International Journal of Medical Informatics 105 (2017) 121–129

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Automated problem list generation and physicians perspective from a pilot study

Murthy V. Devarakonda[a,*], Neil Mehta[b], Ching-Huei Tsou[a], Jennifer J. Liang[a], Amy S. Nowacki[b], John Eric Jelovsek[b]

[a] IBM Research, USA
[b] Cleveland Clinic, USA

ARTICLE INFO

Keywords:
Electronic health records
Longitudinal patient records
Problem list
Machine learning
Natural language processing
IBM Watson

ABSTRACT

Objective: An accurate, comprehensive and up-to-date problem list can help clinicians provide patient-centered care. Unfortunately, problem lists created and maintained in electronic health records by providers tend to be inaccurate, duplicative and out of date. With advances in machine learning and natural language processing, it is possible to automatically generate a problem list from the data in the EHR and keep it current. In this paper, we describe an automated problem list generation method and report on insights from a pilot study of physicians' assessment of the generated problem lists compared to existing providers-curated problem lists in an institution's EHR system.

Materials and methods: The natural language processing and machine learning-based Watson[1] method models clinical thinking in identifying a patient's problem list using clinical notes and structured data. This pilot study assessed the Watson method and included 15 randomly selected, de-identified patient records from a large healthcare system that were each planned to be reviewed by at least two internal medicine physicians. The physicians created their own problem lists, and then evaluated the overall usefulness of their own problem lists (P), Watson generated problem lists (W), and the existing EHR problem lists (E) on a 10-point scale. The primary outcome was pairwise comparisons of P, W, and E.

Results: Six out of the 10 invited physicians completed 27 assessments of P, W, and E, and in process evaluated 732 Watson generated problems and 444 problems in the EHR system. As expected, physicians rated their own lists, P, highest. However, W was rated higher than E. Among 89% of assessments, Watson identified at least one important problem that physicians missed.

Conclusion: Cognitive computing systems like this Watson system hold the potential for accurate, problem-list-centered summarization of patient records, potentially leading to increased efficiency, better clinical decision support, and improved quality of patient care.

1. Background and significance

Electronic Health Records (EHRs[2]) are expected to improve patient outcomes by providing the most important patient information in a single location [1]. A common way to provide the holistic information is in the form of a patient-centered problem list [2], by itself or as part of a summary [3,4]. Ideally, the problem list would include all clinically significant issues that a care provider should consider when managing a patient. It should be accurate, inclusive, and up to date [5,6]. At the same time, it should not contain redundant or irrelevant items that distract the clinician, reduce efficiency, or lead to inappropriate actions [7–9].

Existing EHR systems provide problem lists that are populated by care providers, but they tend to suffer from the problems outlined. A patient's problem list in an EHR system requires the clinician to routinely maintain and update it with each encounter. Unfortunately, this is haphazardly performed, leading to inaccurate, incomplete and duplicative lists.

* Corresponding author.
E-mail address: mvd@acm.org (M.V. Devarakonda).

[1] ***Watson***, mentioned here, refers to ***new methods*** developed ***for patient record text analytics***, including the automated problem list generation, based on the core Watson text processing tools for sentence segmentation, parsing, and named entity linking.

[2] In this article, we use the terms ***EHR*** and ***EHR system*** to mean commercial and non-commercial electronic health record systems, and we use the term ***patient record*** to mean all the patient data, including clinical notes, reports, medications ordered, procedures, labs, and demographics.

http://dx.doi.org/10.1016/j.ijmedinf.2017.05.015
Received 28 December 2016; Received in revised form 21 May 2017; Accepted 23 May 2017

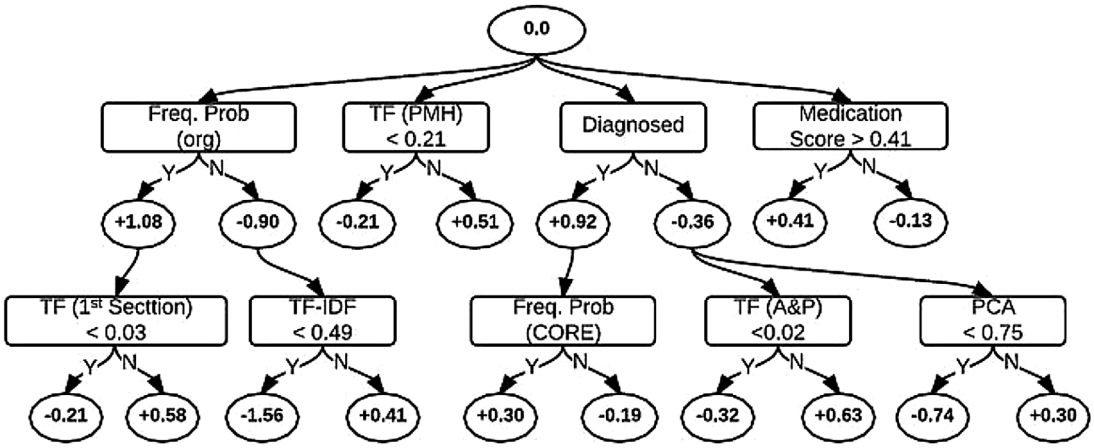

Fig. 1. The first two levels of the Watson problem list ADT model. Rectangular nodes (splitting nodes) represent conditions while oval nodes (prediction nodes) contain numeric scores that are used to make a prediction. The full model is 7 levels deep and contains 50 splitting nodes.

Efforts to automate the creation of the problem list based on EHR data have been made, but were limited to 80 prespecified problems and utilized techniques that may not scale to all possible problems [10–12]. With the advent of advanced natural language processing (NLP) and machine learning techniques [13–15], it is possible for automatic creation and updating of a patients' problem list based on the data in the EHR. We have developed a system based on these technologies, present its key features and report a pilot study of its usefulness from a physician's perspective. The study was conducted in USA.

2. Method: automated problem list generation

The Centers for Medicare and Medicaid Services [16] of the USA defines a patient's problem list as, "a list of current and active diagnoses as well as past diagnoses relevant to the current care of the patient" for the federal meaningful use program. We made it more concrete by assuming the context to be a comprehensive health assessment and by including clinically relevant past procedures in the problem list.

