sided Mann-Whitney U test was applied to compare the P@10 and NGR scores of ranking methods. Methods that significantly outperformed the baseline method (Index 1) are marked with \*\* (p-Value < 0.001) and \* (p-Value < 0.05).

<b>Index</b>	<b>Ranking Method</b>	<b>NGR</b>	<b>P@10</b>
<b>1</b>	<b>Rank by the number of keywords</b>	0.02	0.60
<b>2</b>	<b>Rank by the lengths of documents</b>	0.17**	0.42
<b>3</b>	<b>Rank by the number of similar words</b>	0.00	0.76*
<b>4</b>	<b>Rank by the normalized number of similar words</b>	0.17**	0.42
<b>5</b>	<b>Rank by the usage similarity of note types to query</b>	0.17**	0.42

Table 6.6: IR performances of different ranking methods in the AMI project with document set size 200 and a positive ratio of 20%. One-sided Mann-Whitney U test was applied to compare the P@10 and NGR scores of ranking methods. Methods that significantly outperformed the baseline method (Index 1) are marked with \*\* (p-Value < 0.001) and \* (p-Value < 0.05).

<b>Index</b>	<b>Ranking Method</b>	<b>NGR</b>	<b>P@10</b>
<b>1</b>	<b>Rank by the number of keywords</b>	0.02	0.61
<b>2</b>	<b>Rank by the lengths of documents</b>	0.05**	0.60
<b>3</b>	<b>Rank by the number of similar words</b>	0.24**	0.33
<b>4</b>	<b>Rank by the normalized number of similar words</b>	0.05**	0.62
<b>5</b>	<b>Rank by the usage similarity of note types to query</b>	0.06**	0.40

Table 6.7: IR performances of different ranking methods in the Diabetes project with document set size 200 and a positive ratio of 20%. One-sided Mann-Whitney U test was applied to compare the P@10 and NGR scores of ranking methods. Methods that significantly outperformed the baseline method (Index 1) are marked with \*\* (p-Value < 0.001) and \* (p-Value < 0.05).

## 6.5.2 Critical Document Prediction

The result shows that most of the machine learning models provide poor average AUC scores in critical document prediction (Table 6.8). The bag-of-similar-words (BOSW) feature space provides relatively higher AUC scores compared to other feature spaces, which means the number of similar words has a meaningful impact on the importance of a document during chart reviews.

<b>Index</b>	<b>Feature Space</b>	<b>Average AUC</b>
<b>1</b>	<b>bag-of-words (BOW)</b>	0.50
<b>2</b>	<b>bag-of-similar-words (BOSW)</b>	0.62**
<b>3</b>	<b>Average word vectors (word2vec)</b>	0.50
<b>4</b>	<b>Document vectors (doc2vec)</b>	0.50
<b>5</b>	<b>Average usage vectors (usage vecotr space)</b>	0.50

Table 6.8: Average AUC scores (10-fold cross validation) of predicting critical documents with different feature spaces in the dataset constructed with minimum time difference cutoff as 30 seconds.

## 6.6 Discussion

This chapter proposes two novel ranking metrics based on the behavior analysis of how crowd workers interact with medical notes in chart reviews.

The first metric, negative guarantee ratio (NGR), focuses on measuring the ability of a ranking method in filtering out useless documents for crowd workers. The evaluation result shows that a ranking method with high traditional IR performance (e.g., P@K) may not have high NGR score and a ranking method with high NGR score may not have high traditional IR performance.

It is interesting that ranking by the length of documents has high IR performance and NGR score in some cases. We gathered the idea by interviewing crowd workers and medical researchers. We also identify some complex ranking strategies from the activity log of crowd workers using sequential pattern mining. As shown in Table 6.9, we identified the top frequent document access patterns from the activity log of the Crohn project. The analysis shows that we may identify better ranking methods from the access patterns. For example, we may conclude that the crowd workers made a positive decision by reviewing the “Gastroenterology Clinic Visit” document and then review the “Clinical communication” document to confirm the decision. Therefore, as the next step, we may consider crowdsourcing ranking methods and behavior pattern mining as two potential sources for developing efficient ranking methods.

The other metric, critical document, focuses on measuring the performance of a learning-to-rank approach in predicting if a document is critical for making a decision in a chart review. A critical document could be i) a document that at least two crowd workers spent a significantly different time in reviewing and ii) a document that was only selected and reviewed by one worker. The result shows that it is much more difficult to predict if a document is critical (maximum AUC score around 0.62 across all feature spaces) than to predict if a crowd worker would review a document or not (AUC score  $> 0.75$ ). Therefore, there is plenty of room to improve the prediction of critical documents. Possible future

work may consider identifying more efficient features to enhance the power in predicting critical documents.

<b>Index</b>	<b>Frequent Document Access Pattern</b>	<b>Final Decision</b>
<b>1</b>	Gastroenterology Clinic Visit → Clinical Communication	Positive
<b>2</b>	Outpatient Visit Gastroenterology → Clinical communication	Positive
<b>3</b>	Gastroenterology Clinic Visit → Gastroenterology Clinic Visit	Negative
<b>4</b>	Gastroenterology IBD Center Clinic Visit → Clinical communication	Negative

Table 6.9: Top frequent document access patterns for making a decision in the Crohn project.

## 6.7 Conclusion

In this chapter, we presented two novel ranking metrics for evaluating ranking methods for supporting chart reviews. The two ranking metrics, negative guarantee ratio, and critical document are based on the behavior analysis of crowd workers in a serial of chart review projects. We selected ranking methods identified from different sources, including crowds, literature, and brainstorm. The evaluation of the ranking methods shows the current ranking methods and learning-to-rank methods are not efficient enough to support chart reviews. Future research is needed to develop better ranking methods and learning-to-rank methods to support chart reviews.

## Chapter 7

### OVERALL CONCLUSION

#### 7.1 Summary

Clinical chart reviews are one of the most critical components of medical research, which provide high-quality labeled medical data sets, such as patient cohorts with specific medical conditions (e.g., with diabetes), or medical notes that provide specific information (e.g., relevant to diabetes or not). Traditional chart reviews have two limitations. They are slow (e.g., an average of ten hours per patient) and expensive (average payment around \$109 per hour). Therefore, specific strategies and tools are needed to better support clinical chart reviews.

This dissertation systematically discusses the challenges (Chapter 1 and 2) in: i) doing chart reviews fast and cheap; and ii) developing efficient information retrieval tools to support clinical chart reviews. Based on the discussion, we presented our approaches to support clinical chart reviews:

1. Building a light-weight crowdsourcing framework and maintaining a professional worker pool for labeling medical data sets (Chapter 3);
2. Providing high quality clinically similar terms to enhance the query expansion feature of EMR search engine (Chapter 4);
3. Developing method to adaptively adjust to users' semantic preference during chart reviews (Chapter 5).

Moreover, we are the first to do a deep and fine-grained analysis on crowd workers' behavior during chart reviews and propose two novel ranking metrics as the future direction for building high-quality document ranking methods and learning-to-rank methods for clinical chart reviews (Chapter 6).

