NLP to identify and rank clinically relevant information from EHRs in the Emergency Department

Investigators:	Foster Goss DO, MMSc, Tom Korach, MD, Kevin Bretonnel Cohen, PhD, Kelly Bookman MD, Lawrence E. Hunter PhD, Alyssa Witeof BS, Li Zhou MD, PhD
Organization	University of Colorado Denver
Dates	9/1/2016 – 8/31/2018
Project Officer	Christine Dymek
Acknowledgment of Agency Support	Agency for Healthcare Research and Quality
Grant Number	R21 HS24541

ABSTRACT:

***Purpose**: Timely identification and extraction of relevant or "need to know" clinical information about a patient's history in the emergency department (ED) setting is critical for patient safety and medical decision-making.*

***Scope**: Develop a corpus of relevant information elements for the complaints of chest pain and back pain and then develop a search tool to automatically identify these items within the EHR.*

***Methods**: The relevant information elements for each of these complaints were developed using subject matter experts(SME) in Emergency Medicine, Cardiology and Orthopedics. We used the Medical Term Extraction and Reasoning System (MTERMS) and SOLR/Lucene to extract and pre-process these information elements from clinical notes. We developed and evaluated both manual (SME) and unsupervised machine learning ranking methods and compared their accuracy head-to-head on 1,010 medications and 2,913 problems from 99 patients with a chief complaint of chest or back pain, where each item was manually labeled as relevant or not to the chief complaint. A graphical user interface was developed.*

***Results**: For chest pain, the following relevant items were identified: 12 risk factors, 40 diagnoses, 20 diagnostic tests, 14 procedures, and 10 therapeutic drug classes. For back pain, the following relevant items were identified:13 risk factors, 25 diagnoses, 16 diagnostic tests, 4 procedures, and 5 therapeutic drug classes. Using mean average precision (MAP), the manual ranking out-performed the unsupervised methods on medications (92.2%-98.2% vs 84.8%-90.8%) but underperformed on problems (57.6%-84.9% vs 71.7%-88.7%). The results demonstrate how information retrieval methods using NLP and unsupervised machine learning can provide a reasonably accurate, low-effort, and scalable method for situation-specific clinical relevancy ranking.*

Keywords: Information Storage and Retrieval, Electronic Health Records, Emergency Service, Hospital, Natural Language Processing, Patient Safety.

PURPOSE:

The objectives of this study were 1) to develop a complaint-specific history of relevant information that, if found in the Electronic Health Record (EHR) of an ED patient, should be presented to the treating physician, and 2) to develop an NLP and machine learning search tool to automatically identify and clinically rank this information (both free and structured text) based on relevance to the ED provider. The proposed study will create a novel way to automatically compile clinically and contextually relevant information using Natural Language Processing (NLP) and machine learning, thereby increasing the value of information available for providers' medical decision-making and improving both quality and safety of patient care.

SCOPE:

Background

Timely identification of relevant or "need to know" clinical information about a patient's history in the emergency department (ED) setting is critical for patient safety and medical decision-making. Relevant information, however, is often buried in unstructured or free-text narratives within the Electronic Health Record (EHR), making it time consuming to access. Current search tools within an EHR are often based on key-word search, which is inefficient (e.g., does not consider context) and simplistic (e.g., only captures exactly matched terms). Furthermore, these search tools are often unable to rank or evoke the relevance of information for a particular problem or complaint. One solution is to automatically identify clinically relevant information using natural language processing (NLP) and machine learning. NLP has been widely used to identify both free-text and structured information across a variety of clinical domains(1-5), including adverse drug events(6-9), risk factors for surgical site infections(10), and biosurveillance(9, 11). Unlike traditional EHR search methods, newer search tools can use NLP and machine learning to automatically process, filter, and rank free-text information that providers need to know for a patient's complaint (e.g., chest pain) and present this alongside structured or coded information so that providers have a "snapshot" of relevant information.

Information retrieval (IR) using machine learning from the medical records has been the focus of research and public competitions, such as CLEF and TREC, and several tools have been published. However, they mostly focused on searching for patients in a population, rather than searching for specific details within a

single patient's record. EMERSE was implemented at University of Michigan in 2005, but is used mostly for research and other secondary-uses, and does not employ any clinical relevancy ranking(12). Similarly, STRIDE, from Stanford University(13) and StarTracker(14), which focus on structured data querying, are also designed to identify patient cohorts and do not describe any clinical relevancy ranking. CISearch(15), from Columbia University, allows ad-hoc queries over narrative texts but does not perform any clinical relevancy ranking and does not include structured information elements.

In general IR tasks, term frequency-inverse document frequency (TF-IDF), is a popular ranking mechanism underlying (along with similar mechanisms such as Okapi BM25) many IR solutions(16). For a given term, it incorporates both its probability to occur in a given document (TF) and the amount of information that it represents (IDF). Although TF-IDF is a cornerstone of relevancy ranking in textual search engines, a fundamental characteristic of clinical information renders it less useful for EHR IR: TF-IDF and its derivatives rely on the presence of a term in the searched documents to capture the document's relevancy. For clinical information, however, this assumption does not hold. The ranked items themselves (e.g. medications) rarely contain any of the words defining the clinical situation itself. For example, while aspirin is a highly relevant item for chest pain, aspirin prescriptions rarely contain the phrase "chest pain" or any of its synonyms. Essentially, from the perspective of the medical record, the relevancy is latent and the association of items with a clinical situation is not manifested in the occurrence pattern of the search terms. Moreover, such association may be of a higher order: anticoagulants ("blood-thinners"), a medication class relevant to chest pain, despite not being used to treat chest pain itself. Rather, they are used to prevent blood clot formation in atrial fibrillation, which itself does not typically manifest as chest pain, but instead stems from conditions that also cause chest pain (e.g. ischemic heart disease). Therefore, clinical information ranking requires a way to capture latent topical associations.