2.1. Framing the machine learning task

We framed the problem list generation as a supervised binary classification task, which decided if a candidate problem is a true problem or not. Candidate problems were identified from the textual narrative of clinical notes using NLP (details in next subsection). Several features were extracted for each candidate, and a trained model was applied to the feature values to determine a confidence score for each candidate problem. If the score was above a learned threshold, then the candidate problem was considered a true problem. We engineered features based on clinical, lexical, structural, temporal, and epidemiological aspects of candidate problems, and the features were extracted using NLP from narratives in all types of clinical notes and reports, and from structured parts of a patient record. The machine learning model was trained on a gold standard that was manually created by medical experts (who were not participants in the pilot study) using 399 of the total 996 de-identified patient records obtained from Cleveland Clinic (Cleveland, Ohio, USA) under an Institutional Review Board approval. The following paragraphs provide details of this overview, starting with the step of candidate problems identification.

2.2. Candidate problems identification

Clinical notes were first preprocessed to detect and link medical concepts to the Unified Medical Language System (UMLS®) dictionary entries [17] using Watson NLP components, which provided a function similar to cTAKES [18], but with additional refinements (that are unimportant for the discussion here). Out of these concepts, the subset belonging to the semantic groups *Disorders, Procedures, Physiology,* and *Living Beings* were considered as candidate problems, so long as these concepts had a mapping in the CORE (Clinical Observations Recordings and Encoding) subset of SNOMED CT (Systematized Nomenclature of Medicine − Clinical Terms) [19]. The CORE subset represents a candidate problem space of more than 6,000 problems. Therefore, the candidate problem list would be typically very large (about 15–1) compared to a patient's actual problems. The goal of the machine learning task was to reduce this large set to the actual problem list.

2.3. Features

We engineered a set of features for the machine learning model which would reduce the candidate problems to the actual problem list. At the time of writing this article, the model used 211 features: 18 multi-valued categorical features (e.g. ICD9 category), 28 binary features (e.g. did the candidate problem appear in a clinical note in the last 3 months?), 66 numerical features (e.g. term frequency), and 100 resulting from normalization of the numerical features in various ways (e.g. normalized to the length of the record). For each candidate problem, the features were extracted from the passages (in the clinical notes) where the candidate problem was mentioned, its surrounding context, as well as from the structured data elements related to the problem in the patient record (e.g. diagnosis codes, medications, laboratory test results, and procedures). Details of the features were described in [20]. However, later in this article, we will further discuss some of the features which show how our model represents clinical thinking of medical experts, and therefore played an important role in our trained model.

2.4. Machine learning model

We used the Alternating Decision Tree (ADT) [21], a boosting-based discriminative classifier, to model our classification task due to its accuracy and interpretability. ADT alternates between two types of nodes: a prediction node and a splitter node, represented by an oval and a rectangle in Fig. 1, respectively. During each training step, all features are considered and the best split (i.e., minimizing the training error) is added to the tree. ADT can be viewed as a collection of rules, which predict whether a candidate problem is a true problem or not by adding up the scores of prediction nodes in all valid paths. For example, referring to the ADT in Fig. 1, if a candidate problem is *diagnosed* by a physician and is in the *frequent problems of CORE* then one path for the candidate yields a score of (0.0 + 0.92 + 0.30 =) 1.22. If the aggregate

Table 1
Key features of the Watson problem list generation model.

Feature	Description
Freq Prob (Org)	Frequent problems (top 25% of most frequently diagnosed problems in our patient records data set)
TF (PMH)	Term frequency; frequency of the problem in the past medical history section of the patient record
Diagnosed	Problem has ever been diagnosed (structured data)
Medication Score	Problem is related to any active medication
TF (1st Section)	Term frequency; frequency of the problem in the 1st section of a clinical note
TF-IDF	Term frequency multiplied by Inverse document frequency
Freq Prob (CORE)	The average frequency of occurrence of the problem across all institutions involved in contributing to the SNOMED CT CORE subset
TF (A & P)	Term frequency; frequency of the problem in the "assessment and plan" section of the patient record
PCA	Probability that the given term (covered text) is linked to the UMLS concept

score of all such paths is above an automatically learned threshold, then the problem is a true problem as per the model. In our method, ADT was trained and tested using 10-fold cross validation on 399 de-identified patient records and ground truth problem lists. The top two levels of the model are shown in Fig. 1 and the features, mentioned in the rectangles of Fig. 1, are described in Table 1.

3. Modeling clinical thinking

The key to developing the automated problem list generation method was engineering features that model the clinical thinking of medical experts and physicians. In this section, we describe how the resulting ADT model accomplished this, using the top two levels of the model, shown in Fig. 1, and features in Table 1.

3.1. Modeling the context of problem discussion

In a clinical note, a candidate problem may be discussed in many different contexts. But when it is discussed in the past medical history or in the assessment and plan for the present visit, the candidate problem is likely to be a true problem. A human reader of the note, such as a physician or a nurse, would look in these areas of the note for the patient's medical problems. To represent this intuition, we engineered features that represent how frequently a candidate problem is discussed in these areas, having first developed NLP analytics to automatically distinguish these areas (i.e. "sections") of a note. The features that represent the term frequency in the past medical history section (*TF (PMH)* in Fig. 1) and in the assessment and plan section (*TF(A & P)* in Fig. 1) turned out to be effective in the model.

3.2. Modeling disease prevalence in a population

Epidemiological studies show that certain diseases are more common in certain populations of patients. An automated system should model this observation to be effective. Therefore, we used two features that represent whether a candidate problem is a common disease in the population: (1) Whether it is in the top 25% of the most frequent problems in our patient records data set (all 996 records); (2) Occurrence frequency in the CORE subset of SNOMED CT, aggregated from the hospitals contributing to the CORE subset development. For example, *diabetes mellitus* is one of the most common diseases, so *diabetes mellitus* as a candidate problem will have a high value for these features. These prior probability features, denoted as *Freq Prob (org)* and *Freq Prob (CORE)* in Fig. 1, proved to be effective in our model.

