Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department

Louise Deleger,[1] Holly Brodzinski,[2] Haijun Zhai,[1] Qi Li,[1] Todd Lingren,[1] Eric S Kirkendall,[1,3,4] Evaline Alessandrini,[2,4] Imre Solti[1,4]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001962).

[1]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[2]Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[3]Division of Hospital Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[4]James M. Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Correspondence to
Dr Imre Solti, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA; imre.solti@cchmc.org

Received 19 April 2013
Revised 20 September 2013
Accepted 27 September 2013
Published Online First
15 October 2013

To cite: Deleger L, Brodzinski H, Zhai H, *et al*. *J Am Med Inform Assoc* 2013;**20**:e212–e220.

ABSTRACT

Objective To evaluate a proposed natural language processing (NLP) and machine-learning based automated method to risk stratify abdominal pain patients by analyzing the content of the electronic health record (EHR).

Methods We analyzed the EHRs of a random sample of 2100 pediatric emergency department (ED) patients with abdominal pain, including all with a final diagnosis of appendicitis. We developed an automated system to extract relevant elements from ED physician notes and lab values and to automatically assign a risk category for acute appendicitis (high, equivocal, or low), based on the Pediatric Appendicitis Score. We evaluated the performance of the system against a manually created gold standard (chart reviews by ED physicians) for recall, specificity, and precision.

Results The system achieved an average F-measure of 0.867 (0.869 recall and 0.863 precision) for risk classification, which was comparable to physician experts. Recall/precision were 0.897/0.952 in the low-risk category, 0.855/0.886 in the high-risk category, and 0.854/0.766 in the equivocal-risk category. The information that the system required as input to achieve high F-measure was available within the first 4 h of the ED visit.

Conclusions Automated appendicitis risk categorization based on EHR content, including information from clinical notes, shows comparable performance to physician chart reviewers as measured by their inter-annotator agreement and represents a promising new approach for computerized decision support to promote application of evidence-based medicine at the point of care.

OBJECTIVE

The objective of this study is to evaluate a proposed natural language processing (NLP) and machine-learning (ML) based automated method to risk stratify abdominal pain patients by analyzing the content of the electronic health record (EHR). Our approach relies on both the structured data and the narrative clinical text (physician notes) of the EHR.

BACKGROUND AND SIGNIFICANCE

Identifying the small minority of patients with acute appendicitis among the nearly 2 million annual pediatric emergency department (ED) visits for abdominal pain is challenging.[1] A recent study reported that the rate of acute appendicitis was 10.6/10 000 in children (0–19 age group).[2] In our institution's pediatric ED, the proportion of patients with acute appendicitis among those with abdominal pain was estimated at 12.7% in the year 2010. The use of CT in the diagnostic workup of abdominal pain has become widespread,[3] despite the small but significant associated increase in the lifetime risk of malignancy from ionizing radiation,[4 5] unclear benefit,[6] and increases in costs of care.

Patients who ultimately undergo appendectomy theoretically represent those with the highest risk of appendicitis based on clinical findings, and therefore have less of a need for diagnostic imaging (compared to patients with more equivocal clinical findings). Therefore, limiting the use of CT to cases where it is necessary, evidence-based, and helpful (ie, where clinical findings are equivocal) would improve care by decreasing variation, as well as improving safety and value.

Clinical scoring systems, such as the validated Pediatric Appendicitis Score (PAS)[7] have been developed to stratify patients with suspected appendicitis and avoid unnecessary imaging. While the use of scores as stand-alone diagnostic tools is controversial, many studies have concluded that they could help in the decision-making process.[8–11] The PAS is the most broadly validated tool for appendicitis in children[12] and was used at our institution to establish the risk stratification pathway on which our work is based.

Clinical decision support (CDS) systems can improve adherence to guidelines and support efficiency by decreasing utilization of healthcare resources.[13] A computerized CDS tool to stratify patients according to risk of appendicitis could help reduce the unnecessary use of diagnostic imaging.

Automated EHR-based CDS tools extract relevant information from the medical record in real time and present recommendations as soon as possible in the encounter. In many cases, a large part of the clinically relevant information is located in the physician notes. Accessing this information necessitates information extraction via NLP. Several NLP methods have been developed for identifying or classifying patients with a decision support objective, and have shown promising results for a variety of clinical conditions, including pneumonia, tumor status, heart failure, and cervical cancer.[14–20] Such NLP approaches have been knowledge-based, using linguistic rules and lexica,[14 15 18 20] and/or ML-based.[16 17 19]

To our knowledge, our study is the first to investigate a method for automated appendicitis risk stratification. It is an example of information fusion where structured EHR information (ie, laboratory values) and unstructured information extracted from clinical notes via NLP are integrated to achieve a better performance in computerized decision support.[21 22]

MATERIAL AND METHODS

PAS and risk-stratification pathway

Risk stratification was defined using an evidence-based clinical pathway created at our institution through collaboration among the divisions of emergency medicine, surgery, and radiology and based on the validated PAS, a 10-point scale that assesses appendicitis risk based on history, physical exam, and lab findings.[7] Variables used to calculate this score are presented in table 1.

