RC25615 (WAT1607-040) July 26, 2016
Computer Science

IBM Research Report

Physicians Assessment of IBM Watson Generated Problem List

Murthy V. Devarakonda[1], Neil Mehta[2], Ching-Huei Tsou[1], Jennifer L. Liang[1], Amy S. Nowacki[2], John Eric Jelovsek[2]

[1]IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

[2]Cleveland Clinic

Research Division
Almaden – Austin – Beijing – Brazil – Cambridge – Dublin – Haifa – India – Kenya – Melbourne – T.J. Watson – Tokyo – Zurich

Physicians Assessment of IBM Watson Generated Problem List

Murthy V Devarakonda,[a] PhD, Neil Mehta,[b] MBBS, Ching-Huei Tsou,[a] PhD, Jennifer J Liang,[a] MD,
Amy S Nowacki,[b] PhD, John Eric Jelovsek,[b] MD MMEd
[a]IBM Research and [b]Cleveland Clinic

Abstract

Objective: An accurate, comprehensive and up-to-date problem list can help clinicians focus on providing patient-centered care. In this study, we report on physicians' assessment of IBM Watson generated problem lists and comparison with an existing manually curated problem list in an institution's EHR system.

Materials and Methods: Fifteen randomly selected, de-identified patient records from a large healthcare system were analyzed using Watson. Ten internal medicine physicians each reviewed five randomly selected patient records and created their own problem lists (P) for each patient record. Then, they evaluated the Watson generated problem lists (W), and rated the overall usefulness of P and W, as well as the existing EHR problem lists (E). The primary outcome was the physicians' usefulness ratings of the problem lists on a 10-point scale and their pairwise comparisons.

Results: Six out of the 10 invited physicians completed 27 assessments of P, W, and E, consisting of 732 Watson generated problems and 444 problems in the EHR system. As expected, physicians rated their own lists, P, best. However, they rated W higher than E. In 89% of the assessments, Watson identified at least one important problem that the physicians missed. The higher ratings of W relative to E were influenced by the number of problems missing from E.

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

Background and Significance

Despite the potential to improve healthcare, Electronic Health Records (EHRs[1]), have failed to significantly improve patient outcomes [1]. Physicians struggle to assimilate vast amounts of data, and continue to report workflow disruptions, decreased productivity and low satisfaction with using EHR systems [2]. A simple but a key function of any medical record is to present a comprehensive problem list that summarizes a patient's medical conditions [3]. The problem list offers many benefits, including helping practitioners provide holistic, customized care for a patient, and has potential use for quality improvement and research [4] [5]. While Weed's seminal paper on problem-oriented medical records [3] established the importance of the problem list in patient care, curating an accurate problem list has remained a challenge for many reasons. Some of the known reasons include different "attitudes" towards the problem list arising out of lack of clarity on policies [6] [7], the requirement of broad clinical expertise, and the imposition of significant demands on physicians' time. In fact, a recent report notes that electronic

[1] 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 all clinical notes, reports, medications ordered, procedures ordered, and demographic data; Patient record here always refers to longitudinal and complete patient data stored in an EHR system, although occasionally we prefix it with *longitudinal* for emphasis.

"paper work" in commercial EHR systems so overwhelming to physicians that it is affecting patient care [8] [9] and putting them at higher risk of professional burnout [2].

Existing EHR systems allow for manual creation and maintenance of such problem lists, but often these lists are inaccurate or incomplete, particularly when managed in large multi-provider health systems. There have been a few attempts to study automated problem list generation and its usefulness. These include efforts to define better coding systems to represent medical problems [10] and even more recent activity to define a new coding system based on a subset of SNOMED CT [11]. The only other system for automated problem list generation [12] [13] [14] can only identify a patient's medical problems from a pre-specified list of 80 problems.

Cognitive computing systems, such as IBM Watson [15], based on natural language processing (NLP), information retrieval, knowledge representation, and machine learning (ML) have the potential to improve the use of patient records by automatically generating unconstrained problem lists for clinician review [16]. Research in NLP, ML, and their applications to clinical data has advanced beyond merely extracting a few biomedical concepts from the clinical notes in patient records. We can now solve far harder problems in clinical informatics with this technology [17] [18]. Since winning the Jeopardy! championship, IBM Watson has been adapted to the medical domain [19], and even beyond this, an initiative was started at IBM to extend Watson to provide cognitive assistance to physicians in using longitudinal patient records.

