Ground Truth Creation for Complex Clinical NLP Tasks
– an Iterative Vetting Approach and Lessons Learned

Jennifer J. Liang, MD, Ching-Huei Tsou, PhD, Murthy V. Devarakonda, PhD
IBM Research, Yorktown Heights, NY, USA

Abstract
Natural language processing (NLP) holds the promise of effectively analyzing patient record data to reduce cognitive load on physicians and clinicians in patient care, clinical research, and hospital operations management. A critical need in developing such methods is the “ground truth” dataset needed for training and testing the algorithms. Beyond localizable, relatively simple tasks, ground truth creation is a significant challenge because medical experts, just as physicians in patient care, have to assimilate vast amounts of data in EHR systems. To mitigate potential inaccuracies of the cognitive challenges, we present an iterative vetting approach for creating the ground truth for complex NLP tasks. In this paper, we present the methodology, and report on its use for an automated problem list generation task, its effect on the ground truth quality and system accuracy, and lessons learned from the effort.

Introduction
Despite the potential to improve healthcare, EHR systems have failed to significantly improve patient outcomes. Physicians struggle to assimilate vast amounts of data, and continue to report workflow disruptions, decreased productivity and low satisfaction with using EHR systems. Natural language processing (NLP) holds the promise for effectively analyzing patient record data to help physicians and clinicians in patient care, clinical research, and hospital operations management. A critical part of developing an automated system using NLP and machine learning is a “ground truth” or “gold standard” dataset, which is needed to train and test the supervised machine learning technique commonly used in the methods (even methods that use semi-supervised and unsupervised techniques require a good test dataset). The process and methodology for creating reliable ground truth, especially for inducing higher-level, complex semantics beyond simple concept recognition, has not been studied.

In this paper, we present an iterative vetting approach for ground truth development, and report on the experience of using the approach to train and test an automated problem list generation method which uses NLP and supervised machine learning for distinguishing problems from non-problems. While Weed’s seminal paper on problem-oriented medical records established the importance of the problem list in patient care, curating an accurate problem list has remained a challenge for many reasons. Therefore, the IBM Watson team has set out to develop an automated method of problem list generation. The focus of the paper is the ground truth generation methodology and its impact on the accuracy of the automated method.

Specifically, the ground truth should facilitate training the machine learning model of the automated method, testing (measurement) of method accuracy using the trained model, and determination of the “headroom” (the largest achievable gap) for improvement. These steps are repeated using the ground truth until a desired level of accuracy is reached. For automated problem list generation, the ground truth is a list of true problems for each patient record. There are multiple challenges in creating accurate and complete problem lists for the ground truth:

1. Problems must be coded to a standard vocabulary that is used in the automated system.
2. The same problem may be referred to in a patient record using various terminologies, and may map to multiple related concepts in the standard vocabulary.
3. Human errors (e.g. coding errors, missed problems) are common in any manual process, especially when working with large patient records involving hundreds of clinical documents.

Prior studies evaluating automated problem list systems have used either ICD-9 codes associated with the patient record as ground truth, or created the ground truth from a review of a subset of all clinical documents in the patient record for a given set of problems. Both of these methods create incomplete ground truth; the former method assumes all problems are coded (which is not always the case), and the latter method does not cover all problems a patient may have and is prone to human error resulting in missed problems.

Iterative vetting is a classic approach in computer science and mathematics. In fact, one of the earliest known examples appears in Rhind Mathematical Papyrus, which was assumed to be written around 1700 BC. Iterative vetting involves using the output of one stage to determine the next stage, incrementally reaching a final objective. Here, we present our iterative vetting approach, show how it was used for problem list ground truth generation, and
quantify the improvements in the ground truth and system accuracy through multiple iterations. We also summarize lessons learned from our experiences and discuss broader perspectives beyond this current example.

**Methods**

We acquired de-identified patient records through a research collaboration agreement with the Cleveland Clinic. From this dataset, 399 random adult patients were selected for ground truth problem list generation. Over 17 fourth year medical students served as medical annotators and contributed to the problem list creation, mapping, and vetting process described here. They were hired on as supplemental employees and were compensated for their work. The available time and cost of experienced physicians are prohibitive for this task but we have an MD supervising and guiding the students’ work.