The risk stratification pathway defines pediatric patients, aged 3–21, as:

- *High-risk* for acute appendicitis with a PAS ≥7
- *Equivocal-risk* with a PAS of 3–6
- *Low-risk* with a PAS ≤2

Clinical data selection and creation of a gold standard

This is a retrospective observational study (conducted under an approved IRB protocol) using patient records from a pediatric ED in an urban, quaternary care children's hospital. The ED serves approximately 120 000 visits per year and has 24-h radiology (including ultrasound) and subspecialty services available.

We selected our study sample from the 6500 patient–physician encounters for abdominal pain in all age groups that occurred during a 12-month period (January 1, 2010–December 31, 2010). We included all patients who had appendectomy as a consequence of the ED visit (534 patients), and we added a random sample of 1566 patients from the remainder of the abdominal pain patients, to have a total of 2100 patient records (1000 per annotator pair, with 100 extra records common to all annotators). Each record included all ED physician and nursing notes as well as structured entries corresponding to relevant labs and vital signs (complete blood counts with differential values and temperature recordings). Other non-ED notes (such as surgical consultation or admission notes) were excluded, to avoid capturing information that would not be available during the decision-making portion of the ED visit.

Records were reviewed and manually double-annotated by four senior pediatric emergency medicine fellows. For each record, elements needed to calculate the PAS were manually tagged in the clinical notes. Annotators were instructed to mark all mentions of these elements (separate categories were created for negated elements, eg, a 'no fever' category to catch negations such as '*negative for fever*'). After reviewing a record, the physicians computed the PAS (from 0 to 10) based on the elements they found in the notes, the labs, and vital sign entries. Then they assigned a risk class (low, equivocal, or high) to the patient according to the pathway. Records were presented to the annotators without showing surgery or pathology notes, to avoid introducing bias. Annotators (A1 through A4) were paired up for the study; A1 was paired with A2 and A3 with A4. Each pair was given 1000 records to review, plus 100 records common to all four annotators (the 'common sample') to measure *inter*-annotator agreement across all annotators. Additionally, each annotator was given a random sample of 100 records to annotate a second time (the 'repeat sample'), to measure *intra*-annotator agreement. Annotation was supervised by a medical informatics researcher (LD) with text annotation experience and by two faculty physicians (a pediatric emergency medicine specialist (HB) and a pediatric hospitalist (EK)).

Table 1 Variables of the Pediatric Appendicitis Score

Pain with cough, percussion or hopping	2 points
Right lower quadrant tenderness	2 points
Anorexia	1 point
Fever >38°C	1 point
Nausea/vomiting	1 point
Migration of pain	1 point
Leukocytosis: >10 000 white blood cells per μL	1 point
Left shift: absolute neutrophil count >6750	1 point

Automated risk-stratification

Our approach consists of two steps (summarized in figure 1):

1. Information extraction: identifying the PAS elements in the clinical notes.
2. Risk stratification: assigning a PAS and a risk class to each record using the elements identified in step 1.

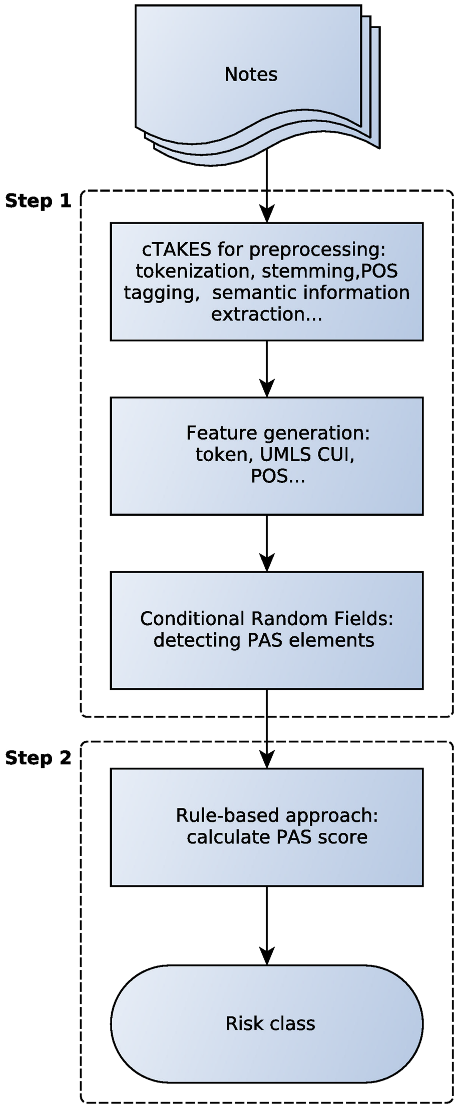

Notes
Step 1
cTAKES for preprocessing: tokenization, stemming,POS tagging, semantic information extraction...
Feature generation: token, UMLS CUI, POS...
Conditional Random Fields: detecting PAS elements
Step 2
Rule-based approach: calculate PAS score
Risk class

Figure 1 Description of the risk stratification system (cTAKES=clinical Text Analysis and Knowledge Extraction System; POS, part-of-speech; UMLS, Unified Medical Language System; CUI, concept unique identifier; PAS, Pediatric Appendicitis Score).

Information extraction

We extracted PAS elements from clinical notes based on a hybrid NLP approach using an ML core algorithm (conditional random fields (CRF),[23] as implemented in the MALLET toolkit[24]) with post-processing rules.