IBM Watson generates a problem list from a longitudinal patient record by analyzing the free-text clinical notes and using the structured data in the patient record [20] (see the Appendix for an overview of the method). Unlike the previous work, IBM Watson can identify any of 6,166 problems in the version of SNOMED CT CORE subset (201508) we employed. We trained and tested the algorithm using a gold standard created by medical experts. The method achieves a high level of accuracy on the gold standard. Beyond the gold-standard-based analysis, it is, however, important to study physicians' perspective of the generated problem list and the value physicians attribute to it in patient care.

Objective

The primary objective of this study was to compare physicians' perceptions of the usefulness of automatically generated Watson problem lists with the pre-existing, manually curated problem lists in the EHR system. We hypothesized that clinicians would perceive the automatically generated problem lists as more useful than the manually curated problem lists. The secondary objective was to conduct additional exploratory analysis of the assessment data to identify factors influencing physicians' ratings and to determine Watson accuracy in terms of recall, precision, and F score.

Methods and Materials

Study Design

The experiment was conducted in a five-week time period in late 2015 at Cleveland Clinic. Institutional Review Board approval was obtained and a convenience sample of 10 internal medicine attending physicians and senior residents were recruited to participate in this study. Fifteen randomly selected, de-identified longitudinal patient records from the healthcare institution were also selected. In order 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. Patient records were extracted from the commercial EHR system at the healthcare institution, and were de-identified before being forwarded to IBM Watson for automatically generating the problem lists. The Watson generated problem

list for each patient was made available to the physicians via a Web application, accessed in a standard Web browser. Physicians were given a key to map the Watson ID for a patient record to the patient record number (e.g. MRN) that can be used to access the patient record in the healthcare institution's EHR

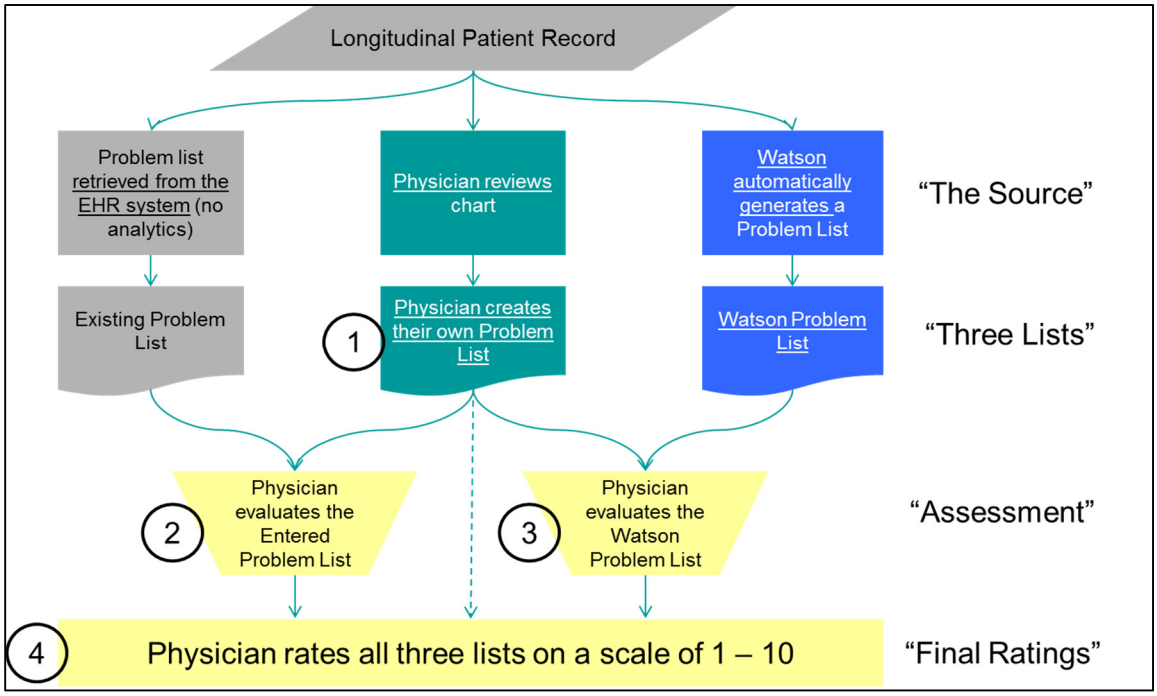

Figure 1. The assessment for a patient record consisted of a series of steps each physician carried out, including creating their own problem list, evaluating the existing problem list in the EHR, evaluating the Watson generated problem list, and finally rating all the three problems lists on a 10-point response scale.

system. They were each randomly assigned to review 5 of the 15 patient medical records in the EHR system like they would prior to a comprehensive health assessment of a patient new to them and were asked to create a problem list for each patient record. They were then 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 response scale of 1 to 10 on their usefulness in patient care.