**Problem List Definition**

The problem list is formally defined by the Center for Medicare and Medicaid Services (CMS) for the federal meaningful use program as “a list of current and active diagnoses as well as past diagnoses relevant to the current care of the patient”. As this description allows for much subjective interpretation on what is considered “current”, “active” or “relevant”, we revised this definition to be more concrete and specific by (1) placing the problem list in the context of a comprehensive health assessment, and (2) providing examples of types of problems that should be included on the problem list, such as chronic diseases (e.g. hypertension, diabetes mellitus), past cardiovascular events (e.g. history of myocardial infarction, pulmonary embolism), and history of drug or alcohol abuse, among others.

**Initial Ground Truth Generation**

Following the above definition, two annotators independently reviewed each patient record. Annotators were asked to individually, (1) review all clinical notes in chronological order, (2) identify problems that should be on the problem list for a comprehensive health assessment, (3) map these problems to SNOMED CT CORE concepts in a one-to-many mapping, and (4) rank each CORE concept based on how closely it represents the problem. Then, they were asked to jointly adjudicate their individual problem lists and CORE mappings to produce an adjudicated problem list, which was then reviewed by one of the authors, an MD (Liang). The inter-annotator agreement is reported in the “Results” section.

Clinical notes were reviewed chronologically to help build a coherent story of the patient’s health and medical care as the backdrop of a comprehensive health assessment. Since a CORE concept may not adequately represent a specific problem, we proposed a ranking for each code mapped that indicates how well the code reflects the annotator’s concept of the problem in the patient record: rank 1 indicates an exact semantic match, rank 2 indicates that the concept is not an exact match but is an acceptable alternative, and rank 3 indicates a concept that is not what one would want to see in a problem list but still provides useful information on the problem. Rank 1 and rank 2 CORE concepts were considered part of the ground truth, while rank 3 was not considered part of the ground truth and was retained mainly for future error analysis. If no CORE concept exists which adequately represents an identified problem, the problem remained in the ground truth as an unmatched problem. Table 1 shows an example of a ground truth problem list with mapping.

**Table 1.** Ground truth problem list with mapping to SNOMED CT CORE for a single patient record. PID is the patient record identifier; GT Problem Name is a description of the problem in the words of the human annotator; SNOMED_CID and SNOMED_FSN are the mapped SNOMED CT code and corresponding fully specified name.

| PID | GT | GT Problem Name | SNOMED_CID | SNOMED_FSN | Rank |
|-----|----|-----------------|------------|------------|------|
| 3811 | 1 | Ulcerative colitis | 64766004 | Ulcerative colitis (disorder) | 1 |
|     |    | Ulcerative colitis | 444548001 | Ulcerative pancolitis (disorder) | 1 |
|     |    | Ulcerative colitis | 441971007 | Chronic left-sided ulcerative colitis (disorder) | 1 |
|     |    | Ulcerative colitis | 445243001 | Left sided ulcerative colitis (disorder) | 1 |
|     |    | Ulcerative colitis | 24526004 | Inflammatory bowel disease (disorder) | 2 |
|     |    | Ulcerative colitis | 64226004 | Colitis (disorder) | 3 |
|     | 2 | Hypertension | 1201005 | Benign essential hypertension (disorder) | 1 |
|     |    | Hypertension | 10725009 | Benign hypertension (disorder) | 1 |
|     |    | Hypertension | 38341003 | Hypertensive disorder, systemic arterial (disorder) | 1 |
|     |    | Hypertension | 59621000 | Essential hypertension (disorder) | 1 |

Since we were working with de-identified patient records outside of a commercial EHR system, clinical notes were made available to the annotators as HTML files in a web browser (Figure 1), and spreadsheets were provided to them to record problems identified and their corresponding mapped CORE codes and ranks.
Using the adjudicated problem lists, we trained our automated problem list generation system developed using machine learning and NLP techniques and measured standard precision, recall, and F1 score as accuracy metrics.