3.3. Modeling the patient's medical treatment

Physicians often modify or create a patient's problem list when they encounter the patient for the first time by working through the patient's current medications. A treatment requiring medication usually indicates an active and current problem. We incorporated this as a feature, designated as *Medication Score* in Fig. 1, by automatically scoring a clinical association [22] between a candidate problem mentioned in a clinical note and the patient's active medications at that time. A high association score with any of the active medications results in a high feature score for the problem. As an example, a patient being on *Atorvastatin* gives a high score to the candidate problem *Hyperlipidemia*.

3.4. Modeling formal diagnoses by physicians

After each patient's visit, a physician provides one or more diagnostic codes in the record for billing purposes. These codes represent the physician's diagnostic decisions for the visit. Although not all visit diagnostic codes should go on a problem list, some of them do belong there. Therefore, this data provides additional evidence (albeit noisy) informing the problem list, and is represented by the feature, *Diagnosed*, in Fig. 1. For example, chronic and acute problems addressed during the visit, such as *hypertension* and *upper respiratory infection*, would be recorded in the visit diagnostic codes, and they will have high scores for the feature. This strengthens the case for hypertension (as well as for upper respiratory infection). Other features of the model weaken transient problems, such as upper respiratory infection (see section below). Modeling the visit diagnoses provides evidence supporting formally diagnosed problems, while weakening the case for candidate problems that were never formally diagnosed.

3.5. Modeling chronic conditions

Many patients suffer from chronic and recurring conditions such as hypertension and asthma, so an automated system needs to be able to distinguish these persistent problems from transient medical problems such as influenza. We captured the chronic nature of problems with the classic information retrieval features involving term frequency (TF) and inverse document frequency (IDF), *TF-IDF*, in Fig. 1. Further, we specialized the term frequency by the context in which candidate problems were mentioned in the clinical notes. For example, if a candidate problem was discussed in the beginning of the note (*TF (1st Section)*), or in the assessment and plan part of the note (*TF (A & P)*), it suggested the problem was being actively discussed and managed by a physician, and present at the time of the note. One might suggest denoting some medical problems as always chronic and others as always transient, but it should be noted that certain medical conditions (e.g. urinary tract infection) can be transient in some patients while recurrent or chronic in others. The term frequency features and their spatial distributions play an important role in the model, and thus enabled our model to capture chronic conditions.

3.6. Modeling the confidence in clinical terms recognition

Since we used UMLS for named entity linking and since UMLS is a metathesaurus of multiple vocabularies, a term in a clinical note could be linked to multiple concept identifiers in UMLS. For example, the term *diabetes* could refer to a disease (C0011849) or it could refer to a finding (C0241863). To assist in resolving these identifiers, each implying a different meaning to the term, we calculated a popularity score, designated as *PCA* in Fig. 1, for each UMLS concept identifier of the term. PCA was higher if the number of vocabularies that contributed this term to UMLS was higher. As it helped to resolve the meaning of the term in a clinical note, it proved to be another effective feature in our model.

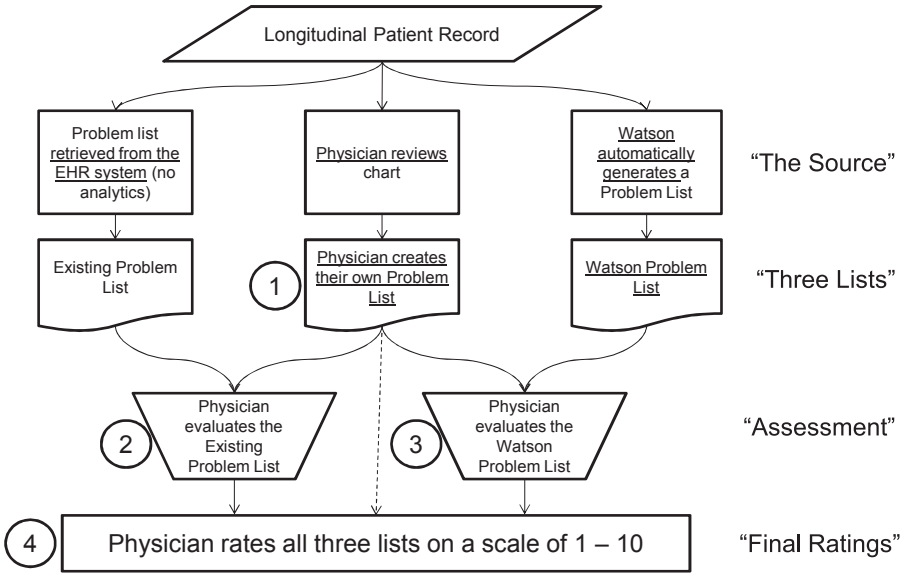

Fig. 2. The assessment for a patient record consisted of a series of steps each physician carried out, including creating their own problem list, evaluating the existing problem list in the EHR, evaluating the Watson generated problem list, and finally rating all the three problems lists on a 10-point response scale.

4. Methods and materials: pilot study

The pilot study was conducted for five weeks in late 2015 at Cleveland Clinic. (The study is illustrated in Fig. 2 and described throughout this section.) A convenience sample of 10 internal medicine attending physicians and senior residents were recruited to participate in the study. Their participation was entirely voluntary and they were not compensated. Fifteen randomly selected, de-identified longitudinal patient records from the healthcare institution were also selected. To be considered for inclusion in this study and to ensure sufficient data for analysis, each patient record was required to have a minimum of three encounters and 200 clinical notes. The selected, de-identified patient records were extracted from the commercial EHR system and forwarded to the Watson method [20], in the form of customized HL7 Continuity of Care Documents [23], for generating problem lists. The EHR system is commonly used in the USA and in several other parts of the world, and claims to hold over 50% of US patient records.

The Watson generated problem lists were made available to the physicians in a standard Web browser. They were given a key to map a patient's Watson ID to the patient record number (e.g. MRN) in the institution's EHR system. Each physician was randomly assigned 5 of the 15 patient records. First, they were asked to create a problem list for each patient record, in the context of a comprehensive health assessment. Then, they were asked to compare the existing EHR problem list and the Watson generated problem list to their own problem list, and rate each of the three lists on a scale of 1–10 on their usefulness in patient care.