Assessment Steps

The assessment consisted of a series of steps carried out by physicians (Figure 1) using the Web application that was developed for this experiment and a standard Web browser.

Steps 1 and 2 (Figure 2a)

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 was available to them as a reference source for creating the problem list. The physicians entered each problem in the Web application (Figure 2a), and also indicated whether the problem was present on the existing problem list (E) in the patient record. As a result, physicians provided an assessment of problems in E while creating their own list (P).

Step 3 (Figure 2b)

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. At this stage, they continued to have access to the full patient record from 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 into the Web application in step 1.

(a)

Problems you identified (in your preferred clinical terminology)	Is this problem present in the EHR problem list?
Hypertension	(•) Yes ( ) No
Diabetes	( ) Yes (•) No
VSD	(•) Yes ( ) No
	( ) Yes ( ) No
	( ) Yes ( ) No

+ Add additional row - Delete selected problem

(b)

	Watson generated problem list (Click on the problem name to see the corresponding clinical notes)	No Assessment	Correct	Acceptable (too general)	Acceptable (too specific)	Acceptable (but redundant)	Incorrect (too general)	Incorrect (transient and/or resolved)	Incorrect (not a problem)	Is this in your problem list?	How clinically important is this problem for managing this patient?
1	sleep apnea	( )	(•)	( )	( )	( )	( )	( )	( )	( ) Yes ( ) No	-- Select --
2	ventricular septal defect	(•)	( )	( )	( )	( )	( )	( )	( )	( ) Yes ( ) No	-- Select --
3	tetralogy of fallot	(•)	( )	( )	( )	( )	( )	( )	( )	( ) Yes ( ) No	-- Select --
4	carpal tunnel syndrome	(•)	( )	( )	( )	( )	( )	( )	( )	( ) Yes ( ) No	-- Select --

(c)

Now that you've reviewed the Watson generated problem list, please rate the following lists for managing the care of this patient.

	Extremely useful 10	9	8	7	6	5	4	3	2	Not at all useful 1
1. The problem list you created in the first screen	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
2 Existing problem list in the EHR system	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
3. Watson generated problem list	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )

Figure 2. 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 lis 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.

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.

Step 4 (Figure 2c)

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 to 10, 1 being least useful and 10 being most useful.

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 Wilcoxon signed-rank test [21] because of the non-normality of the ratings distributions.

In addition, because we asked physicians to indicate if each Watson generated problem was on their problem list, we determined if physicians missed any problems that Watson found and their clinical importance as perceived by the physician.

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 providing a balanced measure of recall and precision, and F2 providing a higher recall-weighted measure. Specificity, also known as true negative rate, is not useful in tasks 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. 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:

$$R = \frac{T_P}{T_P + F_N} \qquad P = \frac{T_P}{T_P + F_P}$$

$$\mathrm{F1} = \frac{2PR}{(P + R)} \qquad \mathrm{F2} = \frac{5PR}{(4P + R)}$$

The gold standard needed for the accuracy calculations was created using the following process:

- For each patient record, we 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.
- We removed any duplicates added to the list as a result of the above two steps (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, therefore, 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 as was used in the previous study [20].

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

Factors Influencing Physicians' Ratings

We further analyzed the assessment data to identify factors that may have influenced the individual scale ratings of P, W, and E as well as the difference between W and E ratings. To this end, we determined the Pearson correlation coefficient between the ratings and the data we collected through the Web application (Figures 2a through 2c). Only the data that was directly measured and their normalized values (with respect to certain relevant measures such as the number of problems in a list, as will be discussed later) were considered, which were:

- Number of problems physicians missed but Watson found (and by each "importance" category)
- Number of correct (true positives) and incorrect (false positives) problems in E, as determined by physicians
- Number of problems missing (false negatives) from E, relative to P
- Number of correct (true positives) and incorrect (false positives) problems in W, as determined by physicians, and incorrect problems of each type (i.e. "too general", "transient/resolved", or "non-problems")

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 comments:

1. *Please identify one thing that you like about the Watson generated problem list*
2. *Please suggest one improvement for the Watson generated problem list*

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

Results

Among 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 completing all the required steps described above for a patient record. Twelve records were assessed by two participants and three records were assessed by only one participant each. The experiment resulted in evaluations of 732 Watson generated problems and 444 problems in the existing EHR patient records.