The evaluation of our approaches to support clinical chart reviews shows that:

1. A Crowdsourced chart review project with appropriate instructions, training sessions, can provide medical researchers high-quality result as well as save significant time (Chapter 3, section 3.8);
2. Extracting clinically similar terms from multiple EMR-based word2vec embeddings can significantly boost the quality of clinically similar terms (Chapter 4, section 4.5);
3. Leveraging the medical contexts of clinical terms can significantly boost the performance of adjusting to users' semantic preference during chart reviews (Chapter 5, section 5.4);
4. Medical researchers and crowd workers have their own document ranking strategies when doing chart reviews. Therefore, specific ranking methods and ranking metrics are needed for building more efficient ranking systems to support clinical chart reviews (Chapter 6, section 6.5).

The work of this dissertation has two takeaways. First of all, EMRs are more complex than general text data, such as messages in social media and News. Therefore, in future research, we should be careful when introducing natural language processing (NLP) methods that have been proven to be efficient in other areas to the analysis and utilization of EMRs. For example, some topic models [219, 220, 42], semantic embeddings [221] are based on the assumption that the training data set contains consistent and single semantic context. Therefore, the relationships of similar terms in contexts such as News and social media, are more stable compared to medical contexts. For example, in News, the similar terms of “cancer” are different types of cancer, such as “breast cancer” and “lung cancer”. However, in a medical context, the similar terms of “cancer” may include the treatments, diagnosis, medications, and symptoms of different types of “cancer”. Moreover, the similar terms of “cancer” may vary with the medical contexts, such as the age, gender of patients or the

previous medical conditions of the patients. The work in Chapter 4 shows that the clinically similar terms that exist in multiple medical contexts (i.e., note types) are preferred by medical researchers for general search, which directly motivated us to pursue the work in Chapter 5.

Second, the usage vector space model provides a new direction in developing NLP methods to support clinical chart reviews. The usage vector space model (Chapter 5) provides an efficient and explainable tool to better identify clinically similar terms for varying medical contexts without training a new semantic embedding each time for a new chart review.

In summary, this dissertation provides knowledge to better collect and produce labels for medical research questions, thus allowing for more supervised machine learning opportunities in healthcare.

## 7.2 The Impact of this Dissertation

### 7.2.1 The impact of this Dissertation to the Healthcare

There are some potential impacts of this work on healthcare. First of all, it improves the efficiency of clinical chart reviews by reducing the time to construct queries and speeding up the document reading process in clinical chart reviews. Second, the usage vector space has the potential to provide a non-keyword search to enhance hospitals' EMR systems. For example, instead of inputting keyword(s), we may allow the input to be the user type, patient type, and medical events to start searching. Third, the usage vector space has the potential to provide explainable machine learning to make document recommendation, query recommendation more user-friendly. For example, when recommending "EEG" when the user input "epilepsy" and "brain", we may explain the recommendation as "we recommend EEG to you because you may prefer the diagnosis of epilepsy".



### 7.2.2 The impact of this Dissertation to Other Domains

There are some potential impacts of this work to the other domains besides healthcare. First of all, the approach of this dissertation could be generalized to other domains, such as law, finance, social media and online retailer (section 1.4.5). For example, the EMR-subsets method may be generalized to other domains to provide high-quality similar terms. Also, the usage vector space model may also be generalized to other domain by capturing the semantic relationships among the professional terms in that domain. Example contexts are shown as follows.

1. Law domain. Possible contexts include acts, cases, policies and so on.
2. Financial domain. Possible contexts include clients' demographic information, income range and so on.
3. Social media domain. Possible contexts include users' background information, such as ages, IP addresses, browser types, forums, threads and so on.

### 7.3 The Scientific Contributions of this Dissertation

There are three main scientific contributions of this dissertation. First of all, previous research, such as word embeddings, identified similar terms using only the text contexts of words. This dissertation expands the resources to identify similar terms from text contexts to domain-knowledge-based contexts, and systematically evaluated the vector space based on the domain-knowledge-based contexts better capture the semantic relationships among the professional terms in a certain domain.

Second, this dissertation presents computational methods that transform context information, such as the note type context (e.g., the EMR-subsets method) and the medical contexts (i.e., the usage vector space) into vector spaces, which better capture the domain knowledge compared to traditional word embeddings.

Third, this dissertation proposes a novel dynamic query recommendation method based on the user's current input without using any search log or previous data. For example, we may build a tool that learns and recommends query without saving users' search log. Therefore, we may provide such a tool as a lightweight, online service to support information retrieval as well as to protect the privacy of users. Moreover, such a dynamic query recommendation method could be easily generalized to other domains, which require domain-knowledge-based and lightweight information retrieval tools.

## BIBLIOGRAPHY

- [1] Amy H Kaji, David Schriger, and Steven Green. Looking through the retrospec-  
toscope: Reducing bias in emergency medicine chart review studies. *Ann. Emerg.  
Med.*, 64(3):292–298, 2014.
- [2] Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and  
Steven Carroll. Automatic Code Assignment to Medical Text. *BioNLP*, pages 129–  
136, 2007.
- [3] Preethi Raghavan, James L. Chen, Eric Fosler-Lussier, and Albert M. Lai. How  
essential are unstructured clinical narratives and information fusion to clinical trial  
recruitment? *AMIA Summits Transl. Sci. Proc.*, 2014:218, 2014.
- [4] Pfrangcon Robekts. MEDICAL TERMS.
- [5] Pedro L Teixeira, Wei-Qi Wei, Robert M Cronin, Huan Mo, Jacob P VanHouten,  
Robert J Carroll, Eric LaRose, Lisa A Bastarache, S Trent Rosenbloom, Todd L Ed-  
wards, Dan M Roden, Thomas A Lasko, Richard A Dart, Anne M Nikolai, Peggy L  
Peissig, Joshua C Denny, SS. Yoon, Q. Gu, T. Nwankwo, JD. Wright, Y. Hong,  
V. Burt, D. Mozaffarian, EJ. Benjamin, AS. Go, JA. Cutler, PD. Sorlie, M. Wolz,  
T. Thom, LE. Fields, EJ. Roccella, WHO ISH Writing Group, MG. Myers, PA.  
James, S. Oparil, BL. Carter, W-Q. Wei, JC. Denny, RE. Klabunde, JC. Denny,  
L. Bastarache, MD. Ritchie, KM. Newton, PL. Peissig, AN. Kho, KM. Newton, PL.  
Peissig, AN. Kho, SJ. Hebring, M. Rastegar-Mojarad, Z. Ye, J. Mayer, C. Jacob-  
son, S. Lin, JC. Denny, MD. Ritchie, MA. Basford, JC. Denny, L. Bastarache, MD.  
Ritchie, DR. Crosslin, DS. Carrell, A. Burt, DC. Crawford, DR. Crosslin, G. Tromp,  
Huan Mo, JA. Pacheco, LV. Rasmussen, E. Bowton, JR. Field, S. Wang, M. Con-  
way, RL. Berg, D. Carrell, Huan Mo, WK. Thompson, LV. Rasmussen, W-Q. Wei,  
PL. Teixeira, Huan Mo, RM. Cronin, JL. Warner, JC. Denny, E. Birman-Deych, AD.