Various unsupervised methods have been developed to automatically find the important items in a document. TextRank, is an adaptation of Page and Brin's PageRank algorithm to textual units such as words and sentences. Briefly, it is a graph-based ranking algorithm that determines an item's score by the scores of the items pointing to it. For NLP, instead of webpages and their hyperlinks it uses words and their relationships to find important items in documents. It has been used successfully for keyword extraction and document summarization(17). A recent work that evaluated unsupervised methods for ranking the importance of terms in the medical record for patient-oriented IR found that ensemble ranking, including TextRank, achieved high accuracy (0.885 AUC-ROC) compared to a manual ranking of importance(18). In recent years, neural word embeddings have been widely used to solve many problems in natural-language processing. These are distributed representation of symbols, using continuous vector space to represent the meaning of discrete symbols such as words and sentences by processing a large corpus of unlabeled data. They can be tuned to capture topical relationships between words(19), and have been successfully used in IR(20), all while requiring no manual input beside hyper-parameter fine tuning. Considering the challenges to rank clinical information items, we hypothesize that *situation-specific relevancy can be estimated using a general unsupervised method*, and that the factors underlying the situation-specific relevancy are common to many clinical situations and can be estimated using unsupervised methods.

Context:

In the ED, access to the right information about the right patient at the right time is critical to a physician's medical decision-making(21, 22). Patients can have difficulty recalling the important or relevant medical history during periods of stress or exacerbation of an illness(23), requiring the provider to quickly search the EHR. Often the relevant information is buried in unstructured or free-text narratives within the EHR, making it time consuming to access(24). Additionally, it can be difficult to find and filter the information by relevance to the patient's presenting complaint. In the fast-paced setting of an ED or intensive care unit, this may lead to delays in diagnosis (25) or recognition of potential life threats.

Setting:

In this study, we identified a cohort of patients with a chief complaint of "chest pain" or "back pain" who visited the ED of the University of Colorado Hospital (UCH) or Brigham and Women's Hospital (BWH) during January 1st to December 31th, 2016. For each patient, we retrieved their data from the past two years

including lab tests, procedures, diagnoses, and medications. Both structured and free-text data were obtained.

Participants:

There were 9,347 patients' data that were extracted from the EHR between January 1st to December 31th, 2016 that were used to develop and test our automated search tool. 100 patients were set aside for our final testing. We had 7 subject matter experts (SMEs) assist in the study, helping develop and refine the information elements to be extracted from the EHR for the respective chief complaints and then evaluate the user interface that was developed.

METHODS:

Identifying and ranking information items chief complaints (Aim 1)

The relevant information elements for each of these complaints were developed through 1) literature review, 2) domain experts in emergency medicine, and 3) domain experts in specialties of cardiology and orthopedics (i.e. specialists for chest pain and back pain, respectively). We used a modified Delphi method consisting of two rounds until all the relevant items were identified and ranked. For each information element, the SMEs ranked the relevancy values (individual diagnoses, medication classes, etc.) on a Likert scale of 0-4, and the responses were averaged to yield a final continuous-number score in that range. The SME ranking was provided in a non-standard terminology and was translated to the terminology used in the patient charts (First Data Bank for medications and ICD-10-CM for problem lists) manually.

Developing and testing an NLP-based information search tool (Aim 2)

In this aim, we used a combination of artificial intelligence methods to develop our search tool. Our initial step was building a Solr/Lucene index and search function that would allow us to query a document for the presence of relevant information items. We then leveraged the MTERMS system to help identify unstructured data. Lastly, realizing the limitation of having manually developed information elements for each complaint, we explored unsupervised machine learning methods that could be used to search for relevant information elements without any training or supervision necessary. A prototype of the search tool's graphical user interface was created using the Invision app and feedback was obtained from end users.

Solr/Lucene Index Function

For document indexing, we used the Apache Solr, an open source enterprise search platform providing indexing and retrieval mechanism based on the Apache Lucene library. We developed a schema for the search index, to transform clinical data within the EHR into a suitable form and balance between search expressivity and index size (Table 1). The schema dictates how the behavior and semantics of the search process and requires a balance flexibility (handling misspellings and partial matches) vs. accuracy (flexible matching might retrieve unplanned terms) as well as practical considerations of speed and hardware resources.

Table 1: Solr/Lucene schema representing the data elements, type, and purpose.

Element	Field name	Data type	Purpose
Common	Patient	Identifier	Identify the patient to which this record is assigned.
Common	Encounter	Identifier	Identify the encounter to which this record is assigned.
Common	Effective date	Date time	The most relevant date value (attributed date>generated date>entered date).
Common	Code identifier	Descendant path	Identify the code (and coding system) of this item.
Common	Description	Text	Canonical description of the item in this record (e.g. the "preferred term" for a SNOMED CT code, etc.).
Common	Additional information	Text	A default field for concatenation of additional information for this record, e.g. signature for medications, side for procedures and diagnoses, etc.
Common	Additional descriptions	Text	Additional descriptions (e.g. synonyms) of the item described in this record.
Common	Data element	Text	The data element represented by this document (medication, diagnosis, etc.).
Common	Logical path - instance	Descendant path	The classification of the item in this document, e.g. "Medication/diabetes medications/oral hypoglycemics/metformin". Allows searching for items by their clinical definition.