Hypothesis Test Results

Results of the pairwise comparison of the scale ratings using the Wilcoxon signed-rank test are shown in Figure 3 and Figure 4. As expected, physicians rated their own list (P) significantly higher than the Watson generated problem list (W) and the existing manually entered 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 entered problem list, the Watson generated problem list was rated higher in 10 cases, the same in two cases, and lower in three cases.

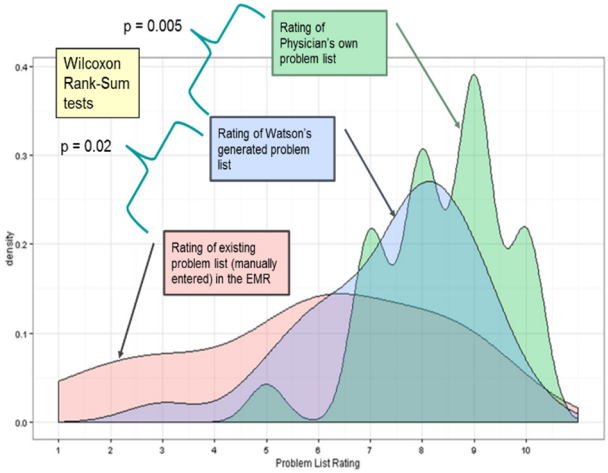

Figure 3. Pairwise comparison of physicians' problem list ratings shown as density functions.

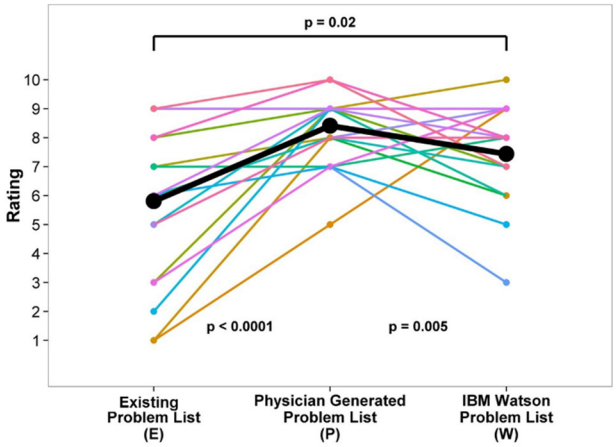

Figure 4. Pairwise comparison of physicians' problem lists ratings shown as a stick diagram and with mean values

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 1).

Table 1. Problems missed in the physicians' problems 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

Watson Accuracy

Table 2a shows the confusion matrix for the Watson problem list accuracy analysis and Table 2b shows the accuracy metrics -- recall, precision, and F scores. The false positives are larger than the false negatives by nearly 3 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 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.

Table 2a. 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	50
	False	139	-----

Table 2b. Watson problem list accuracy analysis from this assessment; Results from the previous study [20] 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

Factors Influencing Physicians' Ratings

Table 3 shows factors correlated with the scale ratings, and all correlations shown are statistically significant at $p<0.01$. The following list summarizes the highest correlation factors for each of the scale ratings of interest:

- Physician's own list ratings (P):
  - Has the highest negative correlation (-0.63) with the number of "very important" problems missed in P relative to W, however, when this factor is normalized with respect to the number of problems in P, the correlation weakens to -0.49
- Watson list rating (W):
  - Has the strongest negative correlation (-0.65) with Watson false positives due to "transient/resolved" problems (relative to P), and when normalized with respect to the number of problems in W, a similar correlation is observed with the total Watson false positives (-0.65)
- Existing EHR list rating (E):
  - Has the strongest negative correlation with the false negatives in E, relative to P, whether the raw scores are considered (-0.77) or the false negatives are normalized with respect to the number of problems in P (-0.87)
- The difference between Watson list and existing EHR list ratings (W – E):
  - Has the strongest positive correlation with the false negatives in E, relative to P, whether the raw counts are considered (0.85) or the normalized false negatives are considered (0.75)

Therefore, the significant results here are the strong correlations between the Watson rating and the Watson false positives and between the Watson and existing EHR ratings difference and the false negatives in E.

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

Table 3. Factors correlated with the problem lists ratings

Factor	Correlation with P	
	Non-normalized	Normalized
Physician missed "very important" problems	-0.63	-0.49
Physician missed "very important" or "important" problems	-0.51	-0.47
Factor	**Correlation with W**	
	Non-normalized	Normalized
Watson false positives ("too general")	-0.45	-0.45
Watson false positives ("transient or resolved")	-0.65	-0.64
Watson false positives (all types)	-0.58	-0.65
Factor	**Correlation with E**	
	Non-normalized	Normalized
Existing problem list false negatives	-0.77	-0.87
Factor	**Correlation with (W – E)**	
	Non-normalized	Normalized
Existing problem list false negatives	0.85	0.75

responses to the question, *please suggest one improvement for the Watson generated problem list.* The following seven common themes were observed in the comments:

1. Watson found diagnoses that physician had missed
2. Watson was very complete/thorough
3. Watson supported clinical reasoning
4. Watson listed a diagnosis that was not well supported
5. Watson list was broad and included redundant and non-active problems
6. Watson missed diagnoses
7. 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.