4.1. Assessment steps

The assessment consisted of a series of steps carried out by physicians (Fig. 2) using the Web application that was developed for this experiment and a standard Web browser, and the steps are described below in detail.

4.2. Steps 1 and 2

For each patient record, physicians were first asked to review the record in the healthcare institution's commercial EHR system and create a problem list. The full patient record in the institution's EHR system, including the existing problem list for the patient, was available to them as a reference source for creating the problem list. The physicians entered each problem in the Web application (Fig. 3a), and indicated whether the problem was present on the existing problem list (E) in the patient record. Thus, physicians provided an assessment of problems in E while creating their own list (P).

4.3. Step 3

Once a participant completed steps 1 and 2 of the experiment, he/she was presented with a new screen containing the Watson generated problem list (Fig. 3b). At this stage, they continued to have access to the full patient record in the institution's EHR system, and they could reference their own problem list created earlier, but were not allowed to change what they had entered in the Web application in step 1. Participants sequentially reviewed and assessed each of the IBM Watson generated problems as correct, acceptable, or incorrect. If acceptable, the participants were further asked to specify if it was acceptable but too general, too specific, or redundant. Similarly, if incorrect, they were further asked to specify if it was too general, transient/resolved, or a non-problem.

For each Watson generated problem, participants also indicated if it was on their problem list, and rated the clinical importance of the problem as very important, important, somewhat important, or unimportant. Clinically important problems are defined as problems that the physician would like to be aware of when taking care of a patient, considering the effects of the problem on patients' risks of future diseases, quality of life, life expectancy, morbidity and mortality.

4.4. Step 4

After assessing the Watson generated problems, the participants were asked to rate each of the three lists – their own list (P), the Watson generated list (W), and the existing EHR system list (E) – for their usefulness, in the context of a comprehensive health assessment, on a response scale of 1–10, 1 being least useful and 10 being most useful. This was also entered in the Web application (Fig. 3c).

5. Hypothesis testing and problems missed

To test our hypothesis, i.e. if physicians rate the Watson problem list (W) better than the existing EHR system list (E), the response scale ratings were compared pairwise using the Wilcoxon signed-rank test because of the ordinal nature of the ratings. P-values less than 0.05 were considered statistically significant.

In addition, because we asked physicians to indicate whether each Watson generated problem was on their problem list, we determined whether physicians missed any problems that Watson found and their

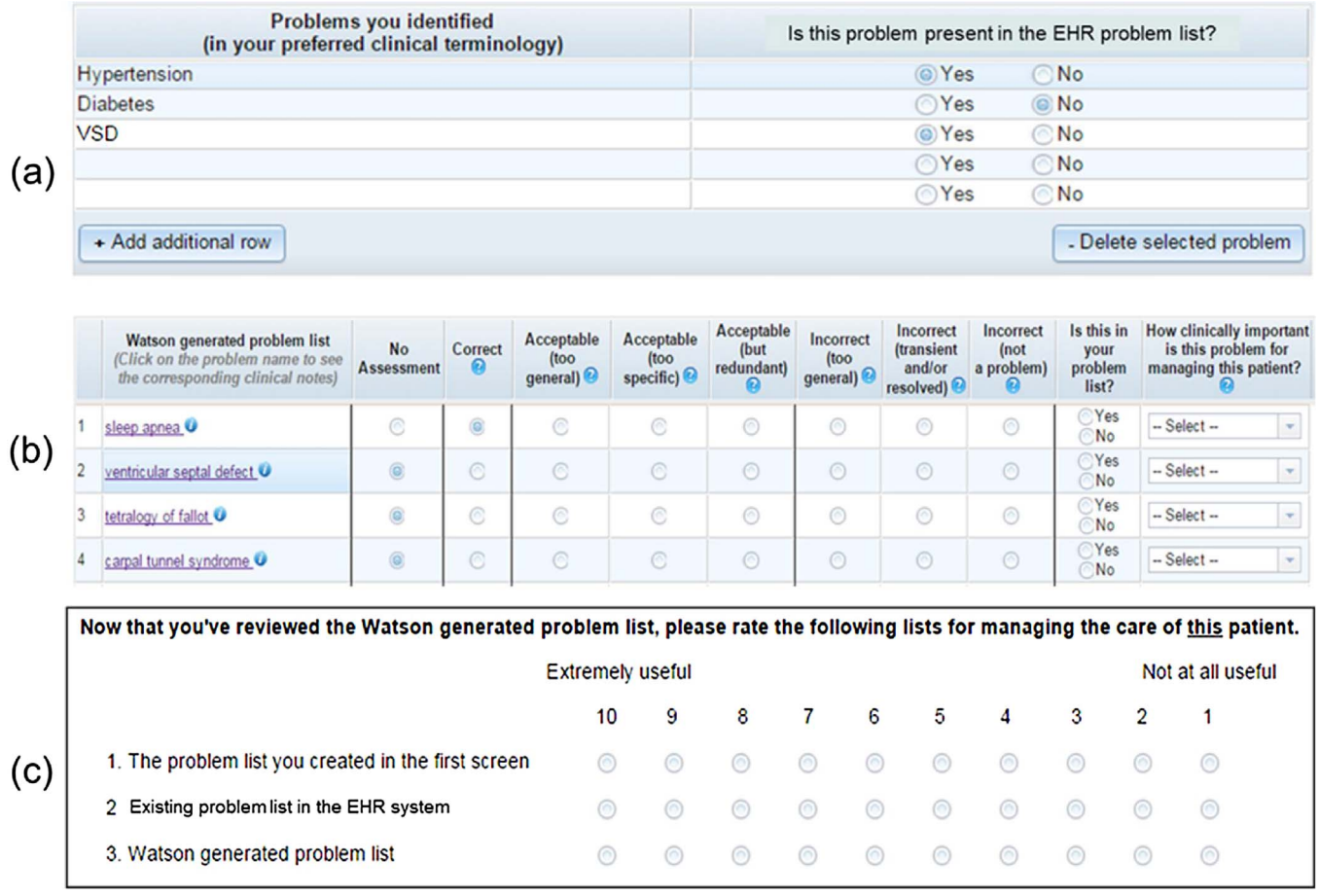

Fig. 3. Screen images of the Web application interface used by the physicians in the assessment; (a) was used by the physician to create their own problem list and evaluate the existing problem list in the EHR, (b) was used by the physician to evaluate the Watson problem list, and (c) to rate all three problem lists on a 10-point scale.

clinical importance as perceived by the physician.