Element	Field name	Data type	Purpose
Common	Creator	Text	The clinician that created this item. Use the full name to allow fast searches without requiring identification of the clinician in the organization's clinician list.
Common	Institute	Text	The institute in which the record was created.
Common	Additional information in EHR	Boolean	A flag that additional important information (e.g. edits, amendments etc.) is available in the EHR.
Medication	Prescription start date	Date time	
Medication	Prescription end date	Date time	
Medication	Chronic/acute prescription	Enumeration	
Medication	signature	Text	
Medication	Last dispensation date	Date time	The last time the prescription was dispensed from the pharmacy.
Laboratory results	Result value	Text	
Laboratory results	Reference range lower bound	Number	
Laboratory results	Reference range upper bound	Number	
Laboratory results	Normal	Boolean	A flag whether this result is normal or not. May not be available for all results.
Diagnosis	Comment	Text	
Problems	Comment	Text	
Procedure	Comment	Text	

MTERMS

To translate the SMEs' ranking to actual ranking logic, we used two approaches: for unstructured data, we created rule-based information extraction logic to capture mentions of these entities from narrative texts. Clinicians sometimes explicitly mention entities that are absent from the clinical situation (e.g. "the patient denies chest pain"). While such negated mentions are important for clinical and medico-legal purposes, they may interfere with patient record search since typically the searching clinician is interested in the existing finding rather than the ruled-out ones. Thus, we implemented the NegEx algorithm to handle and omit negated mentions from the medical text. MTERMS itself is an information extraction, rather than retrieval, tool, while the Solr search engine requires the indexed information to be represented as atomic terms (e.g. words). Therefore, we combined the two tools in tandem with MTERMS, which served as a pre-processor to organize and clean the raw and noisy EHR text for indexing and retrieval. For structured data elements, we translated the desired entities to the terminologies used in the data set.

Machine Learning Unsupervised Ranking

We evaluated two ranking methods, both using an unlabeled corpus of patient records and MAP was calculated.

Pointwise-mutual information

Medications: To better detect appearances of medications with the chief complaints, each ingredient-route combination was expanded to brand names using RxNorm(26). Pointwise-mutual information (PMI) was then calculated based on co-occurrence of any of the original ingredient-route combination or the expanded names in the same note as the chief complaint, using the formula:

$$\text{PMI(med, complaint)} = \log_2 \frac{\frac{f(complaint,med)}{N}}{\frac{f(complaint)}{N} \times \frac{f(med)}{N}}$$

Where $f(complaint)$ is the number of notes mentioning the chief complaint ("chest pain" or "back pain"), $f(med)$ is the number of notes containing the medication name, $f(complaint, med)$ is the number of notes contained in both the chief complaint "complaint" and the medication name *med*, and N is the number of notes. Since each medication was expanded to multiple names yielding multiple PMI values, the maximal value was chosen to represent the medication.

For problems, since the diagnoses names rarely appear verbatim in patient notes, we mapped to their ICD codes and measured the PMI based on co-occurrence of each diagnosis code in the same encounter with notes mentioning “chest pain” or “back pain”.

$$\text{PMI(dx, complaint)} = \log_2 \frac{\frac{f(complaint,dx)}{N}}{\frac{f(complaint)}{N} \times \frac{f(dx)}{N}}$$

Where $f(complaint)$ is the number of notes containing the chief complaints (“chest pain” or “back pain”), $f(dx)$ is the number of diagnosis records, $f(complaint, dx)$ is the number of notes contained in the chief complaint “complaint” from an encounter with a diagnosis *dx*, and N is the number of notes.

TextRank

To capture the relevancy of an item for a chief complaint, we calculated its TextRank score and compared that to the other items. TextRank operates on a graph of items, and each item’s score is calculated based on the scores of all other items connected to it. We followed the original TextRank formula

$$S(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{j,i}}{\sum_{V_k \in Out(V_j)} w_{j,k}} \times S(V_j)$$