Waterman, Y. Yan, DS. Nilasena, MJ. Radford, BF. Gage, GK. Savova, J. Fan, Z. Ye, JFE. Penz, AB. Wilcox, JF. Hurdle, J. Friedlin, M. Overhage, MA. Al-Haddad, JC. Denny, RA. Miller, LR. Waitman, MA. Arrieta, JF. Peterson, DM. Roden, JM. Pulley, MA. Basford, H. Xu, SP. Stenner, S. Doan, KB. Johnson, LR. Waitman, JC. Denny, H. Xu, M. Jiang, M. Oetjens, W-Q. Wei, RM. Cronin, H. Xu, TA. Lasko, L. Bastarache, JC. Denny, CA. Bejan, W-Q. Wei, JC. Denny, N. Shang, H. Xu, TC. Rindflesch, T. Cohen, R. Khare, J. Li, Z. Lu, JC. Denny, A. Spickard, KB. Johnson, NB. Peterson, JF. Peterson, RA. Miller, J. Denny, J. Smithers, JO. Wrenn, DM. Stein, S. Bakken, PD. Stetson, B. Efron, R. Tibshirani, L. Ohno-machado, T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, J. Friedman, T. Hastie, R. Tibshirani, JE. Bickel, MR. Berthold, N. Cebron, F. Dill, CA. McCarty, P. Peissig, MD. Caldwell, RA. Wilke, and AR. Aronson. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J. Am. Med. Inform. Assoc.*, 24(1):162–171, 2017.

[6] Sujana Perera, Amit Sheth, Krishnaprasad Thirunarayan, Suhas Nair, and Neil Shah. Challenges in Understanding Clinical Notes: Why NLP Engines Fall Short and Where Background Knowledge Can Help. *Int. Work. Data Manag. Anal. Healthc.*, pages 21–26, 2013.

[7] Scott Iwashyna, Theodore J; Odden, Andrew; Rohde, Jeffrey; Bonham, Catherine; Kuhn, Latoya; Malani, Preeti, Chen, Lena; Flanders. Identifying Patients with Severe Sepsis Using Administrative Claims: Patient-Level Validation of the Angus Implementation of the International Consensus Conference Definition of Severe Sepsis. *Med Care*, 18(9):1199–1216, 2014.

[8] Magaly Ramirez, Richard Maranon, Jeffery Fu, Janet S Chon, Kimberly Chen, Carol M Mangione, Gerardo Moreno, and Douglas S Bell. Primary care provider adherence to an alert for intensification of diabetes blood pressure medications before

and after the addition of a chart closure hard stop. *J. Am. Med. Informatics Assoc.*, 25(July):1167–1174, 2018.

- [9] Lee S Beers, Leandra Godoy, Tamara John, Melissa Long, Matthew G Biel, Bruno Anthony, Laura Mlynarski, Rachel Moon, and Mark Weissman. Mental Health Screening Quality Improvement Learning Collaborative in Pediatric Primary Care. *Pediatrics*, 140(6):e20162966, 2017.
- [10] Yunzhu Li, Andre Esteva, Brett Kuprel, Rob Novoa, Justin Ko, and Sebastian Thrun. Skin Cancer Detection and Tracking using Data Synthesis and Deep Learning. *arXiv Prepr. arXiv1612.01074*, 2016.
- [11] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.*, 13:8–17, 2015.
- [12] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
- [13] Luis Tari, Varish Mulwad, and Anna Von Reden. Interactive Online Learning for Clinical Entity Recognition. In *Proc. Work. Human-In-the-Loop Data Anal.*, page 8. ACM, 2016.
- [14] S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J. Am. Med. Inform. Assoc.*, 18(2):181–6, 2011.
- [15] Kenneth H Lai, Maxim Topaz, Foster R Goss, and Li Zhou. Automated misspelling detection and correction in clinical free-text records. *J. Biomed. Inform.*, 55:188–95, jun 2015.

- [16] Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, and Martin Duneld. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *J. Biomed. Semantics*, 5(1):6, 2014.
- [17] Jenny J Chen, Natala J Menezes, and Adam D Bradley. Opportunities for Crowdsourcing Research on Amazon Mechanical Turk. *Interfaces (Providence)*, 5:3, 2011.
- [18] Gabriele Paolacci, Jesse Chandler, and Pg Ipeirotis. Running experiments on amazon mechanical turk. *Judgm. Decis. Mak.*, 5(5):411–419, 2010.
- [19] Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon’s Mechanical Turk. *Behav. Res. Methods*, 44(1):1–23, 2012.
- [20] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspect. Psychol. Sci.*, 6(1):3–5, 2011.
- [21] Cheng Ye, Joseph Coco, Anna Epishova, Chen Hajaj, Henry Bogardus, Laurie Novak, Joshua Denny, Yevgeniy Vorobeychik, Thomas Lasko, Bradley Malin, and Daniel Fabbri. A Crowdsourcing Framework for Medical Data Sets. *AMIA Summits Transl. Sci. Proc.*, 2017:273–280, 2018.
- [22] Benjamin M Good and Andrew I Su. Crowdsourcing for bioinformatics, 2013.
- [23] A. Bowyer, C. Lintott, G. Hines, C. Allen, and E. Paget. Panoptes, a Project Building Tool for Citizen Science. In *Assoc. Adv. Artif. Intell.*, 2015.
- [24] Hongwei Li, Bin Yu, and Dengyong Zhou. Error rate analysis of labeling by crowdsourcing, 2013.
- [25] Jennifer Wortman Vaughan. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.*, 18:193—1, 2016.

- [26] Aditya Vempaty, Lav R. Varshney, and Pramod K Varshney. Reliable crowdsourcing for multi-class labeling using coding theory. *IEEE J. Sel. Top. Signal Process.*, 8(4):667–679, 2014.
- [27] Martin Rajchl, Matthew C H Lee, Franklin Schrans, Alice Davidson, Jonathan Passerat-Palmbach, Giacomo Tarroni, Amir Alansary, Ozan Oktay, Bernhard Kainz, and Daniel Rueckert. Learning under Distributed Weak Supervision. *arXiv Prepr. arXiv1606.01100*, pages 1–5, 2016.
- [28] David R Karger, Sewoong Oh, and Devavrat Shah. Efficient crowdsourcing for multi-class labeling. *ACM SIGMETRICS Perform. Eval. Rev.*, 41(1):81, 2013.
- [29] John B Smelcer, Hal Miller-Jacobs, and Lyle Kantrovich. Usability of Electronic Medical Records. *J. Usability Stud.*, 4(2):70–84, 2009.
- [30] David A. Hanauer, Qiaozhu Mei, James Law, Ritu Khanna, and Kai Zheng. Supporting information retrieval from electronic health records: A report of University of Michigan’s nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *J. Biomed. Inform.*, 55:290–300, 2015.
- [31] Travis Goodwin and Sanda M Harabagiu. UTD at TREC 2014 : Query Expansion for Clinical Decision Support. *23rd Text Retr. Conf. (TREC 2014) Proc.*, 1, 2014.
- [32] Dipasree Pal, Mandar Mitra, and Samar Bhattacharya. Exploring Query Categorisation for Query Expansion: A Study. *arXiv Prepr. arXiv1509.05567*, pages 1–34, 2015.
- [33] Jessie Ooi, Xiuqin Ma, Hongwu Qin, and Siau Chuin Liew. A survey of query expansion, query suggestion and query refinement techniques. *2015 4th Int. Conf. Softw. Eng. Comput. Syst. ICSECS 2015 Virtuous Softw. Solut. Big Data*, pages 112–117, 2015.