Iterative Vetting Process

Since problem list creation using a manual review is mentally taxing and therefore prone to errors, we further iterate on the initial ground truth problem list by additional vetting passes. We used the system, trained on the initial adjudicated ground truth, to produce a highly recall-oriented (F3 score optimized) problem list. We compared these system generated highly recall-oriented lists, which are in SNOMED CT, with the ground truth problem lists, to identify any problems that appear in the system generated lists but not in the ground truth (System-only), and any problems that appear in the ground truth but were not found in the system generated lists (GT-only). Both sets of problems, System-only and GT-only, were given to two medical annotators (possibly different ones who created the initial ground truth) for independent vetting.

For the System-only set, two annotators were asked to review each problem and determine if the problem is:

1. false positive (meaning the problem should not be in the ground truth),
2. true positive, a. an existing problem (meaning the problem is already in the ground truth but the code found by the system was missed during the mapping part of the initial ground truth creation), or b. a new problem (meaning the problem should be added to the current ground truth, indicating a missed problem during the initial ground truth creation).

The annotators then adjudicated their vetted problems to agree on a final list of codes for existing problems and new problems to be added to the original ground truth. We then retrained the system using the revised ground truth and measured precision and recall for the retrained system.

For the GT-only set, two annotators were asked to review each problem and provide qualitative feedback on why the system may have missed the problem or, if the problem should be considered for removal from the ground truth, why the problem should be removed. This feedback, collected in free text form, was used by researchers to gain further insight on the information considered and the mental process used by medical experts in creating a problem list.

Figure 1. HTML version of patient record for review of clinical notes.

Figure 2. Search Application used to help locate relevant information pertaining to the problem of interest in a specific patient record. The tool performs a multi-dimensional search on the query term (including concept matching as well as associated medications, tests, treatments and procedures) and presents the results in a tabbed interface. The screenshot above shows matched associated Tests for the query “polyp of colon” issued on a specific patient record.
Figure 3. Evidence Viewer used for viewing evidence behind problems found by the system for a specific patient record. In the screenshot above, the evidence for problem “Depressive disorder” is displayed, including dates when problem was mentioned, clinical notes where problem was mentioned, and associated structured medications.

Annotators reviewed problems primarily using the HTML files from before. Additional tools were also provided to allow annotators to search the patient record for relevant information pertaining to the problem being investigated (Figure 2), and to view any evidence surfaced by the system for each problem (Figure 3).

Theoretically the vetting process can go on indefinitely with many iterations as the system model changes and improves. For this paper, we report on changes to the ground truth problem list after each of two vetting passes (Iteration 1 and Iteration 2), and the results of the system after training on the initial ground truth problem list (Initial GT), the problem list after one vetting pass (Iteration 1), and the problem list after a second vetting pass (Iteration 2). Figure 4 summarizes the entire ground truth generation process.

Figure 4. Problem List Ground Truth Generation Process.
Results

Why can’t Entered Problem List(s) be used as the Ground Truth?

Entered problem lists within patient records are often inaccurate\(^{(13)}\), and therefore not suitable for use as ground truth. In our corpus, each patient record has three coded problem lists: a manually maintained problem list shared by all providers, a diagnosis list used to record diagnosis codes applicable for each clinical encounter, and a medical history list that captures the patient’s significant past medical history. Compared to the initial adjudicated ground truth (Initial GT), all of these entered lists, either individually or combined, have poor accuracy, as shown in Figure 5.

![Figure 5](image)

Figure 5. Accuracy of Entered Lists (problem list, diagnosis list, medical history list, and the union of all 3 lists) as compared to the initial adjudicated ground truth (Initial GT). The entered lists are coded in ICD-9, and the ground truth problems are coded in SNOMED CT CORE. The matching is done automatically using rule-based two-way mapping based on two UMLS resources\(^{(14, 15)}\).

Characteristics of Patient Records Used in the Ground Truth Creation

This study uses 399 randomly selected longitudinal patient records as the corpus, which contains a total of 33,815 clinical notes. The size of a patient record varies from a single note in one visit to more than 900 notes spanning over more than 10 years. The distribution of the record sizes is shown in Figure 6 in terms of number of notes and 1000-character text chunks. More than half of the records contain more than 32 but less than or equal to 128 notes.

![Figure 6](image)

Figure 6. Distribution of patient record sizes.