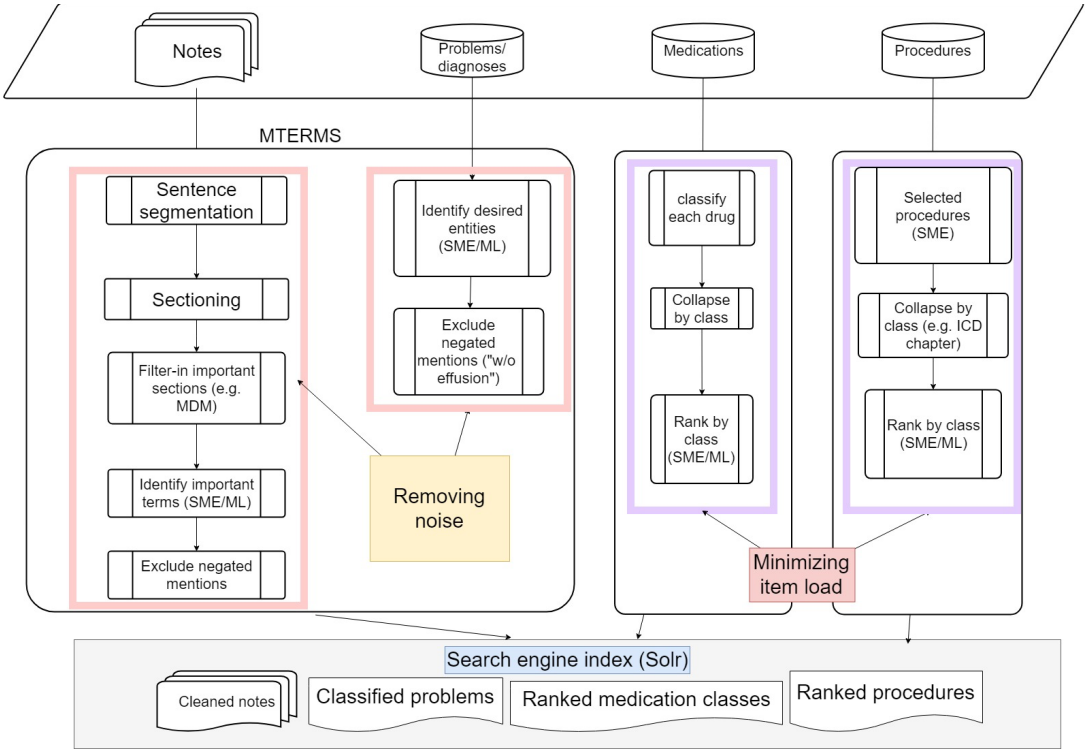
Where $S(V_i)$ is the score of item V_i, $In(V_i)$ and $Out(V_i)$ are the sets of nodes connected to V_i via incoming and outgoing edges, respectively, $w_{j,i}$ is the weight of the connection, and d is a damping factor. The formula represents and assigns a weight to the item using the weights of the items that “point” (i.e. are among the incoming connections) to it and scaling this contribution to the strength of the connection between each pointing item and the scored item. The denominator in the summation element transforms the connection weight $w_{j,i}$ to the range [0-1] (assuming a positive value for all $w_{j,k}$), essentially turning it into a probability. The damping factor d is used to incorporate the probability of diverging from the graph connections and jumping from the pointing item to a random node in the graph.

The weights $w_{j,i}$ between the items are at the core of the TextRank algorithm and while the general principles of TextRank remain the same, the function determining the weights guide the results and differs from use case to use case. In this instance, the graph included all of the items to be ranked. For similarity between items, we used the positive-PMI (PPMI),

$$PPMI = \max(0, PMI)$$

calculated on notes mentioning the chief complaint, to customize the weight to a specific complaint. During the graph iteration stage, we used a damping factor of 0.85 and stopping criteria of a total change in items score of 1E-6 or 200 iterations. Since the problem names used in the patient data rarely appeared verbatim in the patients’ notes, TextRank was not used to rank problems. No training was done for the unsupervised methods and no pre-processing was performed on the utilized corpus of notes.

Figure 1: Overview of search methods showing integration of SOLR/Lucene, MTERMs and Machine Learning.



Limitations

Our study had several limitations in its approach. First, we used the medication history rather than the current medications list, since it is a volatile list (correct only for a point in time), and therefore was not available from the enterprise data warehouse. While the active medications list may differ in item distribution, the challenge of ranking medications based on topical relevance is similar, and the medication history is a valuable information source on its own. Second, while we sampled over a thousand medications and almost three thousand problems, the number of patients was smaller, especially for the group with both chief complaints. This limitation is exacerbated by the high percentage of relevant items, which may mask errors of the ranking problems.

RESULTS:

Delphi Ranking

Initial rankings were obtained by SMEs in Emergency Medicine, Cardiology, and Orthopedics. We underwent two rounds for a duration of 3 months and obtained the information elements relevant to the complaints of chest pain and back pain. An average ranking was generated for each informational element. Examples of these are shown below in Figure 2.

Figure 2: Examples of Delphi method for complaint of chest pain including diagnosis and medications.

Diagnosis	Relevant	Average Ranking
acute coronary ischemia	yes	4.0
acute coronary syndrome	yes	4.0
acute myocardial infarction	yes	4.0
aortic dissection	Yes	4.0
CAD	yes	4.0
congestive heart failure	yes	4.0
coronary atherosclerosis	yes	4.0
heart failure	yes	4.0
NSTEMI	yes	4.0
STEMI	yes	4.0
angina	yes	3.8

Class	Relevant	Average Ranking
Antithrombotics	Yes	3.5
antiarrythmics	Yes	3.3
Anti-hyperlipedemia	Yes	3.3
Vasodilators	Yes	3.3
Beta Blockers	Yes	3.0
Fibrinolytics	Yes	3.0
Inotropes/Vasopressors	Yes	3.0
ACE Inhibitors/ARBs	Yes	2.8
CCB	Yes	2.8
Diuretics	Yes	2.3

For our unsupervised machine learning, we performed a similar approach, where we had the clinicians validate information elements that were identified by Solr/Lucene from clinical notes. For this validation, we had 3 physicians participate, resulting in 96 and 60 ranked elements for chest pain and back pain,

respectively, with little overlap of the items between the two complaints. The results of our machine learning methods are shown below (Table 2).