Before training the CRF model, we performed linguistic preprocessing of the notes, using the clinical Text Analysis and Knowledge Extraction System (cTAKES) toolkit,[25] which included tokenization, stemming, and part-of-speech tagging. cTAKES also contains a lexicon lookup component, matching the text against concepts from the Unified ML System (UMLS) metathesaurus[26] and assigning concept unique identifiers (CUIs) to the matched text tokens. We utilized this semantic information in our CRF model. The feature set is detailed in box 1.

Based on error analysis of the development set, we defined regular expression rules to improve the output of the CRF. For instance, we checked that elements containing numerical values were tagged appropriately (eg, that phrases such as '*temp of 101*' were tagged as 'fever' and not mistakenly tagged as 'no fever').

Risk stratification

We used a rule-based approach for risk stratifying the patients. This method consisted of gathering all the different PAS elements found during step 1 in the clinical notes as well as those contained in the labs and vitals entries, calculating their total number of points (per the PAS scoring system) to generate a PAS score. We then assigned a high-, equivocal-, or low-risk class, based on the computed score.

We also implemented a baseline system by computing the PAS and risk class based on variables obtained from the structured data of the EHR only. This allows us to measure the benefit of using information from the narrative text of the EHR.

Box 1 Feature set for the conditional random fields algorithm

Properties of the current token
- Token itself
- Lower-case version of the token
- Stem of token
- Character length of the token
- POS of the token
- Whether the token is a punctuation
- Whether the token contains a digit
- Whether the token is a number
- Whether the token is alpha-numeric
- Whether the token is capitalized
- Whether the token is upper-case only
- Whether the token contains upper-case and lower-case letters
- Two-character suffix and prefix of the token
- UMLS CUI(s) of the token (if any)

Contextual features
- Two tokens before
- Two tokens after
- Bigram of the current token and token after
- Bigram of the two tokens before
- Trigram of the current token and two tokens after
- Trigram of the three tokens before
- POS of the two tokens before and two tokens after
- Whether the token before is capitalized
- Whether the token after is capitalized
- Note section where the token is appearing (eg, history of present illness, physical exam, etc.)
- Whether a negation phrase (eg, 'negative for') occurs in the sentence containing the token

(POS, part-of-speech; UMLS, Unified Medical Language System).

Experiments

Evaluation metrics

We used *recall*, *precision*, and *specificity* to evaluate the performance of our system.[27] For one of our experiments, we also report the *F-measure* (the weighted harmonic mean of precision and recall).[28]

Additionally, we computed the F-measure for the *intra*-annotator and *inter*-annotator agreements.

Annotator agreement

We assessed the reliability of the gold standard by computing agreement for the risk classification and the annotation of the PAS elements in the clinical notes. For each level, we measured agreement between the paired annotators on the sets of 1000 records and among all four annotators on the common sample of 100 records. We also tested *intra*-annotator agreement to measure agreement between each annotator's earlier and subsequent annotation of the same charts on the repeat sample of 100 records.

Performance of the automated system

Performance of the risk classification

We measured the performance (recall, precision, specificity) of the main step of our system, that is, risk classification, by comparing the system's classifications against the gold standard, in a subset of 1890 patients (randomly selected, 90% of the total). The 1890 records were kept unseen during development and were used to train and evaluate the system in a 10-fold cross-validation setting. Metrics were computed for each risk class, as well as their average on all classes, for both our system and the baseline. To rule out the possibility that the performance difference between the system and the baseline was due to chance, we tested statistical significance, using approximate randomization.[29]

We performed an error analysis of the risk classification on the remaining 210 patient records.

Performance of the detection of PAS elements in notes

We also assessed the performance (recall, precision, F-measure) of the first step of our system (the detection of PAS elements in clinical notes).

Timestamp experiment

A real-time production system is most efficient if the information is recorded in the EHR in real-time (ie, when it is useful for decision making). Consequently, we evaluated the algorithm's risk classification performance using information available at different points of the 48-h study window. More precisely, we ran our algorithm consecutively 48 times using cumulative data available after each hour of a 48-h study window of the ED visit. The purpose of this experiment was to assess the point in time of the ED encounter when high performance could be reached, that is, when enough data were present in the EHR for the algorithm to perform well. Additionally, we also evaluated performance up to the time patients were sent for an abdominal CT scan, to see if the algorithm could prevent unnecessary CT scans.

RESULTS

Annotator agreement

Table 2 shows *inter*-annotator agreement results on the 1000-record sets from the two annotator pairs (A1/A2 and A3/A4), as well as on the common sample annotated by all four annotators. Agreement was strong for the risk classification (above 0.86 in most cases) and slightly lower for the PAS elements (from 0.75 to 0.85). The last three columns detail *inter*-annotator agreement for each risk class. Agreement was highest for the low-risk class and lowest for the equivocal-risk class.

Intra-annotator agreement measured between each annotator's earlier and subsequent annotation on the repeat sample showed similar values to *inter*-annotator agreement values (bottom of table 2), except for the high-risk class, which had the lowest *intra*-annotator agreement.