Inter-Annotator Agreement for the Initial Ground Truth

Because the number of non-problems are significantly larger than the true problems for open-ended problem lists, we use recall, precision, and \(F_1\) score of true problems as surrogates for the typical Cohen’s Kappa in reporting inter-annotator agreement. Using the initial adjudicated result (Initial GT) as the ground truth, the inter-annotator agreement accuracies are shown in Table 2. The recall for one annotator was as low as 0.78 but the average across three annotators was 0.86, which shows that finding all meaningful problems from a longitudinal patient record is not an easy task, even for a knowledgeable human annotator with adequate time.

| Annotator       | TP  | FP  | FN  | Recall | Precision | \(F_1\) |
|-----------------|-----|-----|-----|--------|-----------|--------|
| Annotator A     | 334 | 27  | 98  | 0.78   | 0.93      | 0.84   |
| Annotator B     | 461 | 53  | 61  | 0.88   | 0.90      | 0.89   |
| Annotator C     | 427 | 31  | 44  | 0.91   | 0.93      | 0.91   |
| **Summary/Average** | **1,122** | **111** | **203** | **0.86** | **0.92** | **0.87** |
Ground Truth Vetting Iterations

On average, a patient record in our corpus had 6.5 problems in the adjudicated initial ground truth, excluding the 74 patient records that did not have any problems. On average, 0.8 problems were added to each patient record after the first iteration, and another 0.3 problems were added after the second iteration. The number of problems added for each patient record through the vetting iterations is shown as a distribution of problems per patient record in Figure 7a.

Figure 7a. Changes in the distribution of ground truth problems per patient record through vetting iterations (left). 7b. Problems and codes added in each iteration; “System Negative” indicates candidates classified by the system as non-problems, and “System Positive” indicates candidates classified by the system as “true” problems (right).

As described earlier in the “Iterative Vetting Process” section, annotators reviewed System-only problems in a system generated, highly recall-oriented (F₁ optimized) list as part of the vetting process, whereas the system is configured to generate a balanced (F₁ optimized) or a recall-oriented (F₂ optimized) list in production mode. The F₁ optimized list included both System Negatives (non-problems according to the F₁ or F₂ optimized system) and System Positives (true problems according to the F₁ or F₂ optimized system). This vetting process resulted in the addition of new problems to the ground truth as well as new SNOMED CT CORE codes to existing ground truth problems. Figure 7b shows the problems and the codes added to the ground truth in each iteration from the System Negatives and System Positives. More problems and codes were added from the Negatives than the Positives, and as expected, the additions have reduced from the first iteration to the second.

In addition to reviewing problems identified by the system (System-only), annotators also vetted problems that were in the ground truth but were missed or assigned a very low score by the system (GT-only). About 5% of the problems in the initial ground truth were identified as potential ground truth errors in this step, where at least one annotator considered the problem as a candidate for removal from the ground truth. At the time of writing, no changes have been made to the ground truth to remove these potentially erroneous problems, pending annotator adjudication and review by an MD.

The Impact of Ground Truth Improvements on System Performance

Two iterations of vetting of the System-only problems were conducted on the initial adjudicated ground truth, during which new problems were added and codes were added for existing problems. Using the final version as the reference, the accuracy of the ground truth in each iteration and the accuracy of the problem list generation system trained using the corresponding ground truth are shown in Figure 8a. As the ground truth underwent additional vetting passes, system performance improved, especially its recall. The number of problems in the ground truth increased by about 9%, system recall improved from 0.71 to 0.74 and precision improved from 0.70 to 0.71, resulting in optimum F₁ score improving from 0.71 to 0.72. The improved ground truth also increased the area under the curve in the precision-recall curve as shown in Figure 8b.

Note that the increase in system performance is not as significant as the ground truth itself. This can be understood in two ways. First, the system accuracy has a smaller range comparing to the ground truth accuracy range. Without ground truth, the system can have an F₁ score of 0.4 from the prior knowledge. On the other hand, even with a perfect ground truth, the system performance is still limited by other factors such as the complexity of the model, the type and accuracy of the features, and the number of training samples. Second, problems predicted as both positive and negative by the system were added to the ground truth during vetting. The former increases the true-positive count and decreases the false-positive count in the system, and the latter increases the false-negative count in the system, which
decrease the system’s recall. As can be seen in Figure 7b, although more problems were added from the System Negative category (that reduces the system’s recall), after retraining, the system still achieved a higher accuracy.