Table 2: Distribution of SME-ranked elements and their scores.

Information element	Elements number		Average score (0-4 Likert scale)		Jaccard index
	Chest pain	Back pain	Chest pain	Back pain	
Diagnosis	40	22	3.3	3.1	0.016
Diagnostic tests	20	16	3.5	3.4	0.000
Medication class	10	5	3.0	3.0	0.000
Procedures	14	4	3.3	4.0	0.000
Risk Factors	12	13	2.9	2.7	0.190

Machine Learning:

We performed SME validation of our methods on 99 patients presenting with chest pain or back pain. The distribution of these patients and their associated problems and medications are shown in Table 3. Six patients did not have medications left after filtering and were removed from the gold-standard data set.

Table 3: Patient distribution by chief complaint.

Complaint	Patients	Medications	Problems	Age
Total	99	1,010	2,193	61.2 (13.9)
Back pain	19	195	636	59 (14.7)
Both	5	77	157	51.3 (5.9)
Chest pain	75	738	2,120	62.4 (13.7)

Gold-standard labeling

The lower bounds for the inter-rater agreements achieved by the three clinicians are Light's kappa of 0.74 for medications and 0.58 for problems. These are both probably underestimates of the two agreement values, being based on the highest possible estimate of P(e) (the "expected by chance" agreement). The average information load per patient, in terms of the number of items and the relevant proportion, are shown in Table 4.

Table 4: Information load (number of items and percentage of relevant ones per patient) in the gold-standard data set, by complaint.

	Medications								Problems							
	Back pain		Both		Chest pain		All patients		Back pain		Both		Chest pain		All patients	
	Tot	*Rel %*	*Tot*	*Rel %*	*Tot*	*Rel %*	*Tot*	*Rel %*	*Tot*	*Rel %*	*Tot*	*Rel %*	*Tot*	*Rel %*	*Tot*	*Rel %*
Avg	10.8	49.9%	15	86.7%	11	57.5%	11	57.6%	33.5	52.1%	31.4	83.2%	28.3	72.3%	29.4	69.0%
SD	6.1	5.2%	5.6	26.7%	7.5	25.8%	7.2	27.4%	21.7	5.0%	13.9	11.1%	26.6	21.2%	25.3	22.6%
Med	9	54.2%	16	100.0%	9	54.2%	9	54.5%	32	50.0%	39	82.5%	19	77.8%	23	71.1%
Min	2	0.0%	6	33.3%	1	0.0%	1	0.0%	1	11.1%	4	67.5%	1	27.8%	1	11.1%
Max	25	100.0%	23	100.0%	30	100.0%	30	100.0%	77	100.0%	40	100.0%	125	100.0%	125	100.0%

Avg: average, SD: standard deviation, Med: median, Min: minimum, Max: maxium, Tot: total items, Rel %: percentage

Ranking results

We found that our unsupervised machine learning methods could identify 84.8%-90.8% of relevant medications for the complaints of back pain and chest pain, respectively. This was lower than the SME which ranged from 92.2% – 98.2%. For problems, our machine learning methods actually performed better

than the SME at identifying relevant medical problems with a mean average precision of 71.7%-88.73%. These findings are summarized in Table 5.

Table 5: Mean Average Precision of the ranking methods for medications.

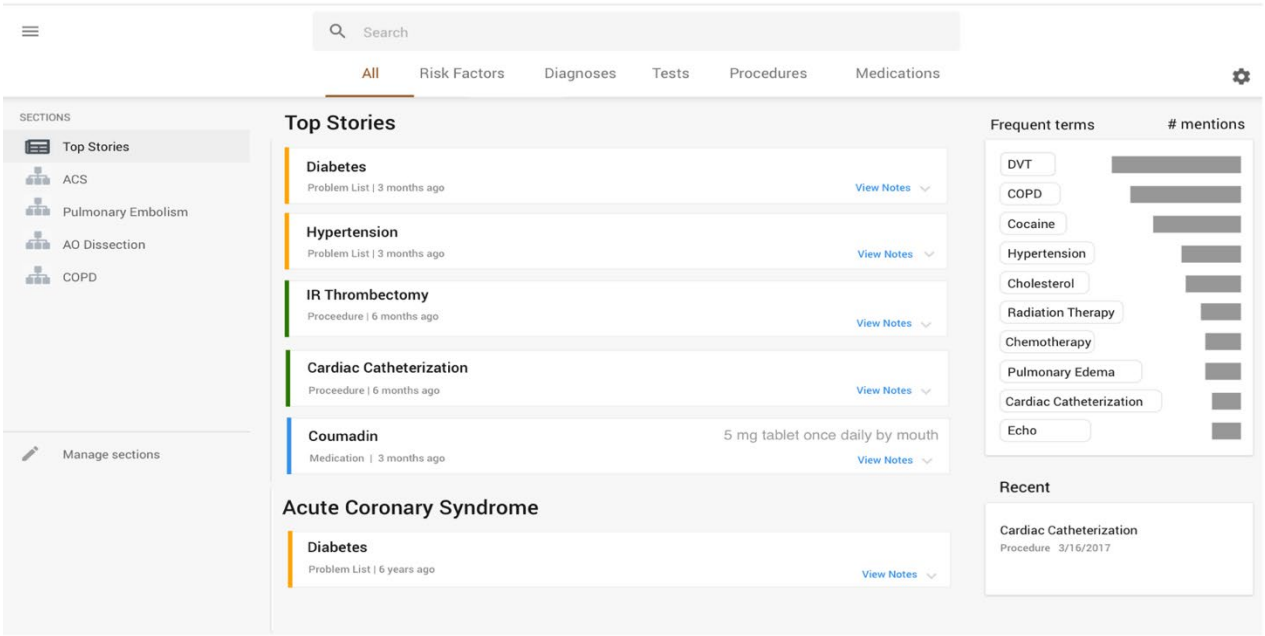
	Medications			Problems	
Complaint	**SMEs**	**TextRank**	**PMI**	**SMEs**	**PMI**
Back pain	92.20%	72.00%*	84.80%*	57.67%	71.70%*
Both	94.00%	90.00%	100.00%	87.05%	89.14%
Chest pain	98.20%	70.40%*	90.80%	84.94%	88.73%*