6. Gold standard creation and Watson accuracy

As is common in information retrieval, we used recall (R), precision (P), and F1 and F2 scores to determine Watson problem list generation accuracy in this study. Recall is also known as sensitivity and precision is also known as positive predictive value. F scores measure the effectiveness of the system in accomplishing the task; F1 provides a balanced measure of recall and precision, and F2 provides a higher recall-weighted measure. Specificity, also known as true negative rate, is not useful in a task like this because true negatives (i.e. non-problems) are significantly larger than true positives (i.e. actual problems of a patient), and so specificity rarely yields a meaningful accuracy distinction in this task. True positives (T_P), false negatives (F_N), and false positives (F_P) were determined based on a gold standard, and the following equations were used to calculate R, P, F1, and F2:

$$R = \frac{T_P}{T_P + F_N} \quad P = \frac{T_P}{T_P + F_P} \quad F1 = \frac{2PR}{(P + R)} \quad F2 = \frac{5PR}{(4P + R)}$$

The gold standard needed for the accuracy calculations was created as follows:

- For each patient record, it was assumed every problem identified by a physician was correct and it was added to the gold standard problem list (note that most patient records were assessed by two physicians).
- If a physician identified a Watson correct problem as missing from his/her list, and rated it as a *very important* or *important* problem, it was also added to the gold standard list for the patient record.
- Any duplicates added to the list because of the above two steps were removed (for example, duplicates can appear if one physician identified a problem, and another physician missed it, but rated it as important).

The gold standard resulting from this process was the set of problems from the physicians' lists, plus any missed problems that were rated as *very important* or *important* for the patient record. Note that this derived gold standard may miss some true problems of the patient, when such a true problem was missed by both physicians and Watson. This may result in a higher recall than using a gold standard that was developed with a process involving adjudication and repeated vetting [24], as was used in the previous study [20], but it is feasible and relevant to the present study.

While the plan called for two physicians to assess each patient record, there was a possibility that some patient records would be assessed by a single physician. In such a case, the patient records assessed twice would contribute more weight to the accuracy calculations than the others. To remedy this, we averaged true positives, false positives, and false negatives for each patient record that had multiple assessments, and showed these averages in the confusion matrix (see below) and used them in calculating the accuracy metrics.

6.1. Free-text write-in comments

At the end of each assessment, physicians were asked to optionally respond to the following open ended questions using free-text responses:

- Please identify one thing that you like about the Watson generated problem list
- Please suggest one improvement for the Watson generated problem list

Physicians were given an option to enter the free-text responses in the Web application. Two of the authors (NM and MVD) identified common themes among the comments, and for each theme, 1–2 insightful and representative comments were selected and reported in Section 7.4.

7. Results

Out of the ten physicians approached for the study, five attending physicians completed assessment of all five of their assigned patient records, one chief resident completed two of the five assigned patient records, and the remaining four senior residents did not complete any reviews. As a result, we obtained a total of 27 assessments from 6 participants, where an assessment means a participant completed all the required steps described above for a patient record. Twelve records were assessed by two participants and three records were assessed by one participant each. The experiment resulted in evaluations of 732 Watson generated problems and 444 problems in the existing EHR patient records.

7.1. Hypothesis test results

Results of the pairwise comparison of the scale ratings are shown in Figs. 4 and 5. As expected, physicians rated their own problem list (P) significantly higher than the Watson generated problem list (W) and the existing manually maintained EHR problem list (E). However, participants also rated W significantly higher than E. The mean (standard deviation) of scale ratings of P, W, and E were 8.4 (1.2), 7.4 (1.6) and 5.8 (2.5), respectively. All pairwise comparisons between the three groups (P-W: $p = 0.005$; P-E: $p < 0.0001$ and W-E: $p = 0.02$) were significant. Out of the 15 patient records, when compared to the existing manually maintained EHR problem list, the Watson generated problem list was rated higher in 10 cases, the same in two cases, and lower in three cases.

7.2. Problems missed

Watson identified an average of 4.33 problems per assessment which physicians missed and were subsequently rated by them as 'important' or 'very important'. In total, physicians missed 117 important/very important problems in the study. They missed at least one important or very important problem that Watson identified, in 24 assessments out of 27 (Table 2).

7.3. Watson accuracy

Table 3a shows the confusion matrix for the Watson problem list accuracy analysis and Table 3b shows the accuracy metrics – recall (sensitivity), precision (positive predictive value), and F scores (harmonic means of recall and precision), which were defined fully in subsection 3.3. The false positives are larger than the false negatives by nearly three times in the confusion matrix. This result is a consequence of configuring Watson to optimize on recall, even at the cost of additional "noise" in the problem list (i.e. reduction in precision). This is also reflected in the F scores, where the F2 score (0.799) is substantially higher than the F1 score (0.740). Using the same gold standard, the accuracy metrics for P (the physician's own problem list) are recall of 0.67 and precision of 1.0 (follows from the gold standard definition), which translates to F1 of 0.79 and F2 of 0.71.