**Discussion**

This ground truth generation process uncovered individual annotator errors accounting for inter-annotator disagreements, and ground truth errors in the initial adjudicated problem list. The ground truth errors discovered during vetting consisted of missed problems, missed codes, and erroneously-identified problems. In the following sections, we discuss these errors in two general groups, problem identification errors and code mapping errors, and explore reasons why these errors occur.

**Errors in Problem Identification**

Review of annotator comments during vetting revealed that missed problems during initial ground truth generation were most often due to infrequent mentions in the patient record. Some examples of such problems include:

1. problems that require monitoring without active treatment (e.g. 1st degree heart block, mitral valve prolapse),
2. asymptomatic problems discovered incidentally (e.g. cyst of ovary), and
3. problems are in remote history but considered relevant for current care (e.g. depressive disorder).

Problems were also left out of the initial ground truth because of annotator confusion on whether the problem is true in the patient (due to conflicting documentation in the patient record), and whether the problem belongs in the problem list based on the guidelines. We expand on these topics in later sections.

Qualitative review of ground truth problems not found by the system (GT-only) also cited conflicting documentation and guideline confusion as the main reasons for considering removal of problems from the initial ground truth. For those “good” ground truth problems that were missed by the system, we are using annotator feedback to explore ways to increase the signal from these problems in the system.

**Errors in Mapping to Standard Vocabulary**

Missed codes, for problems already in the ground truth, were usually more general or more specific variants of the ground truth problem. Some reasons for why these codes were omitted during the initial ground truth generation are detailed below.

Firstly, annotators may miss information within the patient record that would validate certain variants of the ground truth problem. An example is a patient with the problem “COPD,” which was identified and included in the initial ground truth; the system found mentions of “chronic bronchitis,” which describe the specific type of COPD that this patient has, and therefore should be included as a name variant and valid code for the ground truth problem “COPD.”

Another reason is that the annotator simply missed the code for this problem when manually going through the list of 6000+ codes in the current non-retired SNOMED CT CORE subset. This is unavoidable in a manual process due to the long list of coded concepts coupled with annotator fatigue. Also, the same problem may need to be reworded to match a code in CORE. For example, there is no CORE code for the problem “erectile dysfunction”, but there is one for the equivalent concept of “impotence.” Use of a shared mapping resource would help ensure consistent mapping.
and decrease the number of missed codes. However, since some mappings are patient specific while other mappings are not, it is difficult to create a resource that applies to all patients.

Figure 9 shows a subset of CORE codes that may be used to represent the problem “COPD,” and illustrates the various considerations and challenges in identifying all valid codes for a problem. The top seven codes were included in the initial ground truth; some codes are more specific than the target problem, some are more general, and some depend on patient specific disease states. The other codes marked with [X] are additional codes not in the ground truth. However, it may happen that if the system finds one of these, it is correct and should be added to the ground truth. We also show all terms found in notes that are linked to the CORE codes for “COPD.”

![Figure 9: SNOMED CT CORE codes for the problem “COPD”](image)

**Figure 9.** SNOMED CT CORE codes for the problem “COPD”.

**Other Corrections during Vetting**

In addition to missed problems and missed codes, there were also cases where (1) the system-proposed problem was incorrect, but investigation into that problem led the annotators to add a different new problem that was missed to the ground truth, or (2) the system-proposed problem was added to replace an existing problem in the ground truth.

In one case (1), the system-proposed problem “iron deficiency anemia” was (correctly) considered as a false positive as it was mentioned in the context of a test being ordered to investigate for this disease, hence a hypothetical mention and not a definitive diagnosis. However, investigation into this problem revealed that although the patient could not be confirmed to have “iron deficiency anemia,” the patient did have chronic anemia, and therefore a new problem “anemia” was added to the ground truth. In a different case (2), vetting resulted in a new system-proposed problem “steatosis of liver” replacing an existing problem “non-alcoholic fatty liver disease” in the ground truth, because investigating “steatosis of liver” revealed that the disease was not specified in the patient record to be non-alcoholic.