* 95% significant compared to SME ranking

Graphical User Interface:

A representation of our user interface is shown below (Figure 3). There were 18 screens created for the search tool. We employed a user-centered design approach similar to how information is presented on Google News. An example is shown below for a patient presenting with chest pain. On the left of the screen, are the sections which contain the top stories contained within the patient's chart that are relevant to the patients presenting complaint of chest pain. Below are other related diagnosis that would be in the differential diagnosis for the treating provider. These may include but would not be limited to acute coronary syndrome (ACS), pulmonary embolism, or aortic dissection. The user could click on one of these to filter the results for a particular diagnosis. On the top of the screen are the different categories of information elements including risk factors, diagnosis, tests, procedures, and medications. Each is color coded to help the user recognize the category. When the user clicks on the category, they will see cards that contain relevant information elements that are ranked from most important to least important. Cards will contain the name of the information element and also its corresponding value (e.g., Troponin level, dose of medication, etc.)

Figure 3: Graphical user interface of search tool for patient with the complaint of chest pain.



DISCUSSION:

Principal findings and outcomes

The current study focused on tackling the information load of structured information items in the medical record of ED patients using a combination of IR methods and NLP. The gold-standard analysis sheds light on the extent of this information load, showing physicians encounter problem lists as long as dozens of items that need to be processed under the time constraints of clinical practice. The demonstrated information load emphasizes the need for information ranking. On the other hand, the virtually absent overlap between the

items for the two chief complaints (average Jaccard index of 0.041) demonstrate the need for situation-specific ranking, reinforcing the scalability challenges of using SME-based ranking and the resulting need for automated methods.

While SME still outperformed the unsupervised methods for medications, the latter approached SME performance and showed substantial potential for improvement. The higher MAP of the PMI method on medications among patients with both complaints may stem from the fact a higher proportion of the medications of such patients is considered relevant (100.0% vs 51.5% for chest pain and 48.2% for back pain). This gap hints that clinicians' relevancy judgement is additive: a medication will be considered relevant if it is pertinent to any of the complaints, so additional simultaneous complaints result in a higher proportion of relevant items. It also demonstrates the difficulty of SME ranking to handle the vast and unexpected variety of clinical situations.

For problems, the labeling clinicians achieved a lower IRR, (Flight's kappa of 0.58 vs 0.74 for medications). This gap may reflect the higher complexity of problems (compound concepts) compared to medications (an enumeration of relatively atomic entities), which may also be related to the lower accuracy of the SME ranking. The lower MAP achieved by the SME on problems (53.1-86.4% for problems vs. 92.2-98.2% for medications) may reflect the challenge to handle the nuances and complexity of clinical problems, especially when asked to provide a universal ruling ahead of time. Inspection of common SME mistakes revealed certain diagnoses (such as asthma, etc.) that were missing from their ranking, which was very much oriented towards cardiovascular causes of chest pain. This finding demonstrates the limits of manual ranking, including the inflexibility and subjectivity. The higher MAP achieved by the unsupervised method for problems may signal the ability to better handle this complexity. Data-driven methods may possibly overcome such challenges by adapting their knowledge sources to specific clinical situations or perspectives.

While addressing a similar problem, EHR searches present unique challenges from an IR perspective. Unlike web search engines, the scope of documents is predefined to the current patient (search among a group of patients is more relevant to the research and secondary EHR use cases and is targeted by systems like EMERSE [15]) and is much smaller. Thus, recall cannot be ignored. Unlike users of web search engines, there is no commonly accepted length of the result set (e.g. the first page, top 10, etc. [14]), and there is no natural limit to the number of items the clinician should review. While the clinicians are expected to eventually familiarize themselves with the full patient's record, clinically reasonable **ordering** of the items might be more helpful to ease the information load than an arbitrary cutoff, emphasizing the need to evaluate precision and recall across the full spectrum of values (using MAP) rather than at a specific cutoff. EHR search is typically described as an IR task, focusing on answering a specific information need. However, insights gained from clinicians during our work hinted that the task might behave more like a summarization one, with no (or very vague) particular information need. For example, immunosuppressive agents (medications blocking the immune systems used after transplantation) were considered by one clinician as relevant even in the absence of a physiological link to the chief complaint, since such high side effect medications are important by themselves. Such principles are more similar to the summarization task, focusing on the items' general importance independent of a specific information need(27). Still, the SME ranking and the manual labeling revealed a strong complaint specific component underlying the items' relevancy. It is likely that EHR search solutions will involve a mixture of these tasks, with a summarization-like task for the initial view of the chart (even allowing automation such as ranking the records according to the patient's chief complaint) with IR-like tasks for subsequent user-initiated queries.