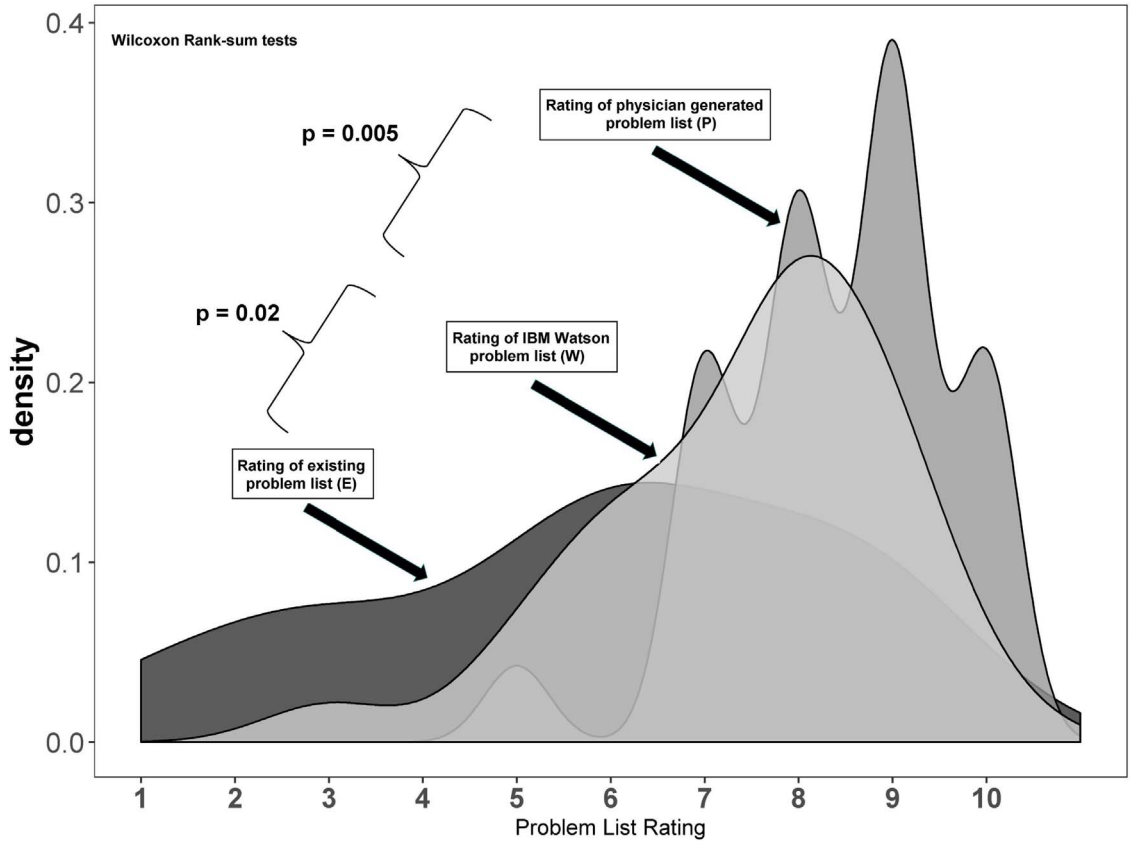

Fig. 4. Pairwise comparison of physicians' problem list ratings shown as density functions.

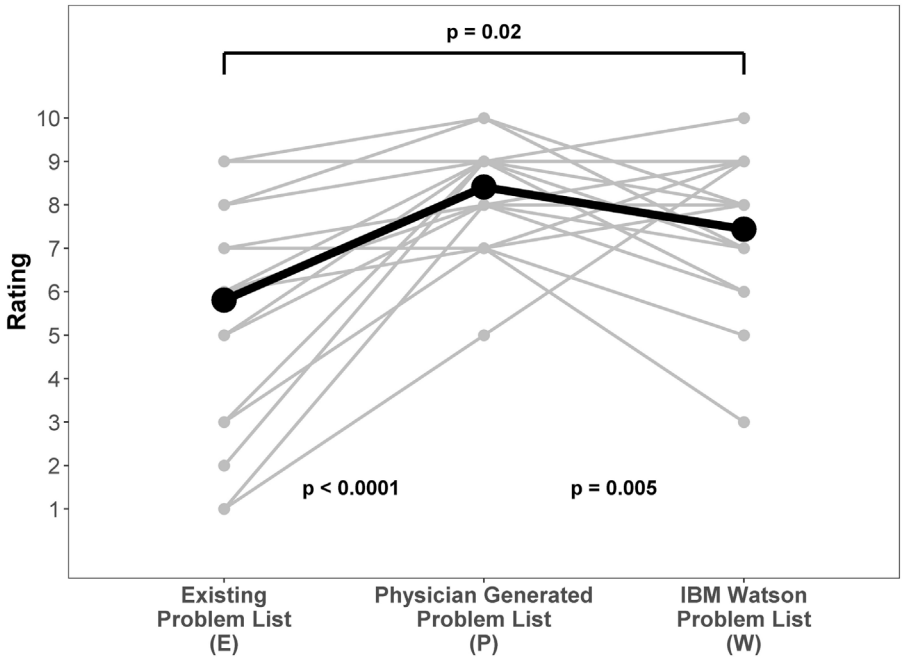

Fig. 5. Pairwise comparison of physicians' problem list ratings shown as a stick diagram; Each line represents the average rating of the problem lists of a single patient record.

Table 2
Problems missed in the physicians' problem lists (total assessments = 27).

Problem Importance as identified by physicians	Number (%) of assessments with missed problems	Number of problems missed	Average number of problems missed
Very Important	13 (48%)	29	1.07
Very Important or Important	24 (89%)	117	4.33

Table 3a
The confusion matrix for the Watson problem list accuracy analysis, showing true positives, false positives, and false negatives.

		Watson	
		True	False
Derived gold standard	True	269 (T_P)	50 (F_N)
	False	139 (F_P)	–

Table 3b
Watson problem list accuracy analysis from this assessment; Results from the previous study [20] (which was based on medical experts generated problem lists that were adjudicated by an MD) are provided for comparison purposes.

	Derived gold standard (in the current study)	Adjudicated gold standard (in the previous study)
Recall	0.843	0.813
Precision	0.659	0.567
F1 Score	0.740	0.668
F2 Score	0.799	0.748

7.4. Free-text write-in comments

Twenty-one out of 27 assessments had free-text responses for the question, *please identify one thing that you like about the Watson generated problem list*, and 23 out of 27 assessments had free-text responses to the question, *please suggest one improvement for the Watson generated problem list.* The following seven common themes were observed in the responses:

- Watson found diagnoses that physician had missed
- Watson was very complete/thorough
- Watson supported clinical reasoning
- Watson listed a diagnosis that was not well supported
- Watson list was broad and included redundant and non-active problems
- Watson missed diagnoses
- Natural language processing errors in Watson

Tables 4a and 4b show insightful and representative comments for each of the themes, as entered by the physicians. The comments suggest that physicians like Watson's thorough analysis of the patient record (which results in identifying problems they sometimes miss) and its potential impact on patient care. The comments also suggest what should be improved in Watson's problem lists, e.g. reducing redundancy, filtering out non-problems, avoiding poorly supported problems, and improving natural language processing.