- [34] F. Guérin and M. Poisson. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.*, 17(5):507–513, 2010.
- [35] EH Chi, Lichan Hong, Michelle Gumbrecht, and SK Card. ScentHighlights: highlighting conceptually-related sentences during reading. *Proc. 10th ...*, pages 6–8, 2005.
- [36] M. Arguello, M.J. Fernandez-Prieto, and J. Des. Extracting and Visualising Clinical Statements from Electronic Health Records. In *Res. Dev. Intell. Syst. XXX*, pages 307–320. Springer, 2013.
- [37] Cheng Ye and Daniel Fabbri. Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. *J. Biomed. Inform.*, 83(April):–, 2018.
- [38] NIH-NLM. SNOMED Clinical Terms® (SNOMED CT®), 2015.
- [39] David Martinez, Arantxa Otegi, Aitor Soroa, and Eneko Agirre. Improving search over Electronic Health Records using UMLS-based query expansion through random walks. *J. Biomed. Inform.*, 51:100–106, 2014.
- [40] Robert Speer and Joshua Chin. An Ensemble Method to Produce High-Quality Word Embeddings. *Arxiv*, 2016.
- [41] Dongqing Zhu, Stephen Wu, Ben Carterette, and Hongfang Liu. Using large clinical corpora for query expansion in text-based cohort identification. *J. Biomed. Inform.*, 49:275–281, 2014.
- [42] Qing T Zeng, Doug Redd, Thomas Rindfleisch, and Jonathan Nebeker. Synonym, topic model and predicate-based query expansion for retrieving clinical documents., 2012.



- [43] Karthik Natarajan, Daniel Stein, Samat Jain, and Noémie Elhadad. An analysis of clinical queries in an electronic health record search utility. *Int. J. Med. Inform.*, 79(7):515–522, jul 2010.
- [44] Karthik Natarajan. *Analysis of Search on Clinical Narrative within the EHR*. Columbia University, 2012.
- [45] Kai Zheng, Qiaozhu Mei, and David A. Hanauer. Collaborative search in electronic health records. *J. Am. Med. Informatics Assoc.*, 18(3):282–291, 2011.
- [46] David A. Hanauer, Danny T.Y. Wu, Lei Yang, Qiaozhu Mei, Katherine B. Murkowski-Steffy, V. G. Vinod Vydiswaran, and Kai Zheng. Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine. *J. Biomed. Inform.*, 67:1–10, 2017.
- [47] Christopher D. Manning. Introduction to Information Retrieval. *Nat. Lang. Eng.*, 16:100–102, 2010.
- [48] P Avino, I Notardonato, G Cinelli, and M V Russo. Aromatic sulfur compounds enrichment from seawater in crude oil contamination by solid phase extraction. *Curr. Anal. Chem.*, 5(4):339–346, 2009.
- [49] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O’Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. Utilizing social media data for pharmacovigilance: A review. *J. Biomed. Inform.*, 54:202–212, 2015.
- [50] Maria Antoniak and David Mimno. Evaluating the Stability of Embedding-based Word Similarities. *Trans. Assoc. Comput. Linguist.*, 6:107–119, 2018.
- [51] G Hripcsak, D K Vawdrey, M R Fred, and S B Bostwick. Use of electronic clinical documentation: time spent and team interactions. *J Am Med Inf. Assoc.*, 18(2):112–117, 2011.

- [52] Lu Chen, Uta Guo, Lijo C Illipparambil, Matt D Netherton, Bhairavi Sheshadri, Eric Karu, Stephen J Peterson, and Parag H Mehta. Racing Against the Clock: Internal Medicine Residents' Time Spent On Electronic Health Records. *J. Grad. Med. Educ.*, 8(1):39–44, 2016.
- [53] Luke V Rasmussen. The electronic health record for translational research. *J. Cardiovasc. Transl. Res.*, 7(6):607–614, 2014.
- [54] Helma van der Linden, Dipak Kalra, Arie Hasman, and Jan Talmon. Inter-organizational future proof EHR systems. A review of the security and privacy related issues. *Int. J. Med. Inform.*, 78(3):141–160, 2009.
- [55] Annemarie G Hirsch, J B Jones, Virginia R Lerch, Xiaoqin Tang, Andrea Berger, Deserae N Clark, and Walter F Stewart. The electronic health record audit file: the patient is waiting. *J. Am. Med. Informatics Assoc.*, page ocw088, 2016.
- [56] Kai Zheng, Rema Padman, David Krackhardt, Michael P Johnson, and Herbert S Diamond. Social networks and physician adoption of electronic health records: insights from an empirical study. *J. Am. Med. Inform. Assoc.*, 17(3):328–36, 2010.
- [57] Bradley Malin, Steve Nyemba, and John Paulett. Learning relational policies from electronic health record access logs. *J. Biomed. Inform.*, 44(2):333–342, 2011.
- [58] Wen Zhang, Carl A Gunter, David Liebovitz, Jian Tian, and Bradley Malin. Role prediction using Electronic Medical Record system audits. *AMIA Annu Symp Proc*, 2011:858–867, 2011.
- [59] Sarah Read-Brown, Michelle R. Hribar, Leah G. Reznick, Lorinna H. Lombardi, Mansi Parikh, Winston D. Chamberlain, Steven T. Bailey, Jessica B. Wallace, Thomas R. Yackel, and Michael F. Chiang. Time requirements for electronic health record use in an academic ophthalmology center. *JAMA Ophthalmol.*, 135(11):1250–1257, 2017.

- [60] Daniel Fabbri and Kristen Lefevre. Explaining accesses to electronic medical records using diagnosis information. *J. Am. Med. Inform. Assoc.*, 20(1):52–60, 2013.
- [61] Martin R Cowie, Juuso I. Blomster, Lesley H. Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P Pell, Mary Ross Southworth, Wendy Gattis Stough, Martin Thoenes, Faiez Zannad, and Andrew Zalewski. Electronic health records to facilitate clinical research, 2016.
- [62] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.*, 24(1):198–208, 2017.
- [63] Xiao Peng, Elke S Nelson, Jessica L Maiers, and Kris a Demali. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif. Intell. Med.*, 65(2):155–166, 2015.
- [64] Znaonui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. Deep learning for healthcare decision making with EMRs. In *Proc. - 2014 IEEE Int. Conf. Bioinforma. Biomed. IEEE BIBM 2014*, pages 556–559, 2014.
- [65] Kristiina Häyrynen, Kaija Saranto, and Pirkko Nykänen. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int. J. Med. Inform.*, 77(5):291–304, 2008.
- [66] Fitbit Inc. Fitbit Official Site for Activity Trackers & More, 2017.
- [67] United States Congress. Health Information Technology (HITECH Act). *Index Excerpts from Am. Recover. Reinvestment Act 2009*, 2009:112–164, 2009.