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 was able to generate a more complete and useful problem list.

Table 4a. Physician's free-text response 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)*

Table 4b. Physician's free-text response 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)*

The fact that physicians missed several important problems is an indication that the problems that were identified by Watson may be of potential importance. Necessary facts are not well organized in a commercial EHR system for easy access, 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 free 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 which they might otherwise overlook.

There is an indication that the number of incorrect problems, especially the transient or resolved problems, produced by Watson has 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 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 seem to react negatively to this increased noise level and it is a topic for further investigation.

The existing EHR problem list rating is negatively correlated with the number of true problems missing from it relative to the physicians' own list, in other words, poor recall is less useful from the physicians' perspective. This may explain why most physicians are reluctant to rely on the EHR problem list [7]. It is important to note that while Watson's rating is negatively influenced by lower precision (even at higher recall), the EHR problem list rating is negatively influenced by its poor recall.

The difference between the Watson and EHR problem list ratings is highly, positively correlated with the number of true problems missing from the EHR problem list (relative to the physicians' own list). This result, at least in part, explains why physicians rated the Watson problem list more useful than the EHR problem list – the Watson problem list includes more true problems than the EHR problem list.

It is instructive to explore how the Watson accuracy measured here compares with the results based on the gold standard developed from the previously reported method [20], where the gold standard was developed involving multiple experts, subsequent 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, in spite of 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, non-problems and so on in the Watson problem list are also reflected in the negative correlation between Watson's false positives and Watson's 10-point scale score, and in the relatively lower precision (compared to the recall) of the Watson problem list in the accuracy analysis.

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 in spite of 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 can automatically present problems for physicians to verify and validate. Physicians clearly value the ability to identify important problems. Therefore, incorporating such a cognitive computing system into the workflow can improve the accuracy of problem lists, 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 as a consequence, physicians don't rely on the problem list in a patient record.
- A natural language processing method can identify a patient's 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 entered EHR problem list.
- Physicians miss important problems when creating their own list, as the task of reviewing a patient record can be tedious and error prone.
- Physicians perceive the existing EHR problem list poorly because of missing important problems.
- Cognitive computing systems can be a foundation for clinical decision support and have the potential to improve the quality of patient care.

Acknowledgement

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 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, McGraw-Hill, 2014.

[2] T. D. Shanafelt, L. N. Dyrbye, C. Sinskye, O. Hasan, D. Satele, J. Sloan and C. P. West, "Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction," *Mayo Clinic Proceedings,* vol. 91, no. 7, pp. 836-848, 2016.

[3] L. L. Weed, "Medical Records That Guide and Teach," *New England Journel of Medicine,* pp. 652-657, March 1968.

[4] C. Holmes, "The Problem List Beyond Meaningful Use, Part I," *Journal of American Health Information Management Association,* vol. 81, no. 2, pp. 30-33, 2011 Feb.

[5] C. Holmes, "The Problem List beyond Meaningful Use, Part 2," *Journal of American Health Information Management Association,* vol. 81, no. 3, pp. 32-35, 2011 Mar.

[6] H. Casey, M. Brown, D. S. Hilaire and A. Wright, "Healthcare provider attitudes towards the problem list in an electronic health record: a mixed-methods qualitative study," *BMC Medical Informatics and Decision Making,* vol. 12, no. 127, 2012.

[7] A. Wright, F. L. Maloney and J. C. Feblowitz, "Clinician attitudes toward and use of electronic," *BMC Medical Informatics and Decision Making,* vol. 11, no. 36, 2011.

[8] D. Murphy, M. Ashley, E. Russo, D. F. Sittig, L. Wei and H. Singh, "The Burden of Inbox Notifications in Commercial Electronic Health Records," *JAMA Internal Medicine,* vol. 176, no. 4, pp. 559-560, 2016.

[9] T. Brown, "When hospital paperwork crowds out hospital care," *New York Times,* p. SR11, 19 December 2015.

[10] J. R. Campbell and T. H. Payne, "A Comparison of Four Schemes for Codification of Problem Lists," San Francisco, 1994.

[11] 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].