8. Discussion

This study of automatically generated Watson problem lists suggests that cognitive computing systems can generate problem lists which physicians find more useful than the manually maintained EHR problem lists. By using natural language processing, machine learning, information extraction, and other advanced analytics on a longitudinal patient record, Watson could generate a more complete and useful problem list when compared to an EHR problem list that is only manually maintained. The method is not limited to a subset of diseases, and therefore broadly scalable compared to identification of problems from a prespecified list of 80 problems [10,11].

Note that the manually maintained EHR problem list is accessible to all the providers, and when one provider updates it, the updates are visible to all providers. Being a multi-specialty hospital, a patient record is accessed and updated by all specialties, including cardiology, nephrology, internal medicine, behavioral health, rheumatology, and orthopedics. The patient problem list is coded using ICD-9 with the help of the EHR system, and it is retrieved by the physicians using the overview panels provided by the EHR system.

The fact that physicians missed several important problems that Watson identified, also demonstrates its potential value. Necessary facts are not well organized or easily accessible in a commercial EHR system,

Table 4a
Physicians' free-text responses to what they liked about the Watson generated problem list.

Theme	Comments (physician's anonymized id in the parenthesis)
Physician missed diagnoses	*It was able to search significantly more thoroughly the past medical records than I was. I only look at the most recent, but Watson was able to pick up on a very remote DVT (2003) and very remote pre-malignant polyp (most recent was only hyperplastic) (3452)* *(Watson) found the history of recurrent UTIs (3807)* *(Watson) found hx of hyperparathyroidism (3807)*
Complete/Thorough	*With a multitude of records to inspect, I look at higher-yield documents like discharge summaries, outpatient notes, procedures, etc. Watson can look at every line of text and pick up on things the physician who discharges the patient may not have even known about. I quickly saw how sick the patient was, and ignored many of the insignificant facts in the chart (such as cataracts…) (3452)* *Comprehensive – won't miss a diagnosis. (5413)*
Supported clinical reasoning	*Made me rethink the reasons for urinary incontinence which was not on my problem list – may be related to a procedure (prostatectomy) listed below. (4472)*

and humans tend to perform poorly when the task requires foraging through a long and poorly organized patient record. The task is not only tedious and time consuming, but also requires significant expertise (and even a dialog among experts). There is a clear need to relieve physicians from this laborious task, while allowing them to verify and validate the outcome of an automated system. Therefore, Watson problem list generation may complement physicians' efforts by identifying important problems that they might otherwise overlook.

This study identified several areas for improvement. The number of incorrect problems, especially the transient or resolved problems, produced by Watson negatively impacted physicians' perception of its usefulness. While improving the Watson algorithms has the potential to decrease this number, Watson can also be configured to reduce the number of incorrect problems at the risk of missing some correct problems. As described in the earlier report [20], Watson uses a threshold to filter out non-problems from (what Watson considers as) true problems. This threshold can be set to maximize the F2 score (recall-oriented) or the F1 score (recall-precision balanced). For this study, we configured the threshold to maximize F2, with the assumption that it is easier for physicians to reject non-problems presented to them than to search for true problems buried in the vast amount of data. Physicians seemed to react negatively to this increased noise level as seen from some of the free-text responses, and it is a subject for further investigation.

It is instructive to explore how the Watson accuracy, measured here, compares with the results based on the gold standard as reported earlier [20], where the gold standard was developed involving multiple medical experts, adjudication of their work, and final vetting based on the Watson output. Watson list accuracy is somewhat higher in this study than in the previous study, but they are relatively close, despite significant differences in the data set size and the gold standard creation approach.

The physicians' free-text responses explain and support several observations from the data discussed so far. Positive comments about Watson's thoroughness in problem list generation are consistent with the fact that physicians sometimes missed true problems (and could be helped by Watson) and with the high recall of the Watson problem list in the accuracy analysis. Their concerns about the redundancy and non-problems in the Watson problem list (due to NLP errors) are also reflected in the ratings, and in the relatively lower precision (compared to the recall) of the Watson problem list.

9. Conclusions

Physicians are burdened with the task of assimilating vast amounts of information in the EHR systems. Despite spending a lot of time and effort, and despite their best intentions, they tend to miss important problems. The existing problem lists in patient records are inaccurate and maintenance of the problem lists is not currently a part of the physician workflow. An accurate problem list can have significant benefits and a cognitive computing system which effectively models clinical thinking, including disease chronology, can automatically present problems for physicians to verify and validate. Physicians clearly value the ability to identify important problems, however, redundancy and non-problems are a challenge that should be resolved. Therefore, incorporating such a cognitive computing system, with enhancements to minimize redundancy and non-problems, into the workflow will be well received by physicians, and may improve patient care.

Summary points

What was known before this study?

- The structured and unstructured data (plain text clinical notes) of a longitudinal patient record contain valuable information about a patient's medical status and treatment, and NLP can be used successfully to extract various medical concepts, assertions, and relations about them using the UMLS® Metathesaurus® of biomedical concepts.
- While a patient's medical problem list can be at the core of successful management and treatment, maintaining a correct problem list remains a challenge, and therefore, physicians don't rely on the problem list in a patient record.
- A natural language processing method can identify a patient's

Table 4b
Physicians' free-text responses to what should be improved in the Watson generated problem list.