Descriptive statistics

The final gold standard (after resolution of disagreements) consisted of 31 478 PAS element annotations (see online supplementary table for statistics per category), 1039 low-risk patients, 637 equivocal-risk patients, and 424 high-risk patients. Figure 2 depicts the distribution of cases in the gold standard (figure 2A) and in the predictions of the system (figure 2B) between the three risk classes for all patients. Compared to the annotators, the system classified more patients as equivocal-risk (34% in the system output vs 30% in the gold standard).

Automated risk stratification

Performance of the risk classification

The system obtained average recall of 0.869 and precision of 0.868 (table 3). It shows best performance at classifying low-risk patients (0.897 recall; 0.952 precision) and has the lowest performance in the equivocal-risk class (0.854 recall; 0.766 precision). The baseline system (using only structured data) obtained much lower performance (average recall and precision of 0.353 and 0.385, respectively), and is unable to correctly classify any high-risk patients. The difference between our system and the baseline was statistically significant (last column, $p<0.05$).

Performance of the detection of PAS elements in notes

The first step of our system achieved 0.853 recall and 0.878 precision against the gold standard. Detailed performances are given in figure 3.

Timestamp experiment

Figure 4A shows risk classification performance (F-measure) for each class and on average at different points in time (from 1 to 48 h after presentation to the ED, in 1-h time intervals). We can see that performance curves are high after 2–3 h and approach the plateau after 4 h, which means that approximately 4 h of data is enough to achieve optimal performance. Figure 4B gives a closer look at the first 4 h, showing F-measures for each risk class. While the low-risk class has good performance from the start, the high- and equivocal-risk classes start having reasonably good performance at 3 h.

Figure 5 shows how classification changes over time compared to the eventual system's predictions. For each eventual risk class, we report the predictions at each hour (how many patients are assigned a different risk class than the eventual one and how many have their eventual risk class). Percentages in the graphs show the proportion of patients with a classification identical to the eventual prediction. Figure 5A shows that not all eventual high-risk patients are identified from the start (many are classified as low-risk and equivocal-risk), but a high proportion is reached at 3 h (80.4%) with only a small number classified as low-risk. A similar tendency is observed for eventual equivocal-risk patients with a majority identified at 3 h (figure 5B). Eventual low-risk patients are almost all identified from the start (figure 5C). We can also observe from the low-risk and equivocal-risk graphs that patients classified as high-risk never end up as low-risk and almost never as equivocal-risk, which means a high-risk prediction is final from the start.

Figure 6 displays the CT time distribution for patients who had a CT scan (267). Average time is 241 min after ED arrival, with the median at 239 min. Half the CT scans happened more than 4 h after ED arrival. At 3 h, 42% patients have been sent to CT.

DISCUSSION

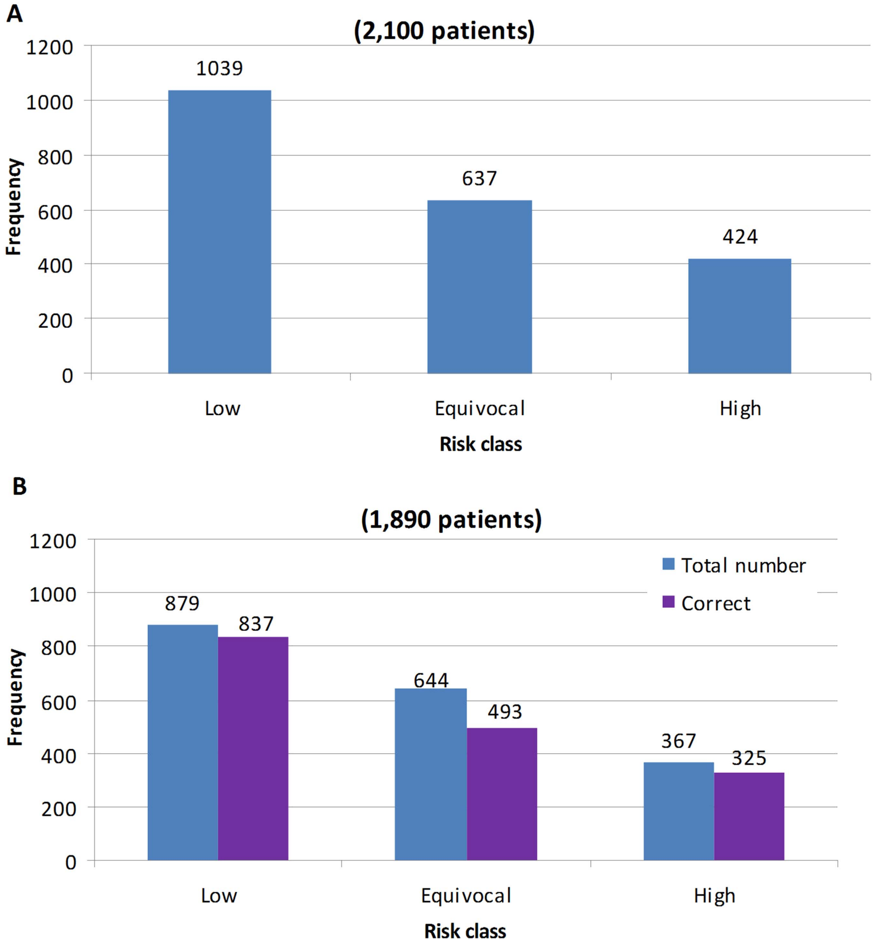
The main goal of our work was to automatically assign the appropriate appendicitis risk classification (high-, equivocal- or low-risk) to pediatric abdominal pain patients. The risk classification system performed very well overall, achieving only 3% percent lower performance than the average *inter*-annotator