The advantage SMEs had over the unsupervised methods on medications came at a high price: curation of the ranking was an expensive and lengthy process and required an additional step of translating the (largely informal) SME input to the specific data structures and terminologies used by the institute's EHR. The unsupervised methods, on the other hand, required no training or manual input at all, and use the existing patient data as is without any need for terminology mapping. Clinicians and organizations might opt to forego a certain level of accuracy to get a cheaper and ready-to-use solution. Moreover, general web search engines demonstrated that relevancy logic can be effectively harvested from user's decision (click-through data)(28). However, to maintain an active user base, a search engine must produce satisfactory and well-

ranked results. Unsupervised methods may help overcoming this Catch-22 and provide an acceptable baseline ranking logic at low cost.

Additional methods to estimate topical similarity exist. Word embeddings have been widely used to reveal latent semantic relationships, both taxonomic and topical. However, standard word embedding learning algorithms such as word2vec (both skip-gram and continuous-bag-of-words)(29), treat every word as an atomic and uniform unit. Therefore, learning representations that are specific to a chief complaint is not a straight forward task. Our early experiments with word embeddings for problems ranking revealed poor results, and therefore we did not pursue this path. PMI, despite its simplicity, provides a robust and interpretable ranking method, while remaining easy to use.

Much of the work on EHR search engines focuses on the narrative part of the medical record. The current study demonstrates both the need to address the structured elements and the ability of data-driven methods to answer this need. Future directions for our work include: expansion to additional data elements (laboratory tests and imaging reports), finding and ranking information appearing in narrative texts, and lastly, implementation of a real-life EHR search solution. Despite their usefulness for improving the relevancy logic, user activity data (e.g. which items are opened) are not always available from the EHR. Implementation of an EHR search solution will allow us to learn more about clinicians' actual needs, test our methods in real-life settings, and learn how to harvest user feedback to improve the relevancy logic.

Conclusion:

Data-driven unsupervised machine learning methods leveraging NLP and SOLR can efficiently approximate or outperform manual methods for relevancy ranking of clinical information. Clinical relevancy is situation specific and incorporates elements of summarization (non-specific importance) in addition to specific information needs.

Significance:

This work is significant in several ways. First, relying on SMEs to define all the information elements for a particular problem is not scalable and would require extensive time commitments to curate all the relevant information items and then rank them. Some are using this approach, such as the University of Wisconsin, for a variety of common diagnoses or conditions(30). We feel that a more automated approach using artificial intelligence (NLP and machine learning) that can learn from the user's interactions represent the future of information retrieval in EHRs. Second, the fact that our approach was able to retrieve 70-90% of relevant information for a particular complaint without any training is significant. This task of summarization of relevant information has been of interest to leaders in the field of search (e.g., Google). To the best of our knowledge, few if any have achieved our performance in EHR summarization and ranking. Lastly, most clinicians are not able to fully comprehend 70-90% of a patient's chart in a matter of seconds. Manually reviewing a patient's chart to reach this level of comprehension would typically take hours to days. Our methods, reduce this process to seconds, helping bring the relevant information to the clinicians' fingertips in a user-friendly graphical user interface.

Implications:

This research has several implications. The first most immediate implication is in the field of data summarization in EHRs. This is an important area that EHRs are currently inadequate or have not effectively solved. Providers are "drowning" in data and have limited time to both spend with the patient and review/identify all the salient and relevant information in the patient's chart. Our solution, we feel would be of direct benefit to the treating physician in routine patient care. Second, we feel this type of information retrieval could be of interest to researchers who might be interested in understanding relationships between different types of data as they relate to a particular patient context. Our solution may help uncover new patterns or relationships within the data that could provide valuable insights to a clinical problem or a predictor of a patient outcome. With the rapid innovation taking place in artificial intelligence, we think solutions like ours will be increasingly valuable to both clinicians and researchers, helping them improve not only the quality of care they deliver, but also improving patient safety.

Progressing to an R01 application:
Both the successes and the failures of this project point us toward the elements of an R01 application. The specific experimental leads to be followed include the following:

1. **Data science questions:** Our findings suggest that progress can be made in this research area by applying text mining and machine learning to EHRs. A natural follow-up question would be: can even more be learned by scaling to Big Data? The **scientific premise** behind this comes from a growing body of work on a variety of social media platforms, much of it from colleagues at the University of Pennsylvania and in the mental health domain.
2. **Deep learning questions:** Numerous papers have publicized the many recent successes of deep learning/neural networks. Our data thus provides us with the opportunity to investigate a topic that may be on the verge of being hot: when and why does deep learning in EHRs learning *fail?* The **scientific premise** here comes from broad literature on reproducibility failures in general, from a significant body of literature on reproducibility failures in the computational sciences, and from a small but growing body of literature on reproducibility problems in NLP.
3. **Upper-bound questions:** Our good inter-rater agreement results for medications suggests that we developed a good annotation methodology. Why, then, were the inter-rater agreement results so much lower for problems? Why were the machine learning methods better than the SME? There are at least two reasonable hypotheses here: (1) the calculation of inter-rater agreement is flawed (presumably because of assumptions about calculating the probability of chance agreement in the standard formulae), or (2) the task is harder. If it is the case that then task of annotating problems is more difficult than the task of annotating medications, why? Being able to answer that question could tell us a lot about how to generalize the approach that we are developing for search in EHRs beyond back and chest pain to pain in general; beyond pain to other symptomatology; and beyond the Emergency Department to other areas of clinical practice. The **scientific premise** here comes from research on inter-rater agreement going back to the early days of wartime propaganda detection and continuing up to last year's MEDINFO conference.