[12] S. Meystre and P. Haug, "Improving the Sensitivity of the Problem List in an Intensive Care Unit by Using Natural Language Processing," in *AMIA Annual Symposium Proceedings*, Washington, DC, 2006.

[13] S. Meystre and P. J. Haug, "Natural language processing to extract medical problems," *Journal of Biomedical Informatics,* vol. 39, pp. 589-599, 2006.

[14] S. M. Meystre and P. J. Haug, "Randomized controlled trial of an automated problem list with improved sensitivity," *International Journal of Medical Informatics,* vol. 77, pp. 602-612, 2008.

[15] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer and C. Welty, "Building Watson: An overview of the DeepQA project," *AI Magazine,* vol. 31, no. 3, pp. 59-79, 2010.

[16] M. V. Devarakonda and N. Mehta, "Cognitive Computing for Electronic Medical Records," in *Healthcare Information Management Systems, 4th Edition*, A. C. Weaver, J. M. Ball, R. G. Kim and M. J. Kiel, Eds., Springer International, 2015.

[17] R. Pivovarov and N. Elhadad, "Automated Methods for the Summarization of Electronic Health Records," *Journal of the American Medical Informatics Association,* vol. 22, no. 5, pp. 938-947, 2015.

[18] S. Velupillai, D. Mowery, B. R. South, M. Kvist and H. Dalianis, "Recent Advances in Clinical Natural Language," *IMIA Yearbook of Medical Informatics,* vol. 10, no. 1, pp. 183-193, 2015.

[19] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek and R. T. Mueller, "Watson: Beyond Jeopardy!," *Artificial Intelligence,* pp. 93-105, 2013.

[20] M. Devarakonda and C.-H. Tsou, "Automated Problem List Generation from Electronic Medical Records in IBM Watson," Autin, TX, 2015.

[21] Wikipedia, "Wilcoxon signed-rank test," Wikipedia, the free encyclopedia, [Online]. Available: https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test. [Accessed 5 June 2016].

Appendix: An Overview of the IBM Watson Problem List Generation Method

IBM Watson problem list generation is a binary classification method that uses supervised machine learning. The goal is to identify a patient's diseases/syndromes, major unresolved symptoms, and significant procedures that require medical care and management. It starts with an automatically identified subset of UMLS concepts from clinical notes of a patient record as potential candidate problems, and proceeds to classify them as true problems or non-problems. The UMLS subset includes any mentioned concept in the clinical notes belonging to the UMLS *Disorders* semantic group and a few belonging to the *Procedures*, *Physiology*, and *Living Beings* semantic groups, subject to there being a mapping to a concept in the CORE subset of SNOMED CT. The machine learning model is trained on a gold standard manually created by medical experts using de-identified patient records from Cleveland Clinic.

Candidate Problems Identification

In Watson, UMLS medical concepts are identified in all parts of a patient record – both in the plain text clinical notes and in the remaining semi-structured clinical data – resulting in terms (words and phrases) in a patient record being assigned one or more Concept Unique Identifiers (CUIs) from the UMLS Metathesaurus. Using the context around a term is often necessary to obtain a CUI that represents the concept more accurately, because the CUI and the term spaces are large and the mapping is many-to-many. In addition to the standard NLP and UMLS lookup, we use additional contextual and sentence structural information to obtain a better mapping. A numerical score indicates how confident Watson is that a CUI represents the original term, and the score is used as a feature in problem list generation. As a result of CUI mapping, the terms are also categorized into semantic groups, e.g. as Disorders, Chemicals & Drugs, Procedures, etc. Each of these groups is further subcategorized, for example, Disorders are sub-grouped as Diseases or Syndromes, Signs or Symptoms, Findings, and others. Once one or more CUIs for a concept are identified, the CUIs are then mapped to a SNOMED CT CORE concept. If there is no exact match, we climb the UMLS hierarchy until the closest parent or a sibling concept, that is also a SNOMED CT CORE concept, is reached. This set of SNOMED CT CORE concepts identified from the clinical notes forms the candidate problem list.

For a typical patient record, usually a few hundred candidate problems are identified. When compared to the final list, the problems generated in the first step would have high recall (~90%) but poor precision (~10%). The subsequent steps attempt to improve precision of the problem list without substantial loss of recall.

Features and Feature Extraction

In the second step, the method produces feature values for use in the machine learning model. We manually engineered a large number of features which can be broadly categorized as lexical, clinical, frequency-based, structural, and temporal. At this time, we use 260 features, some of which are expanded further from multi-valued categorical features to binary valued one-hot-vectors depending on the classification method (e.g. support vector machines). However, our present preferred classification method, Alternating Decision Trees, selects the most informative features among these during the training step, and currently consists of 31 features. The feature categories are described below using sample features. The key features of the current ADT model are listed in Table A-1.