Table 2 Inter-annotator agreement (F-measure) on the 1000-record sets and on the (100-record) common sample and intra-annotator agreement on the repeat sample

Record set	Annotator(s)	Risk class	PAS elements	High-risk	Equivocal-risk	Low-risk
1000-record sets (inter-annotator agreement)	Pair 1 (A1/A2)	0.896	0.830	0.898	0.841	0.938
	Pair 2 (A3/A4)	0.897	0.783	0.906	0.845	0.927
Common 100-record sample (Inter-annotator agreement)	A1/A2	0.900	0.855	0.947	0.833	0.918
	A3/A4	0.864	0.789	0.829	0.812	0.921
	A1/A3	0.861	0.814	0.821	0.818	0.903
	A1/A4	0.836	0.748	0.895	0.769	0.851
	A2/A3	0.889	0.836	0.878	0.813	0.946
	A2/A4	0.843	0.765	0.900	0.762	0.872
Repeat sample (intra-annotator agreement)	A1	0.901	0.811	0.800	0.862	0.946
	A2	0.895	0.844	0.880	0.841	0.933
	A3	0.929	0.858	0.857	0.899	0.962
	A4	0.890	0.783	0.818	0.844	0.930

PAS, Pediatric Appendicitis Score; Risk class, low or equivocal or high risk for appendicitis; PAS elements, PAS terminology annotated in the text (eg, 'fever').

Figure 2 Distribution of cases in the gold standard (A) and system output (B). Correct cases are determined against the gold standard of Pediatric Appendicitis Score-based risk classification.



agreement for the annotators. The system had a higher performance for overall risk strata than four of the six physician pairs on the common 100 charts. In addition, the system achieved only 2% lower performance than two of the physicians, 3% lower than one of the physicians, and 5% lower than the fourth physician—measured by their *intra*-annotator agreement on repeat chart reviews. From a practical point of view, the system behaved as a fifth physician and its performance was not distinguishable from the four experts.

In this type of task, the tradeoff among recall, precision, and specificity is an important issue. In order to be useful to physicians, one would expect a system to be able to identify patients in the low-risk and high-risk category with a high precision; patients can then, with certainty, either avoid a CT scan for the low-risk class or proceed to surgical consultation (thus potentially bypassing imaging) for the high-risk class. However, recall should also be reasonably high (enough patients should be classified as high- and low-risk), otherwise the system will identify too few cases to be of any use. For the low-risk class, our system has high specificity and precision (respectively, 0.956 and 0.952), and a recall that is lower, but still high (0.897). For the high-risk class, it has a very high specificity (0.972), and good precision and recall (0.889 and 0.855, respectively), although there is room for improvement. However, the system had comparable results to human annotators.

Error analysis on the 210-record development set showed that all misclassified low-risk patients (11) and high-risk patients (4) were classified as equivocal, while misclassified equivocal patients (6) were classified as low-risk in some cases and as high-risk in others. The equivocal category is the most challenging, particularly for patients with a PAS corresponding to the lower bound (PAS=3) and upper bound (PAS=6) of the class. The error analysis revealed that misclassification occurred most often when the difference between the gold standard PAS and the PAS computed by the system was only one (4 cases) or two (12 cases) points. Thus, in most cases, errors are due to only one or two PAS element categories being missed or erroneously detected at the first step (identification of PAS elements). The analysis also showed that none of the patients misclassified as low-risk actually had an appendectomy as a result of their ED visit, which is a very promising finding for our algorithm.

The large difference in performance between relying solely on structured data (eg, temperature and lab values) and additionally incorporating clinical variables via NLP demonstrates the importance of using information extracted from clinician notes. Clinical notes relay important medical information needed for risk stratification as well as communicating clinical care. The ability to capitalize on this information using NLP to impact real-time decision-making and clinician behavior (in lieu of a plethora of check boxes) is a critical step for information-enabled solutions to improve healthcare. In clinical scenarios such as suspected appendicitis, use of structured data alone (such as can be extracted via traditional EHR-based decision support methodologies) is helpful but not sufficient to support diagnostic decision making.

The timestamp experiment showed that the data required for optimal performance of our risk stratification system was entered in the EHR within 4 h of the patient's arrival at the ED, on average, when evaluating all of the patient's data in aggregate. In general, the time between ED arrival and evaluation by

Table 3 Automated risk stratification performance

	Recall				Precision				Specificity				Stat. sig. of the difference between baseline and full system (p value)
	Baseline	95% CI	System	95% CI	Baseline	95% CI	System	95% CI	Baseline	95% CI	System	95% CI	
High	0	0 to 0.003	0.855	0.815 to 0.888	NaN	NaN	0.886	0.847 to 0.915	1	0.988 to 1	0.972	0.962 to 0.980	NaN
Equivocal	0.059	0.042 to 0.082	0.854	0.822 to 0.882	0.236	0.171 to 0.315	0.766	0.731 to 0.798	0.916	0.900 to 0.930	0.885	0.866 to 0.902	0.0001*
Low	1.000	0.995 to 1	0.897	0.875 to 0.916	0.534	0.511 to 0.558	0.952	0.935 to 0.965	0.150	0.129 to 0.175	0.956	0.941 to 0.968	0.0001*
Average	0.353		0.869		0.385		0.868		0.689		0.938		0.0001*