- [68] Chun-Ju Hsiao, Esther Hing, Thomas C Socey, and Bill Cai. Electronic Medical Record/Electronic Health Record Systems of Office-based Physicians : United States , 2009 and Preliminary 2010 State Estimates. *Natl. Cent. Heal. Stat.*, pages 2001—2011, 2010.
- [69] David S. Ludwig Gilman W. Matthew. The Meaningful Use Regulation for Electronic Health Records. *Perspective*, 363(1):1–3, 2010.
- [70] Ajit Appari, M. Eric Johnson, and Denise L. Anthony. Meaningful use of electronic health record systems and process quality of care: Evidence from a panel data analysis of U.S. acute-care hospitals. *Health Serv. Res.*, 48(2 PART1):354–375, 2013.
- [71] Matt Vassar and Matthew Holzmann. The retrospective chart review: important methodological considerations. *J. Educ. Eval. Health Prof.*, 10:12, 2013.
- [72] Morgan Harrell, Daniel Fabbri, and Mia Levy. Evaluating EHR Data Availability for Cohort Selection in Retrospective Studies. *2016 IEEE Int. Conf. Healthc. Informatics*, pages 380–387, 2016.
- [73] Shelli L Feder. Data Quality in Electronic Health Records Research. *West. J. Nurs. Res.*, 1:019394591668908, 2017.
- [74] Julie W Doberne, Ze He, Vishnu Mohan, Jeffrey A Gold, Jenna Marquard, and Michael F Chiang. Using High-Fidelity Simulation and Eye Tracking to Characterize EHR Workflow Patterns among Hospital Physicians. *AMIA Annu. Symp. Proc.*, 2015:1881–9, 2015.
- [75] Michael J Mitchell and Michael R King. Secondary Use of Clinical Data: the Vanderbilt Approach. *J. Biomed. Inform.*, 52:28—35, 2014.
- [76] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsu-

- pervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.*, 6(April):26094, 2016.
- [77] Christopher Kotfila and Özlem Uzuner. A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases. *J. Biomed. Inform.*, 58:S92–S102, 2015.
- [78] Evelene M Carter and Henry Ww Potts. Predicting length of stay from an electronic patient record system: A primary total knee replacement example. *BMC Med. Inform. Decis. Mak.*, 14(1):26, 2014.
- [79] A Mortona, Eman Marzban, Georgios Giannoulis, Ayush Patel, Rajender Aparasu, Ioannis A Kakadiaris, April Morton, Eman Marzban, Georgios Giannoulis, Ayush Patel, Rajender Aparasu, and Ioannis A Kakadiaris. A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. *Proc. - 2014 13th Int. Conf. Mach. Learn. Appl. ICMLA 2014*, pages 428–431, 2014.
- [80] Parag C Pendharkar and Hitesh Khurana. Machine learning techniques for predicting hospital length of stay in pennsylvania federal and specialty hospitals. *Int. J. Comput. Sci. Appl.*, 11(3):45–56, 2014.
- [81] Philip Yoon, Ivan Steiner, and Gilles Reinhardt. Analysis of factors influencing length of stay in the emergency department. *CJEM Can. J. Emerg. Med. care = JCMU J. Can. soins médicaux d’urgence*, 5(3):155–61, 2003.
- [82] Aya Awad, Mohamed BaderElDen, and James McNicholas. Patient length of stay and mortality prediction: A survey. *Heal. Serv. Manag. Res.*, 0(0):095148481769621, 2017.
- [83] Manuel Amunategui, Tristan Markwell, and Yelena Rozenfeld. Prediction

- Using Note Text: Synthetic Feature Creation with word2vec. *arXiv Prepr. arXiv1503.05123*, page 13, 2015.
- [84] Jonathan H Chen, Tanya Podchiyska, and Russ B Altman. OrderRex: Clinical order decision support and outcome predictions by data-mining electronic medical records. *J. Am. Med. Informatics Assoc.*, 23(2):339–348, 2016.
- [85] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: {A} {Convolutional} {Net} for {Medical} {Records}. *arXiv1607.07519 [cs, stat]*, 2016.
- [86] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare. *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '15*, pages 1721–1730, 2015.
- [87] Shipeng Yu, Faisal Farooq, Alexander van Esbroeck, Glenn Fung, Vikram Anand, and Balaji Krishnapuram. Predicting readmission risk with institution-specific prediction models. *Artif. Intell. Med.*, 65(2):89–96, 2015.
- [88] Melih Kandemir and Fred A. Hamprecht. Computer-aided diagnosis from weak supervision: A benchmarking study. *Comput. Med. Imaging Graph.*, 42:44–50, jun 2015.
- [89] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *J. Am. Med. Informatics Assoc.*, 21(2):231–237, 2014.
- [90] Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. Large Scale Diagnostic Code Classification for Medical Patient Records. In *Proc. Int. Jt. Conf. Nat. Lang. Process.*, pages 877–882, 2008.

- [91] Yu Cheng, Fei Wang, Ping Zhang, Hua Xu, and Jianying Hu. Risk Prediction with Electronic Health Records : A Deep Learning Approach. In *SIAM Int. Conf. Data Min.*, 2016.
- [92] M. M. Hansen, T. Miron-Shatz, A. Y. S. Lau, and C. Paton. Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives. *IMIA Yearb.*, 9(1):21–26, 2014.
- [93] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.*, 2(1):3, 2014.
- [94] Jingfang Xu, Chuanliang Chen, Gu Xu, Hang Li, and Elbio Renato Torres Abib. Improving quality of training data for learning to rank using click-through data. *Proc. third ACM Int. Conf. Web search data Min. - WSDM '10*, page 171, 2010.
- [95] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. *Adv. Neural Inf. Process. Syst.*, pages 2234—2242, 2016.
- [96] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 25(5):845–869, 2014.
- [97] Cory B. Giles, Chase A. Brown, Michael Ripperger, Zane Dennis, Xiavan Roopnarinesingh, Hunter Porter, Aleksandra Perz, and Jonathan D. Wren. ALE: Automated label extraction from GEO metadata. *BMC Bioinformatics*, 18(Suppl 14), 2017.
- [98] Text Summarization. An Automatic Label Extraction Technique for Domain-Specific Hidden Web Crawling (LEHW). In *Comput. Eng. Syst. 2006 Int. Conf.*, pages 454–459. IEEE, 2006.
- [99] Travis Murdoch and Allan Detsky. The Inevitable Application of Big Data to Health Care. *JAMA Evid.*, 309(13):1351–1352, 2013.

- [100] Robin E Gearing, Irfan A Mian, Jim Barber, and Abel Ickowicz. A methodology for conducting retrospective chart review research in child and adolescent psychiatry. *J. Can. Acad. Child Adolesc. Psychiatry*, 15(3):126–34, 2006.
- [101] Jessica Zeitz Self, Radha Krishnan Vinayagam, J T Fry, and Chris North. Bridging the Gap between User Intention and Model Parameters for Human-in-the-Loop Data Analytics. In *HILDA@ SIGMOD*, pages 1–6, 2016.
- [102] Yue Wang, Kai Zheng, Hua Xu, and Qiaozhu Mei. Interactive medical word sense disambiguation through informed learning. *J Am Med Inf. Assoc*, 0(March):1–9, 2018.
- [103] Marta Poblet, Esteban García-Cuesta, and Pompeu Casanovas. Crowdsourcing roles, methods and tools for data-intensive disaster management, 2017.
- [104] Munindar P. Singh. Crowdsourcing Ground Truth for Medical Relation Extraction. In *IJCAI Int. Jt. Conf. Artif. Intell.*, volume 2015-Janua, pages 4207–4211, 2015.
- [105] Yaliang Li, Nan Du, Chaochun Liu, Yusheng Xie, Wei Fan, Qi Li, Jing Gao, and Huan Sun. Reliable Medical Diagnosis from Crowdsourcing. In *Proc. Tenth ACM Int. Conf. Web Search Data Min. - WSDM '17*, pages 253–261, 2017.
- [106] Amazon.com Inc. Amazon Mechanical Turk - Policies, 2016.
- [107] Jonathan M Mortensen, Mark A Musen, and Natalya F Noy. Crowdsourcing the verification of relationships in biomedical ontologies. *AMIA Annu. Symp. Proc.*, 2013:1020–9, 2013.
- [108] Haijun Zhai, Todd Lingren, Louise Deleger, Qi Li, Megan Kaiser, Laura Stoutenborough, and Imre Solti. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *J. Med. Internet Res.*, 15(4):73–731, 2013.