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

Feature	Description
Freq Prob	Frequent problems (top 25% of most frequently diagnosed problems).
Diagnosed	Has ever been diagnosed (structured data)
S_PMH	Problem is found in the past medical history section
Med Score	Problem is related to any active medication
TF-IDF	Term Frequency and Inverse document frequency
TF (1st section)	TF; problem appears in the 1st section of the EMR
TF (A&P)	TF; problem appears in the "assessment and plan" section of the EMR
PCA	Probability of Concept for a given term (covered text)
Freq Prob (CORE)	SNOMED CORE usage: the average usage percentage among all institutions
1st Date	Date (normalized) the problem is first occurred
TF	Term-frequency
TF (Recent)	TF; prob. appears in the recent (last 3 mo.) notes
TF (Recent, RoS)	TF; problem appears in the "Review of System" section in the recent (last 3 months) notes

Lexical Features

Standard TF (term frequency) and TF-IDF (term frequency multiplied by the inverse document frequency) are examples of lexical features we use. TF-IDF reflects how important a term is to a document in a corpus. In our case, a term is a candidate problem. Depending on the goal, a document can be a clinical note or an entire patient record. When generating the problem list for a patient, an entire patient record is the document and the collection of all patient records is the corpus. When deciding which is a relevant note to a selected problem, the clinical note becomes the document and a patient record becomes the corpus.

Unlike a normal text document, a patient record is a longitudinal record and therefore, more recent notes are likely to better represent the patient's medical problems. Also, each note in the patient record has implicit sections, and so a concept (e.g. hypertension) appearing in different sections (e.g. family history vs. assessment and plan) may have different implications. Because of this, in addition to calculating TF at the patient record level, TF is also calculated for each note section, assertion type (e.g. negation, family history, hypothetical), and for different time periods (e.g. last 3 months, 6 months, and one year).

Clinical Features

The terms in the patient record's semi-structured data are also mapped to UMLS concepts so that we can use the UMLS relations on these terms. Medications turned out to be one of the most important features, whereas the lab tests, results, and procedure orders seemed to be less useful. The first reason is that the medication names are relatively standardized and UMLS concepts can be reliably found for them. But, labs and procedures are often specified in institution-specific abbreviations instead of standardized terms or codes, such as CPT and LOINC, and are therefore harder to accurately map to UMLS concepts. Second, medications are prescribed to treat problems, while lab tests and procedures are often ordered to explore diagnosis of a problem and extensive domain knowledge is needed to interpret their results. The relation between a medical problem and a medication is obtained from an ensemble of techniques including

distributional semantics, UMLS relationships, and data mining of structural (coded) data from millions of patient records (details beyond the scope of this Appendix).

Frequency Features

Problem frequency in the general population can be thought of as the prior probability that the patient may have it. Two sources of the frequency are used as features in our model. The first is the SNOMED CT CORE usage, which represents the frequency in a broad population. The second is the problem frequency in the diagnosed problems (as ICD-9 codes) in our collection of patient records (about 1,000).

Structural Features

A disorder such as "diabetes mellitus" appearing in the Assessment and Plan part of a physician's progress note is a much stronger indicator that it is indeed a medical problem for the patient, than the same concept detected in the family history part of a nursing note. Therefore, the section in which a disorder is mentioned in a clinical note and the note type are two useful features. Since clinical notes are unstructured plain text, IBM Watson detects the logical sections of a note with a learned SVM classifier combining regular expressions, heuristic rules, and n-grams as features. Note type is an optional metadata and is often missing, so the note type is determined using a supervised maximum entropy classifier that is based on several medical and lexical features from the note text and (available) metadata features. These structural features are combined with term-frequencies to weight each occurrence, as explained earlier in the lexical features section.

Temporal Features

The span of a patient record varies from a single day to several decades. Most temporal features in our experiments are normalized to prevent bias towards longer patient records, but the absolute value is also used to define certain features, e.g. note recency, where the recency is defined as the number of days from the latest patient contact.

Temporal data is used in three ways. First, it is used as features directly. Temporal features considered include the first and last mention of a problem, the duration of a problem, and other statistics that capture the characteristics of distribution of the occurrences. Second, it is used to align semi-structured data and structured data, e.g. a medication prescribed before a problem is mentioned in a note is not considered as evidence to the problem. Third, temporal data is used to divide notes into bins on the timeline so that frequency can be counted by intervals, e.g. TF in recent notes vs. TF in earlier notes.