Stat. sig., statistical significance (* indicate statistically significant values); NaN, 'not a number', used to indicate values that cannot be computed.

a clinician is approximately 1 h. In a teaching institution, the diagnostic evaluation begins after this time period and may be delayed additionally in the event that residents are involved in care. Accounting for a lab turnaround time of 30–60 min, it makes sense that the data is most complete and prediction most accurate approximately 4 h after ED arrival. The average length of stay in our ED (door to disposition) for patients with appendicitis is 257 min (4.28 h); thus, maximal performance of the system at 3–4 h from ED arrival would enable the system to provide risk stratification to clinicians in real time as decisions are being made.

Due to the clinical nature and evolution of appendicitis, patients presented to the ED in varying states along that course. Some patients may have been diagnosed before this 4 h had elapsed (eg, a high-risk patient whose diagnosis may be obvious), while others may be exhibiting the very beginning of onset of symptoms. In our hospital, most patients will have had vitals taken, blood work drawn and resulted, and likely have had basic imaging such as an x-ray (which is usually non-specific), but often have not yet received a surgical consultation or advanced, definitive imaging (such as a CT). In our study sample, the average time to send a patient to CT was 4 h. The research question was 'at what time does our system have the data it needs to perform well', the answer to which is 4 h. The clinical timeliness of this mark will differ across institutions, but 4 h is early enough that a fair proportion of patients will not have had a definitive diagnosis when the algorithm could provide the most accurate classification. It can also be argued that the system already has a high performance level after 3 h and might already be useful to a physician as a CDS tool, especially to identify definite high-risk patients as high-risk predictions are final from the start. Moreover, in our current culture of wanting to minimize unnecessary radiation, clinicians would likely be willing to wait for a prediction before ordering a CT, since the system is predictive reasonably early.

The automated algorithm has several clinical advantages and applications. First, it can be viewed as a safety and quality tool that can be embedded within an EHR to prompt providers to supply additional information if the algorithm is unable to detect the appropriate data elements to classify the patient. This check would ensure that physicians have considered all of the appropriate signs, symptoms, and laboratory values needed for an accurate assessment. Second, automated appendicitis CDS could function as a second opinion of sorts, automatically risk classifying the patient and presenting the result to a provider to validate and be an adjunctive to their own clinical judgment and gestalt. Third, automated risk stratification could occur 'on the fly' to increase provider efficiency, providing CDS without requiring providers to manually input PAS variable values into an electronic spreadsheet or form.

The performance of the system is based on the manually created gold standard, with the assumption that this standard is reliable. High agreement was reported among the annotators for the risk classification and is an argument in favor of the strength of the gold standard. However, agreement was fair but lower for the annotation of PAS elements in clinical notes, which is a crucial intermediate step towards classification; performance may be affected if the system is trained using a gold standard with modest reliability. Previous studies have also observed that agreement between physicians on patient history and physical examination findings of patients with possible appendicitis was variable.[30 31] This is a limitation not only for an automated system, but also for clinical pathways.

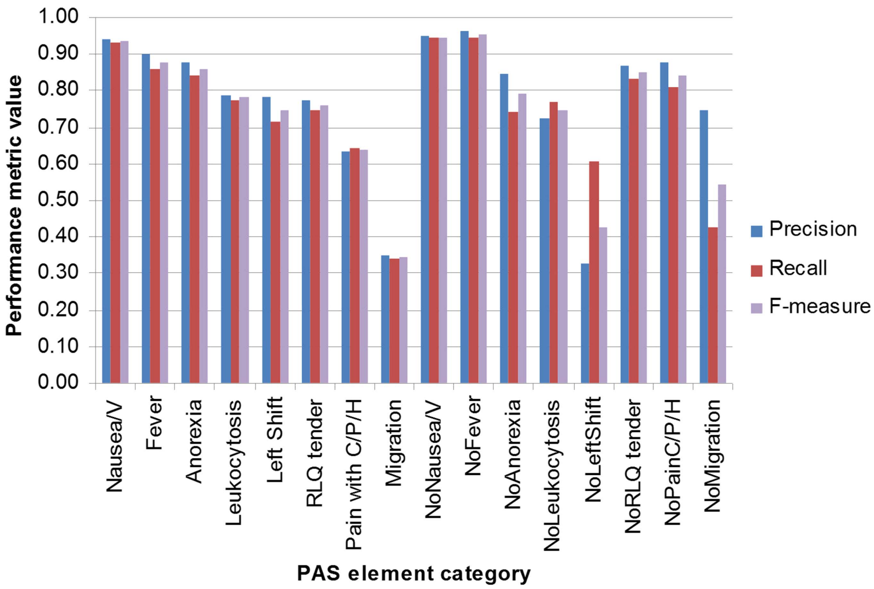


Figure 3 Performance of the detection of Pediatric Appendicitis Score (PAS) elements in clinical notes.