- [109] Ferda Ofli, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant, Pauline Millet, Friedrich Reinhard, Matthew Parkan, and Stéphane Joost. Combining Human Computing and Machine Learning to Make Sense of Big (Aerial) Data for Disaster Response. *Big Data*, 4(1):47–59, 2016.
- [110] Marc Cuggia, Nicolas Garcelon, Boris Campillo-Gimenez, Thomas Bernicot, Jean François Laurent, Etienne Garin, André Happe, and Régis Duvauferrier. Roogle: An information retrieval engine for clinical data warehouse. *Stud. Health Technol. Inform.*, 169(i):584–588, 2011.
- [111] P. Biron, M. H. Metzger, C. Pezet, C. Sebban, E. Barthuet, and T. Durand. An information retrieval system for computerized patient records in the context of a daily hospital practice: the example of the Léon Bérard Cancer Center (France). *Appl. Clin. Inform.*, 5(1):191–205, 2014.
- [112] Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. HARVEST, a longitudinal patient record summarizer. *J. Am. Med. Inform. Assoc.*, 22(2):263–74, 2015.
- [113] Jennifer G. Stadler, Kipp Donlon, Jordan D. Siewert, Tessa Franken, and Nathaniel E. Lewis. Improving the Efficiency and Ease of Healthcare Analysis Through Use of Data Visualization Dashboards. *Big Data*, 4(2):129–135, jun 2016.
- [114] Adam Perer, Fei Wang, and Jianying Hu. Mining and exploring care pathways from electronic medical records with visual analytics. *J. Biomed. Inform.*, 56:369–378, 2015.
- [115] Eric R. Tkaczyk, Joseph R. Coco, Jianing Wang, Fuyao Chen, Cheng Ye, Madan H. Jagasia, Benoit M. Dawant, and Daniel Fabbri. Crowdsourcing to delineate skin affected by chronic graft-vs-host disease. *Ski. Res. Technol.*, pages 1–6, 2019.

- [116] Savita Gandhani and Nandini Singhal. Content Based Image Retrieval : Survey and Comparison of CBIR System Based on Combined Features. *Int. J. Signal Process. Image Process. Pattern Recognit.*, 8(11):417–422, 2015.
- [117] Christos Faloutsos and Dougals W. Oard. A survey of information retrieval and filtering methods. *Comp. A J. Comp. Educ.*, 8958546:1–24, 1998.
- [118] M. Rami Ghorab, Dong Zhou, Alexander O’Connor, and Vincent Wade. Personalised Information Retrieval: Survey and classification. *User Model. User-Adapted Interact.*, 23(4):381–443, 2013.
- [119] Jan Brophy and David Bawden. Is Google enough? Comparison of an internet search engine with academic library resources. *Aslib Proc. New Inf. Perspect.*, 57(6):498–512, 2005.
- [120] Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Jeff Donahue, and Sarah Tavel. Visual Search at Pinterest. *Knowl. Discov. Databases*, 2015.
- [121] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita. ViSE: Visual search engine using multiple networked cameras. *Proc. - Int. Conf. Pattern Recognit.*, 3:1204–1207, 2006.
- [122] Jing Han, E. Haihong, Guan Le, and Jian Du. Survey on NoSQL database. *Proc. - 2011 6th Int. Conf. Pervasive Comput. Appl. ICPCA 2011*, pages 363–366, 2011.
- [123] R Arun Kumar, M A Jabbar, and Y V Bhaskar Reddy. Information Retrieval systems and Web Search Engines: A Survey. *Natl. Conf. Trends Eng. Technol.*, 25(October):123–125, 2017.
- [124] Xiaoqing Zheng, Yiling Gu, and Yinsheng Li. Data extraction from web pages based on structural-semantic entropy. *Proc. 21st Int. Conf. companion World Wide Web - WWW '12 Companion*, page 93, 2012.

- [125] Sho Takase, Naoaki Okazaki, and Kentaro Inui. Fast and Large-scale Unsupervised Relation Extraction. *29th Pacific Asia Conf. Lang. Inf. Comput.*, pages 96–105, 2015.
- [126] Hua Xu, Shane P. Stenner, Son Doan, Kevin B. Johnson, Lemuel R. Waitman, and Joshua C. Denny. MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Informatics Assoc.*, 17(1):19–24, 2010.
- [127] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S. Zettlemoyer, and Daniel S Weld. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proc. HLT*, pages 541–550, 2011.
- [128] Henry R Ehrenberg, Jaeho Shin, Alexander J Ratner, Jason A Fries, and Christopher Ré. Data Programming with DDLite : Putting Humans in a Different Part of the Loop. In *HILDA*, 2016.
- [129] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data Programming: Creating Large Training Sets, Quickly. *arXiv*, abs/1605:1–26, 2016.
- [130] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proc. Jt. Conf. 47th Annu. Meet. ACL 4th Int. Jt. Conf. Nat. Lang. Process. AFNLP Vol. 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2010.
- [131] Misha Denil, Alban Demiraj, and Nando de Freitas. Extraction of Salient Sentences from Labelled Documents. *arXiv Prepr. arXiv1412.6815*, 2014.
- [132] Michael C. Tackenberg, Jeff R. Jones, Terry L. Page, and Jacob J. Hughey. Tau-independent Phase Analysis: A Novel Method for Accurately Determining Phase Shifts. *J. Biol. Rhythms*, 33(3):223–232, 2018.

- [133] Vivian L West, David Borland, and W Ed Hammond. Innovative information visualization of electronic health record data: A systematic review. *J. Am. Med. Informatics Assoc.*, 22(2):330–339, 2015.
- [134] William Gregg, Jim Jirjis, Nancy M Lorenzi, and Dario Giuse. StarTracker: an integrated, web-based clinical search engine. *AMIA Annu. Symp. Proc.*, 2003(1):855, 2003.
- [135] Ginni Aggarwal and Mukesh Rawat. Ranking of Web Documents for Domain Specific Database. *Int. J. Comput. Appl.*, 135(6):975–8887, 2016.
- [136] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. Using Word Embeddings for Automatic Query Expansion. *arXiv1606.07608 [cs]*, 2016.
- [137] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Inf. Process. Manag.*, 43(4):866–886, 2007.
- [138] Yukikazu NAKAMOTO. A Short Introduction to Learning to Rank. *IEICE Trans. Inf. Syst.*, E94-D(1):1–2, 2011.
- [139] Leonid Boytsov and Anna Belova. Evaluating Learning-to-Rank Methods in the Web Track Adhoc Task. In *Trec*, 2011.
- [140] Jesse M Lingeman and Hong Yu. Learning to Rank Scientific Documents from the Crowd. *arXiv Prepr. arXiv1611.01400*, 2016.
- [141] D Sculley. Large Scale Learning to Rank. *NIPS 2009 Work. Adv. Rank.*, pages 1–6, 2009.
- [142] Thorsten Joachims. Optimizing search engines using clickthrough data. *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '02*, page 133, 2002.