These cases illustrate how error-prone the initial ground truth generation process can be. They also demonstrate how even if the system-proposed problem is incorrect, it may provide useful information and guide the annotators to other problems that may have been missed during the initial review of the patient record. All in all, an iterative approach to ground truth generation improved the quality of ground truth, thereby improving the performance of the system trained on such ground truth. This same approach may be applied to any ground truth generation involving concept extraction and/or relation detection in clinical documentation.

**Difficulties in Defining a Problem List**

Because we had different annotators contribute to different parts of this data generation throughout the ground truth development task, it was essential to have a clearly defined description of what should or should not be included in
the problem list to ensure consistency of the data generated. However, even with our revised guidelines beyond what was described by CMS for meaningful use, there is still room for subjective interpretation by individual annotators.

Our guidelines specify that chronic problems and recurrent problems should be included, while resolved problems should not. However, terms such as “chronic,” “recurrent” or “resolved” are, to a certain degree, subjective. How many times does an acute disease need to occur to be considered recurrent? Can an acute disease be assumed to be resolved if no further documentation of the problem exists? Secondly, the annotators, when creating the problem list, have the context of the entire patient record as the backdrop when considering a problem, which leads to specific cases that may not fit in nicely with our artificial guidelines. For example, we may generalize in the guidelines that certain problems, such as diabetes mellitus and hypertension, are always chronic due to the nature of the problem. However, we encountered in our dataset an obese patient with diabetes mellitus and hypertension, typically chronic diseases, both of which were documented as being resolved after the patient experienced significant weight loss due to gastric bypass surgery. Through this experience, we recognize and confirm the inherent difficulty in defining a problem list given the complexity of clinical information and variability among real patients and their problems.

Conflicting Documentation within the Patient Record

Sometimes there is conflicting evidence within the patient record leading to uncertainty, even by medical experts, as to whether the patient does or does not have a problem. For example, we had a patient record which alternately documented the patient as having Type I Diabetes Mellitus and Type II Diabetes Mellitus; we had another patient that was documented in one electrocardiogram (EKG) to have heart block, and in a subsequent EKG to have normal sinus rhythm, although no intervention was given between the two tests. Conflicting documentation within the patient record may be due to multiple reasons, including misdiagnoses, typos, and copy-paste behavior that propagates such errors. In some cases, we can use other clues such as the note type, note author or related medications to infer whether the problem is true for the patient. However, in other cases it remains unclear. If human experts cannot tell whether the patient does or does not have a problem, it would be unreasonable to expect the system to be able to do so.

Comparison to Existing Annotation Works

Previous work in annotation of longitudinal clinical records have included some element of support from a system. However, the involvement of a system is usually in the form of a static algorithm or script used to check annotator work for errors. For the task of identifying protected health information (PHI) in medical narratives, Stubbs et al. wrote scripts to look for PHI text annotated in one part of the records but not in others16, while Neamatullah et al. used a prototype de-identification algorithm to check for potential ground truth errors17. Our work involving iterative vetting differs in that instead of using a static algorithm, we used the system being trained on each version of the revised ground truth to simultaneously improve our ground truth as well as our trained system.

Future Direction

Currently, we define a general-purpose problem list that is useful for a comprehensive health assessment by primary care practitioners or internists. However, the problem list, as a part of the medical record, is used by multiple users with different interests and workflow processes; what is of interest to an obstetrician is likely different from what is of interest to an ophthalmologist. In addition, although medical professionals are familiar with the concept of a problem list, different providers may have slightly different interpretations on what exactly belongs in the problem list4. Given this, what we have defined as ground truth in the problem list generation effort described here may not align with what an end-user of the system is looking for when they want a problem list.

We are in the process of exploring a new, more flexible structure for ground truth generation that would allow mapping to different problem list definitions. The proposed method would involve not only determining whether a problem belongs in the problem list (a binary classification), but also include annotations on specified attributes that a clinician may consider when deciding whether a problem should go in the problem list. These attributes may be problem-specific (e.g. “dermatologic problem”, “benign neoplasm”) or patient-specific (e.g. “actively presenting with symptoms,” “surgically corrected”). In addition to allowing for a flexible problem list definition, this new approach would also allow us to incorporate patient-specific attributes to capture the problem in the context of a full patient record, and help researchers better understand the domain experts’ thought process.