Our approach is centered on a risk stratification schema based on the PAS, and we only reported performance with regard to this schema, which may or may not be the most accurate one. In a future analysis, it would be interesting to use pathology-proven appendicitis as the gold standard of outcome to train and evaluate a similar system to predict acute appendicitis.

This study uses retrospective data and has not yet been evaluated at the point-of-care and thus it is a first step towards implementation as a CDS tool. Future work includes evaluating the approach in a real time setting. We plan to set up a production version of the automated system that can be launched in the ED, after physicians have examined abdominal pain patients, and will predict an appendicitis risk class to help physicians in their decision-making process in real time. Evaluation of such a system is twofold: (1) assessment of the system's performance in making its predictions in real time will be necessary; and (2) evaluation of the value of the system as a decision support tool will be undertaken (answering questions such as, 'Does it function well within ED workflow?'). In the next stage, we will determine if the system actually helped physicians in the diagnosis of appendicitis and in reducing unnecessary imaging tests.

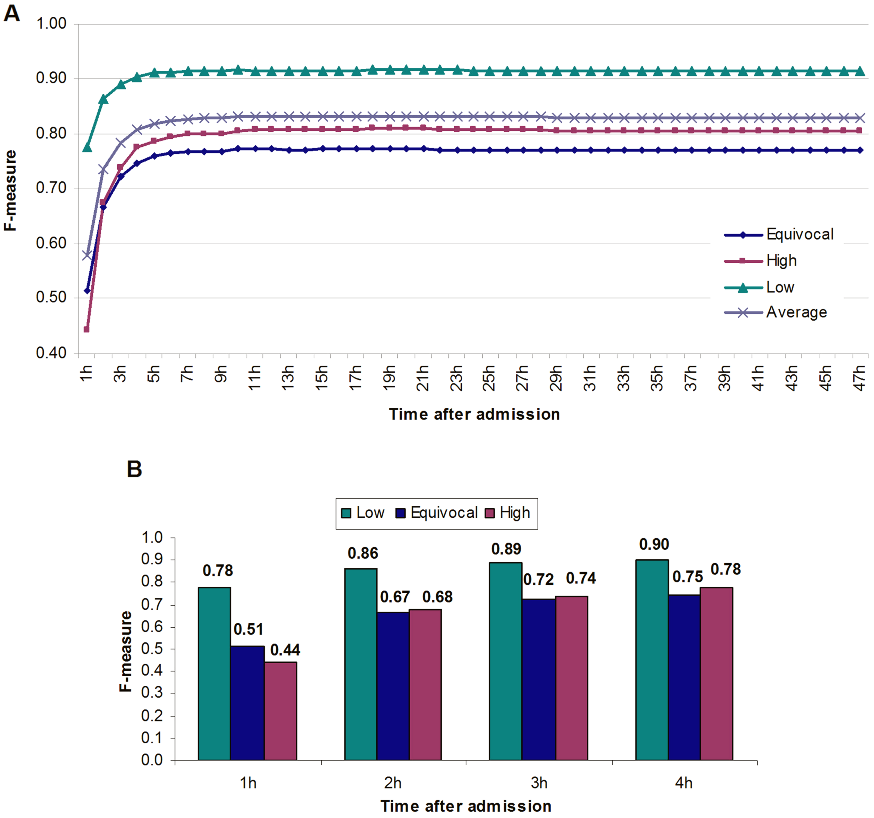


Figure 4 Risk classification performance (F-measure) at different points in time after admission. (A) F-measure over time for the entire 48-hour window; (B) F-measure during the first four hours for each risk class.