- [143] Bing Li, Rong Xiao, Zhiwei Li, Rui Cai, Bao Liang Lu, and Lei Zhang. Rank-SIFT: Learning to rank repeatable local interest points. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 1737–1744, 2011.
- [144] Tom Seymour, Dean Frantsvog, and Satheesh Kumar. History Of Search Engines. *Int. J. Manag. Inf. Syst.*, 15(4):47, 2011.
- [145] Daniel C. Fain and Jan O. Pedersen. Sponsored search: A brief history. *Bull. Am. Soc. Inf. Sci. Technol.*, 32(2):12–13, 2006.
- [146] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Comput. networks ISDN Syst.*, 30:107–117, 1998.
- [147] Thore Graepel and Rherb Microsoft Com. Web-scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft’s Bing Search Engine. In *Proc. 27th Int. Conf. Mach. Learn. Haifa, Isr.*, 2010.
- [148] Ovidiu Dan and Brian D. Davison. Measuring and Predicting Search Engine Users’ Satisfaction. *ACM Comput. Surv.*, 49(1):1–35, 2016.
- [149] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. *Proc. EMNLP*, 85:404–411, 2004.
- [150] Neelam Duhan, A. K. Sharma, and Komal Kumar Bhatia. Page ranking algorithms: A survey. In *2009 IEEE Int. Adv. Comput. Conf. IACC 2009*, pages 1530–1537, 2009.
- [151] Claudio Carpineto and Giovanni Romano. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.*, 44(1):1–50, 2012.
- [152] Sonia Haiduc, Gabriele Bavota, Andrian Marcus, Rocco Oliveto, Andrea De Lucia, and Tim Menzies. Automatic query reformulations for text retrieval in software engineering. *Proc. - Int. Conf. Softw. Eng.*, pages 842–851, 2013.

- [153] Jinxi Xu and W B Croft. Query expansion using local and global document analysis. *SIGIR '96 Proc. ACM SIGIR Conf.*, 19:4, 1996.
- [154] Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. *16th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 11:160–169, 1993.
- [155] Saar Kuzi, Anna Shtok, and Oren Kurland. Query Expansion Using Word Embeddings. In *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag.*, pages 1929–1932. ACM, 2016.
- [156] Hang Cui, Ji Rong Wen, Jian Yun Nie, and Wei Ying Ma. Query expansion for short queries by mining user logs. *IEEE Trans. Knowl. Data Eng.*, 15(4):829–839, 2003.
- [157] Hiteshwar Kumar Azad and Akshay Deepak. Query Expansion Techniques for Information Retrieval: a Survey. *arXiv Prepr. arXiv1708.00247*, 2017.
- [158] Jessie Ooi, Xiuqin Ma, Hongwu Qin, and Siau Chuin Liew. A survey of query expansion, query suggestion and query refinement techniques. In *2015 4th Int. Conf. Softw. Eng. Comput. Syst. ICSECS 2015 Virtuous Softw. Solut. Big Data*, pages 112–117. IEEE, 2015.
- [159] Patrick Marcel and Elsa Negre. A survey of query recommendation techniques for data warehouse exploration. *Eda*, pages 119–134, 2011.
- [160] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query Recommendation Using Query Logs in Search Engines. In *Lect. Notes Comput. Sci.*, pages 588–596, 2004.
- [161] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. Deep Sentence Embedding Using the Long Short Term Memory Network: Analysis and Application to Information Retrieval. *arXiv1502.06922 [cs]*, pages 1–25, 2015.

- [162] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pages 253–260. ACM, 2002.
- [163] Michael Zalis and Mitchell Harris. Advanced search of the electronic medical record: Augmenting safety and efficiency in radiology, aug 2010.
- [164] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, pages 1532–1543, 2014.
- [165] Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32(23):3635–3644, 2016.
- [166] Hannah Bast, Björn Buchhold, and Elmar Haussmann. Semantic Search on Text and Knowledge Bases. *Found. Trends® Inf. Retr.*, 10(1):119–271, 2016.
- [167] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. Learning to rank with (a lot of) word features. *Inf. Retr. Boston.*, 13(3):291–314, 2010.
- [168] Daria Sorokina and Erick Cantú-paz. Amazon Search : The Joy of Ranking Products. *Proc. 39th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '16*, pages 459–460, 2016.
- [169] Jinying Chen, Abhyuday N Jagannatha, Samah J Jarad, and Hong Yu. Ranking medical jargon in electronic health record notes by adapted distant supervision. *arXiv Prepr. arXiv1611.04491*, 2016.
- [170] Jinying Chen and Hong Yu. Unsupervised ensemble ranking of terms in elec-

- tronic health record notes based on their importance to patients. *J. Biomed. Inform.*, 68:121–131, 2017.
- [171] Mengqi Jin, Hongli Li, Christopher H Schmid, and Byron C Wallace. Using Electronic Medical Records and Physician Data to Improve Information Retrieval for Evidence-Based Care. *IEEE Int. Conf. Healthc. Informatics*, 2016.
- [172] Sanjay Krishnan, Jiannan Wang, Michael J Franklin, Ken Goldberg, Tim Kraska, Tova Milo, and Eugene Wu. SampleClean: Fast and Reliable Analytics on Dirty Data. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.*, pages 59–75, 2015.
- [173] Peivand Bastani, Farzane Doosty, Rohollah Kalhor, Samira Alirezaei, Soudabeh Vatankhah, and Omid Khosravizadeh. Factors affecting length of stay in teaching hospitals of a middle-income country. *Electron. physician*, 8(10):3042–3047, 2016.
- [174] Tobias Krahn, Marco Eichelberg, Stefan Gudenkauf, Gokce B.Laleci Erturkmen, and H. Jürgen Appelrath. Adverse drug event notification system: Reusing clinical patient data for semi-automatic ADE detection. *Proc. - IEEE Symp. Comput. Med. Syst.*, pages 251–256, 2014.
- [175] Yuxiang Zhao and Qinghua Zhu. Evaluation on crowdsourcing research: Current status and future direction. *Inf. Syst. Front.*, 16(3):417–434, 2014.
- [176] Nadine Tamburrini, Marco Cinnirella, Vincent A.A. Jansen, and John Bryden. Twitter users change word usage according to conversation-partner social identity. *Soc. Networks*, 40:84–89, 2015.
- [177] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. *Proc. Second ACM Int. Conf. Web Search Data Min. - WSDM '09*, page 132, 2009.