Conclusion

We presented an iterative vetting approach for creating problem list ground truth from patient medical records, for use in developing an automated problem list generation system based on NLP and supervised machine learning. We reported on the methodology used, the data generated, its effect on system performance, and lessons learned from the
experience. Using precision, recall, and F₁ score as metrics, we found that system performance increases with repeated rounds of vetting. An average of 0.8 problems were added per patient record after the first vetting pass, and another 0.3 problems were added after a second vetting pass. After two rounds of vetting, system recall increased from 0.71 to 0.74, precision increased from 0.70 to 0.71, and F₁ score increased from 0.71 to 0.72. Reasons for errors during initial ground truth creation included infrequently mentioned problems, coding errors, difficulties in defining a problem list, and conflicting documentation in patient records. This methodology is applicable for many higher-level methods such as disease progression modeling, clinical factor extraction (such as the history of a symptom or disease), and detection of drug adverse effects in a patient record. Any complex and higher-level semantic concept detection or chronological observation task, which typically requires foraging through a longitudinal patient record and/or assimilating several pieces of evidence, is a good candidate for this machine-in-the-loop (of a human activity) approach. Human fallibility in creating ground truth can be helped by the system being developed, which in turn benefits from the improved ground truth quality.

Acknowledgements

We acknowledge Cleveland Clinic for the patient medical record data, and the students from New York Medical College and SUNY Downstate Medical Center for their enthusiastic and dedicated annotation work. We thank Lauren Mitchell for careful proof reading of the manuscript and for making insightful suggestions to make it better readable.

References

1. Wachter R. The Digital Doctor. McGraw-Hill; 2014.
2. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, West CP. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc. 2016;91(7):836-48.
3. Weed LL. Medical records that guide and teach. N Engl J Med. 1968;278:652-657.
4. Holmes C. The problem list beyond meaningful use, part 1: the problems with problem lists. J AHIMA. 2011;81(2):30-33.
5. Devarakonda M, Tsou CH. Automated problem list generation from electronic medical records in IBM Watson. Proc IAAL. 2015:3942-3947.
6. Bui AA, Taira RK, El-Saden S, Dordoni A, Aberle DR. Automated medical problem list generation: towards a patient timeline.
7. Meystre, SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inf. 2008;77:602-612.
8. Odifreddi P, Cooper SB. Recursive Functions [Internet]. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy (Fall 2016 Edition). Metaphysics Research Lab; 2016 [cited 2016 Sept 22]. Available from: http://plato.stanford.edu/archives/fall2016/entries/recursive-functions/
9. Center for Medicare and Medicaid Services. Eligible professional meaningful use core measures 3 of 13 [Internet]. 2014 [cited 16 Sept 2016]. Available from: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/3_Maintain_Problem_ListEP.pdf
10. The CORE problem list subset of SNOMED CT [Internet]. Bethesda (MD): U.S. National Library of Medicine. 2014 – [cited 2016 Sept 16]. Available from: http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html. Accessed 2016 Sep 16.
11. Prager JM, Liang JJ, Devarakonda M. SemanticFind: Locating what you want in a patient record, not just what you ask for. AMIA Jt Summits Transl Sci Proc. Forthcoming 2017.
12. Tsou CH, Devarakonda M, Liang JJ. Toward generating domain-specific / personalized problem lists from electronic medical records. AAAI 2015 Fall Symp. 2015:66-69.
13. Holmes C. The problem list beyond meaningful use: part 2: fixing the problem list. J AHIMA. 2011;82(3):32-35.
14. ICD-9-CM Diagnostic Codes to SNOMED CT Map [Internet]. Bethesda (MD): U.S. National Library of Medicine. 2015 [cited 2016 Sept 16]. Available from: https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html
15. SNOMED CT to ICD-9-CM Rule Based Mapping to Support Reimbursement [Internet]. Bethesda (MD): U.S. National Library of Medicine. 2009 [cited 2016 Sept 16]. Available from: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd9cm_reimburse.html
16. Stubbs A, Uzuner O. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth Corpus. J Biomed Inform. 2015;58(Suppl):S20-S29.
17. Neamatullah I, Douglass MM, Lehman LH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008;8:32.