Model

We used the Alternating Decision Tree (ADT) technique for its accuracy and clarity of the decision process. We formulated the problem list generation as a binary classification problem, i.e., for each candidate problem in a patient record, the task is to classify it as a problem or a non-problem. We initially used an SVM model with polynomial kernel, but soon favored the more human interpretable model. As the gold standard is expensive to develop and the training data is limited, knowledge coming from the domain experts and error analyses becomes critical to success – and both benefit from models that output a human understandable decision process. The decision tree and the association rules-based classifiers generate models close to the way medical experts think, at the cost of usually lower accuracy. We observed performance similar to our earlier SVM model by using ADT, which outputs an option tree but has its root in boosting. The basic implementation of ADT uses a decision stump as the base learner and adaptive boosting to grow the tree iteratively. During a boosting iteration, ADT adds a splitter node and

corresponding prediction nodes to extend one of the existing paths in the tree. The scores associated with the prediction nodes are obtained from the rules.

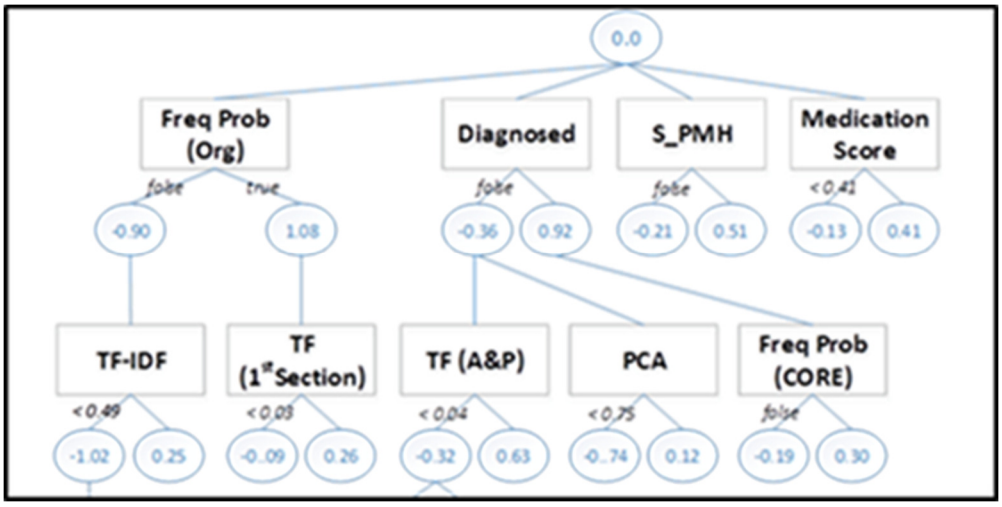

Figure A-1. The first two levels of the Watson problem list ADT model.

Model parameters are selected using 10-fold cross validation. The number of iterations of ADT is set to 40 (from the ROC and the Recall-Precision graphs), and the score threshold is set to 0.5, to maximize the training F2-measure. A subset (some branches are omitted after the first two levels) of the tree generated by our model is shown in Figure A-1. The top level features in Figure A-1 are all intuitive – but it is important to understand that they are not necessarily the most important features to determine whether a candidate problem is, in fact, a patient's active problem – they simply work better for the easy instances. Some less intuitive features also shed some light on how clinical notes are written. For example, it is a positive indicator if a problem appears in the first section of a clinical note, regardless of what the section is. This is because many clinical notes start by stating the patient's active concerns. Another example is the first mention date because a patient's past medical history is often carefully documented in his/her first visit to the hospital.

We note that (as reported in our previously published article) the number of candidate problems generated per patient record, across 399 patient records used in model training and testing, exhibit a nearly normal distribution, with an average of 135 candidate problems and a standard deviation of 33. The machine learning model reduces these candidate problems to an average of nine final problems, a reduction by over 93%.

Gold Standard for Training and Testing the Model

As there is no publicly available gold standard for problem lists, we developed a gold standard of our own, which is used for training and testing the Watson model. The process involved two fourth-year medical students reviewing each de-identified patient record and each student independently creating a problem list for the patient record. Then, an MD reviewed and adjudicated any differences between the students' problem lists. Inter-annotator agreement and further analysis of Watson generated problem lists for the patient records showed that even these lists can be improved further. So, for each patient record in the gold standard, the Watson generated problem list is compared with the adjudicated problem list, and any differences are further reviewed by the students and the MD. After this last step of vetting, the adjudicated problem lists are considered as the final gold standard. The gold standard problem lists are coded in the SNOMED CT CORE subset.