Theme	Comments (physician's anonymized id in the parenthesis)
Not well supported diagnosis	*Watson picks up a lot of text that states no evidence of something, but picks up that word and adds it to the diagnosis (4475)* *Watson documents problems if they are mentioned in the chart, BUT does not appear to require validating evidence to substantiate what someone wrote in a note. Anyone can write in the note, and whether a first day RN, a third year medical student, or a staff these lines of texts look like they are analyzed with equal weight. Some claims need substantiation. (3452)*
Non-active/redundant/ general problems	*If acute diagnosis/condition but no longer a diagnosis on subsequent visits, then Watson should remove from (active) problem list. (5413)* *Eliminate redundancy (maybe run a function that looks for similar problems (obesity, morbid obesity) and removes the least specific one prior to presenting the problem list to the user. (3452)*
Watson missed diagnoses	*Chart review revealed diagnosis of hypertension(and associated medications). This was neither on* (the existing problem) *list nor Watson list. So if medication list used by Watson, this should have been noted. (3807)*
Natural language processing errors	*Confused muscle response depressed with depression (4475)*

medical problems from a pre-specified list of 80 problems with improved sensitivity.

What did this study add to the body of knowledge?

- Physicians found the IBM Watson generated problem list more useful than an existing manually maintained EHR problem list.
- Physicians miss important problems when creating their own problem list, as the task of reviewing a patient record can be tedious and error prone.
- Physicians perceive the existing EHR problem list poorly because important problems can be missing.
- Cognitive computing systems can be a foundation for clinical decision support and have the potential to improve the quality of patient care.

Financial disclosures

This research was self-funded by the institutions involved, IBM Research and Cleveland Clinic, and thus funding was not provided by one to the other. The authors do not have any other financial interests to disclose. The participation in the study by physicians and residents was entirely voluntary—they were not compensated.

Acknowledgements

We thank the physicians and IT staff at Cleveland Clinic who guided definition of the requirements for this application and provided de-identified patient records under an IRB protocol for the study. We also acknowledge the groundbreaking work of the IBM Watson team colleagues, past and present, which made this research possible. We thank Tong-Haing Fin (IBM) for the software engineering support to create and maintain the web application used in the pilot study. Eric W. Brown's (IBM) executive support and Julie Tebo's (Cleveland Clinic) insightful suggestions benefited this study. We gratefully acknowledge the able project management support of Lauren Mitchell (IBM) and Charles “Chip” Steiner (Cleveland Clinic) in this effort.

References

[1] R. Wachter, The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age, McGraw-Hill, 2014.
[2] L.L. Weed, Medical records that guide and teach new england, J. Med. 278 (March (12)) (1968) 652–657.
[3] R. Pivovarov, N. Elhadad, Automated methods for the summarization of electronic health records, J. Am. Med. Inform. Assoc. 22 (5) (2015) 938–947.
[4] M. Devarakonda, D. Zhang, T. Ching-Huei, M. Bornea, Problem-Oriented patient record summary. an early report on a watson application, IEEE HealthCom, Natal Brazil, 2014.
[5] C. Holmes, The problem list beyond meaningful use, part I, J. AHIMA 81 (February (2)) (2011) 30–33.
[6] C. Holmes, The problem list beyond meaningful use, part 2, J. AHIMA 81 (March (3)) (2011) 32–35.
[7] T.D. Shanafelt, L.N. Dyrbye, C. Sinskye, O. Hasan, D. Satele, J. Sloan, C.P. West, Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction, Mayo Clin. Proc. 91 (March (7)) (2016) 836–848.
[8] D. Murphy, M. Ashley, E. Russo, D.F. Sittig, L. Wei, H. Singh, The burden of inbox notifications in commercial electronic health records, JAMA Int. Med. 176 (4) (2016) 559–560.
[9] T. Brown, When Hospital Paperwork Crowds Out Hospital Care, New York Times, 2015, p. SR11 (19 December).
[10] S. Meystre, P.J. Haug, Natural language processing to extract medical problems, J. Biomed. Inform. 39 (2006) 589–599.
[11] S. Meystre, P. Haug, Improving the sensitivity of the problem list in an intensive care unit by using natural language processing, AMIA Annual Symposium Proceedings, Washington, DC, 2006.
[12] S.M. Meystre, P.J. Haug, Randomized controlled trial of an automated problem list with improved sensitivity, Int. J. Med. Inf. 77 (2008) 602–612.
[13] M.V. Devarakonda, N. Mehta, Cognitive Computing for Electronic Medical Records, in Healthcare Information Management Systems, in: A.C. Weaver, J.M. Ball, R.G. Kim, M.J. Kiel (Eds.), 4th edition, Springer International, 2015.
[14] S. Velupillai, D. Mowery, B.R. South, M. Kvist, H. Dalianis, Recent advances in clinical natural language, IMIA Yearb. Med. Inform. 10 (1) (2015) 183–193.
[15] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, R.T. Mueller, Watson: beyond jeopardy, Artificial Intelligence vol. 200, (2013), pp. 93–105.
[16] Department of Health and Human Services, Medicare and Medicaid programs; electronic health record incentive program; final rule., July 2010. [Online]. Available: http://edocket.access.gpo.gov/2010/pdf/2010-17207.pdf (Accessed 17 February 2017).
[17] US National Library of Medicine, UMLS Reference Manual, National Library of Medicine (US), September 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK9675/ (Accessed 15 04 2014).
[18] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inf. Assoc. 17 (5) (2010) 507–513.
[19] US National Library of Medicine, The CORE Problem List Subset of SNOMED CT, 2014. [Online]. Available: http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html (Accessed 16 September 2014).
[20] M. Devarakonda C.-H. Tsou Automated Problem List Generation from Electronic Medical Records in IBM Watson, in Innovative Applications of Artificial Intelligence (AAAI), Austin, TX, USA 2015.
[21] Y. Freund, L. Mason, The alternating decision tree algorithm, Proc of the Sixteenth International Conference on Machine Learning, San Francisco, CA. USA, 1999.
[22] B. Dandala, M. Devarakonda, M. Bornea, C. Nielson, Scoring disease-Medication associations using advanced NLP, machine learning, and multiple content sources, Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), Osaka, Japan, 2016.
[23] HL7 (Health Level Seven International), HL7/ASTM Implementation Guide for CDA® R2 −Continuity of Care Document (CCD®) Release 1, [Online]. Available: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=6. (Accessed 16 March 2017).
[24] J.J. Liang, C.-H. Tsou, M.V. Devarakonda, Ground truth creation for complex clinical NLP tasks −An iterative vetting approach and lessons learned, Proceedings of AMIA Joint Summits on Translational Science, San Francisco, CA, USA, 2017.