- [178] Jiannan Wang, Sanjay Krishnan, Michael J. Franklin, Ken Goldberg, Tim Kraska, Tova Milo, and U C Berkeley. A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data. *Proc. 2014 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '14*, pages 469–480, 2014.
- [179] Lior Turgeman, Jerrold H May, and Roberta Sciulli. Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. *Expert Syst. Appl.*, 78:376–385, 2017.
- [180] Shichao Zhang, Yongsong Qin, Xiaofeng Zhu, Jilian Zhang, and Chengqi Zhang. Kernel-Based Multi-Imputation for Missing Data. *Amt*, pages 106–111, 2006.
- [181] Elijah Mayfield and Carolyn Penstein-Rosé. Using feature construction to avoid large feature spaces in text classification. *Proc. 12th Annu. Conf. Genet. Evol. Comput. - GECCO '10*, page 1299, 2010.
- [182] J Makhoul, F Kubala, R Schwartz, and R Weischedel. Performance measures for information extraction. In *Proc. DARPA Broadcast news Work.*, pages 249–252. Herndon, VA, 1999.
- [183] Jionglin Wu, Jason Roy, and Walter F. Stewart. Prediction Modeling Using EHR Data. *Med. Care*, 48(6):S106–S113, 2010.
- [184] Michael M Dinh, Kendall J Bein, Chris M Byrne, Belinda Gabbe, and Rebecca Ivers. Deriving a prediction rule for short stay admission in trauma patients admitted at a major trauma centre in Australia. *Emerg. Med. J.*, 31(4):263–7, 2014.
- [185] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. Voodoo Machine Learning for Clinical Predictions. *bioRxiv*, page 059774, 2016.
- [186] Paweł Matykiewicz and John Pestian. Effect of small sample size on text catego-

- rization with support vector machines. In *Proc. 2012 Work. Biomed. Nat. Lang. Process.*, pages 193–201. Association for Computational Linguistics, 2010.
- [187] Daniel M. Bean, Honghan Wu, Olubanke Dzahini, Matthew Broadbent, Robert Stewart, and Richard J.B. Dobson. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci. Rep.*, 7(1):1–11, 2017.
- [188] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. *Proc. Sixth ACM Int. Conf. Web Search Data Min.*, pages 193–202, 2013.
- [189] Abhimanu Kumar and Matthew Lease. Learning to rank from a noisy crowd. *Proc. 34th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 800:1221–1222, 2011.
- [190] Shihao Ji, Hyokun Yun, Pinar Yanardag, Shin Matsushima, and S. V. N. Vishwanathan. WordRank: Learning Word Embeddings via Robust Ranking. *arXiv*, pages 1–12, 2015.
- [191] Jinying Chen, Jiaping Zheng, and Hong Yu. Finding Important Terms for Patients in Their Electronic Health Records: A Learning-to-Rank Approach Using Expert Annotations. *JMIR Med. Informatics*, 4(4):e40, 2016.
- [192] Xiang Wan, Jiming Liu, William K Cheung, and Tiejun Tong. Learning to improve medical decision making from imbalanced data without a priori cost. *BMC Med. Inform. Decis. Mak.*, 14(BioMed Central):111, 2014.
- [193] Joan M. Kiel and Laura M. Knoblauch. HIPAA and FERPA: Competing or collaborating? *J. Allied Health*, 39(4):161–165, 2010.
- [194] Public Law. Health insurance portability and accountability act of 1996. *Public Law*, 104:191, 2003.

- [195] John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *Int. J. Med. Inform.*, 79(12):849–859, 2010.
- [196] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proc. Int. Conf. Learn. Represent. (ICLR 2013)*, pages 1–12, 2013.
- [197] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Adv. Neural Inf. Process. Syst.*, pages 3111–3119, 2013.
- [198] D M Roden, J M Pulley, M A Basford, G R Bernard, E W Clayton, J R Balsler, and D R Masys. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin. Pharmacol. Ther.*, 84(3):363, 2008.
- [199] Radim Rehurek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proc. Lr. 2010 Work. New Challenges NLP Fram.*, pages 45–50. ELRA, 2010.
- [200] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. Query Expansion with Locally-Trained Word Embeddings. *arXiv Prepr. arXiv1605.07891*, pages 367–377, 2016.
- [201] Leonard Richardson. Beautiful Soup Documentation. Technical report, PyPI, 2016.
- [202] Stat Consulting Group. MULTINOMIAL LOGISTIC REGRESSION — R DATA ANALYSIS EXAMPLES, 2014.
- [203] Jon Starkweather and Amanda Kay Moske. Multinomial logistic regression. *Multinomial Logist. Regres.*, 51(6):404–410, 2011.

- [204] Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Informatics Assoc.*, 21(2):221–230, 2014.
- [205] Kavita Ganesan, Shane Lloyd, and Vikren Sarkar. Discovering Related Clinical Concepts Using Large Amounts of Clinical Notes. *Biomed. Eng. Comput. Biol.*, 7(Suppl 2):27–33, 2016.
- [206] Joshua C. Denny, Anderson Spickard, Kevin B. Johnson, Neeraja B. Peterson, Josh F. Peterson, and Randolph A. Miller. Evaluation of a Method to Identify and Categorize Section Headers in Clinical Documents. *J. Am. Med. Informatics Assoc.*, 16(6):806–815, 2009.
- [207] Ashutosh Malhotra, Erfan Younesi, Michaela Gündel, Bernd Müller, Michael T. Heneka, and Martin Hofmann-Apitius. ADO: A disease ontology representing the domain knowledge specific to Alzheimer’s disease. *Alzheimer’s Dement.*, 10(2):238–246, 2014.
- [208] Claudia Perlich, Foster Provost, and Jeffrey S Simonoff. Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *J. Mach. Learn. Res.*, 4:211–255, 2003.
- [209] Sebastien Romano, Nathanael and Dubois. Learning Effective Embeddings from Medical Notes. *arXiv Prepr. arXiv1705.07025*, pages 1–10, 2017.
- [210] Yue Wang, Kai Zheng, Hua Xu, and Qiaozhu Mei. Interactive medical word sense disambiguation through informed learning. *J. Am. Med. Informatics Assoc.*, 25(7):800–808, 2018.
- [211] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pages 1135–1144. ACM, 2016.

- [212] Michael W Temple, Christoph U Lehmann, and Daniel Fabbri. Predicting Discharge Dates From the NICU Using Progress Note Data. *Pediatrics*, 136(2):e395–405, 2015.
- [213] D Buscaldi and P Rosso. A bag-of-words based ranking method for the Wikipedia question answering task. *Eval. Multiling. Multi-modal Inf. Retr.*, 4730:550–553, 2007.
- [214] Mercy ME Paul Selvan, A ME Chandra Sekar, and APriya Dharshin. Survey on Web Page Ranking Algorithms. *Int. J. Comput. Appl.*, 4119:975–8887, 2012.
- [215] Xin Zhang, Ben He, Tiejian Luo, Dongxing Li, and Jungang Xu. Clustering-based transduction for learning a ranking model with limited human labels. *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manag. - CIKM '13*, pages 1777–1782, 2013.
- [216] Jinying Chen and Hong Yu. Unsupervised Ensemble Ranking of Terms in Electronic Health Record Notes Based on Their Importance to Patients. *J. Biomed. Inform.*, 2017.
- [217] Aida Nematzadeh, Stephan C Meylan, and Thomas L Griffiths. Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words. *Proc. 39th Annu. Conf. Cogn. Sci. Soc.*, pages 859–864, 2017.
- [218] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus Robert Müller, and Wojciech Samek. "What is relevant in a text document?": An interpretable machine learning approach. *PLoS One*, 12(8):1–19, 2017.
- [219] David Blei, Lawrence Carin, and David Dunson. Probabilistic topic models. *IEEE Signal Process. Mag.*, 27(6):55–65, apr 2010.

- [220] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proc. Lr. 2010 Work. New Challenges NLP Fram.*, pages 46–50. Citeseer, 2010.
- [221] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.)*, pages 238–247, 2014.