List of Publications:

1. Cohen KB, Goss FR, Zweigenbaum P et al. Translational Morphosyntax: Distribution of Negation in Clinical Records and Biomedical Journal Articles. Studies in health technology and informatics. 2017;245:346-50. Epub 2018/01/04. PubMed PMID: 29295113.

2. Cohen KB, Xia J, Zweigenbaum P, Goss FR, et al. Three Dimensions of Reproducibility in Natural Language Processing. LREC International Conference on Language Resources & Evaluation : [proceedings] International Conference on Language Resources and Evaluation. 2018;2018:156-65. Epub 2018/06/19. PubMed PMID: 29911205; PubMed Central PMCID: PMCPmc5998676.

3. Korach TZ, Zhou L, Goss FR, et al. Unsupervised clinical relevancy ranking of structured information (in process).

References:

1. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med. 1999;74(8):890-5.
2. Meystre SM, Savova GK, Kipper-Schuler KC, et al. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics. 2008:128-44.
3. Johnson S. Natural language processing in biomedicine. 2nd ed: Springer Science & Business Media; 2000.
4. Sager N, Lyman M, Bucknall C, et al. Natural language processing and the representation of clinical data. Journal of the American Medical Informatics Association : JAMIA. 1994;1(2):142-60. Pmc116193
5. Spyns P. Natural language processing in medicine: an overview. Methods Inf Med. 1996;35(4-5):285-301.
6. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. Journal of the American Medical Informatics Association : JAMIA. 2005;12(4):448-57. Pmc1174890
7. Penz JF, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. Journal of biomedical informatics. 2007;40(2):174-82.
8. Petratos GN, Kim Y, Evans RS, et al. Comparing the effectiveness of computerized adverse drug event monitoring systems to enhance clinical decision support for hospitalized patients. Applied clinical informatics. 2010;1(3):293-303. Pmc3631899
9. Kilbridge PM, Noirot LA, Reichley RM, et al. Computerized surveillance for adverse drug events in a pediatric hospital. Journal of the American Medical Informatics Association : JAMIA. 2009;16(5):607-12. Pmc2744710
10. Michelson JD, Pariseau JS, Paganelli WC. Assessing surgical site infection risk factors using electronic medical records and text mining. American journal of infection control. 2014;42(3):333-6.
11. Hou JK, Chang M, Nguyen T, et al. Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. Digestive diseases and sciences. 2013;58(4):936-41. Pmc3974588
12. Hanauer DA, Mei Q, Law J, et al. 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). Journal of biomedical informatics. 2015;55:290-300.
13. Lowe HJ, Ferris TA, Hernandez PM, et al., editors. STRIDE–An integrated standards-based translational research informatics platform. AMIA Annual Symposium Proceedings; 2009: American Medical Informatics Association.
14. Gregg W, Jirjis J, Lorenzi NM, et al., editors. StarTracker: an integrated, web-based clinical search engine. AMIA annual symposium proceedings; 2003: American Medical Informatics Association.
15. Natarajan K, Stein D, Jain S, et al. An analysis of clinical queries in an electronic health record search utility. International journal of medical informatics. 2010;79(7):515-22.
16. Aizawa A. An information-theoretic perspective of tf–idf measures. Information Processing & Management. 2003;39(1):45-65.
17. Mihalcea R, Tarau P, editors. Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing; 2004.
18. Chen J, Yu H. Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients. Journal of biomedical informatics. 2017;68:121-31.
19. Levy O, Goldberg Y, editors. Dependency-based word embeddings. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); 2014.
20. Nalisnick E, Mitra B, Craswell N, et al., editors. Improving document ranking with dual word embeddings. Proceedings of the 25th International Conference Companion on World Wide Web; 2016: International World Wide Web Conferences Steering Committee.
21. Allen M, Currie LM, Graham M, et al. The classification of clinicians' information needs while using a clinical information system. AMIA Annu Symp Proc. 2003:26-30.
22. Bates DW, Cohen M, Leape LL, et al. Reducing the frequency of errors in medicine using information technology. J Am Med Inform Assoc. 2001;8(4):299-308.
23. Kessels RP. Patients' memory for medical information. Journal of the Royal Society of Medicine. 2003;96(5):219-22.

24. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Family practice. 2006;23(2):253-63.
25. Pickering BW, Herasevich V, Ahmed A, et al. Novel Representation of Clinical Information in the ICU: Developing User Interfaces which Reduce Information Overload. ACI. 2010;1(2):116-31.
26. RxNorm [Available from: https://www.nlm.nih.gov/research/umls/rxnorm/overview.html.
27. Nenkova A, Maskey S, Liu Y. Automatic summarization. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011; Portland, Oregon. 2002468: Association for Computational Linguistics; 2011. p. 1-86.
28. Joachims T, editor Optimizing search engines using clickthrough data. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining; 2002: ACM.
29. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.
30. Problem List MD [Internet]. [cited Dec 12, 2018]. Available from: https://problemlist.org.