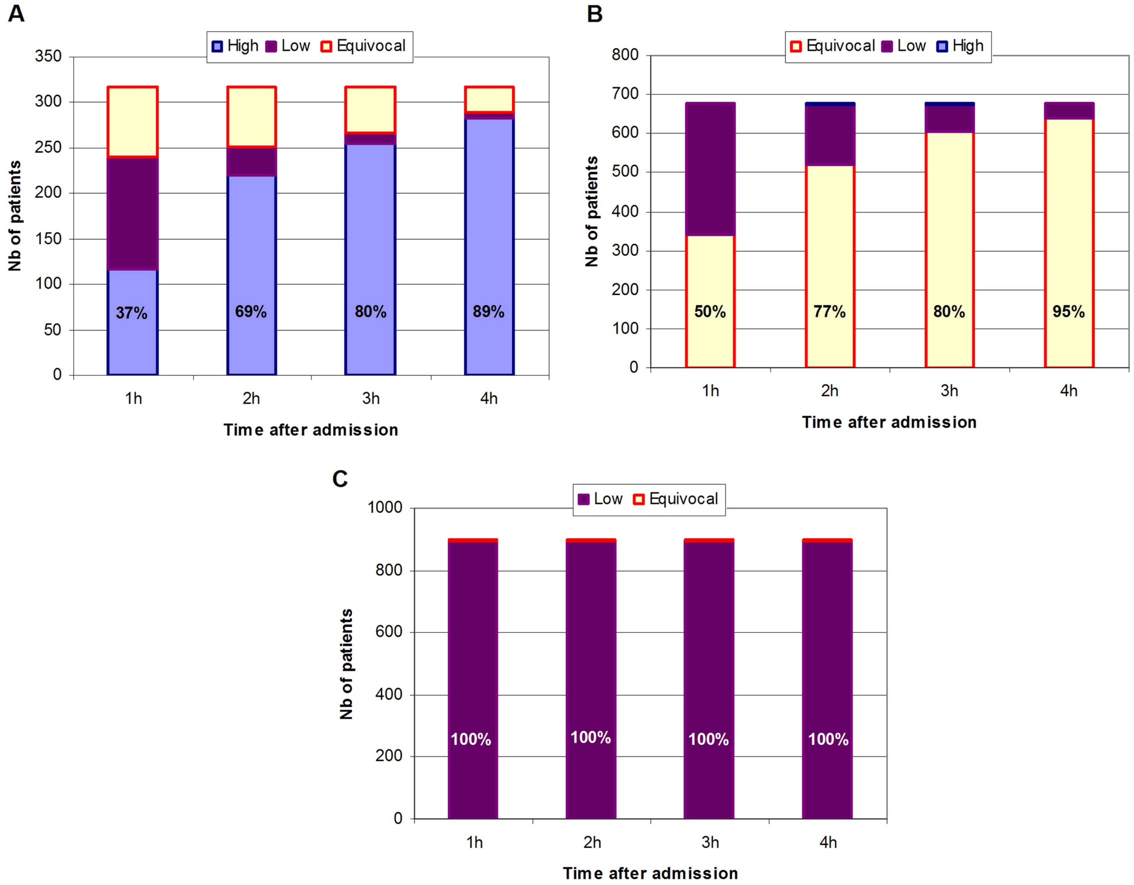


Figure 5 Risk classification over time compared to eventual system predictions for each risk class. (A) Eventual high-risk; (B) eventual equivocal-risk; (C) eventual low-risk.

CONCLUSION

In this work, we investigated automated methods for appendicitis risk stratification of patients with abdominal pain based on the content (both structured entries and text) of the EHR. We proposed a two-step approach by first extracting information from clinical notes based on NLP and then using this information together with the structured data to automatically compute the PAS and assign a risk class (high/equivocal/low) with a rule-based method. We obtained very good performance for the risk classification of patients (average precision and recall of 0.869 and 0.868), comparable to those obtained by physicians performing the same task. Automated appendicitis risk classification based on the content of the EHR, when information from the clinical notes is incorporated, is a promising approach to enhance application of decision support for pediatric abdominal pain ED patients.

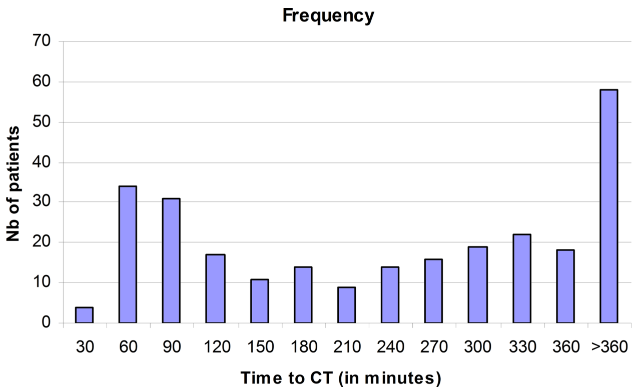


Figure 6 CT time distribution.

Contributors IS and EA conceived the study and obtained research funding. HB provided advice on study design. HB, EK, and LD supervised the creation of the gold standard. LD extracted the data, designed and implemented 90% of the experiments, and analyzed the results. HZ provided statistical advice on the design of the experiments and ran 10% of the experiments. QL and TL contributed to the implementation of the experiments. LD drafted the manuscript and all authors contributed substantially to its revision.

Competing interests The authors and annotators were supported by internal funds from Cincinnati Children's Hospital Medical Center. IS, LD, HZ, QL, and TL were partially supported by grants 5R00LM010227-04, 1R21HD072883-01, and 1U01HG006828-01.

Ethics approval Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Bhuiya FA, Pitts SR, McCaig LF. Emergency department visits for chest pain and abdominal pain: United States, 1999–2008. *NCHS Data Brief* 2010;43:1–8.

2 Buckius MT, McGrath B, Monk J, *et al*. Changing epidemiology of acute appendicitis in the United States: study period 1993–2008. *J Surg Res* 2012;175:185–90.

3 Fahimi J, Herring A, Harries A, *et al*. Computed tomography use among children presenting to emergency departments with abdominal pain. *Pediatrics* 2012;130:e1069–75.

4 Brenner D, Elliston C, Hall E, *et al*. Estimated risks of radiation-induced fatal cancer from pediatric CT. *Am J Roentgenol* 2001;176:289–96.

5 Pierce DA, Shimizu Y, Preston DL, *et al*. Studies of the mortality of atomic bomb survivors. Report 12, Part I. Cancer: 1950–1990. *Radiat Res* 1996;146:1–27.

6 Pines JM. Trends in the rates of radiography use and important diagnoses in emergency department patients with abdominal pain. *Med Care* 2009;47:782–6.

7 Samuel M. Pediatric Appendicitis Score. *J Pediatr Surg* 2002;37:877–81.

8 Zuniga RV, Arribas JL, Montes SP, *et al*. Application of Pediatric Appendicitis Score on the emergency department of a secondary level hospital. *Pediatr Emerg Care* 2012;28:489–92.

9 Goldman RD, Carter S, Stephens D, *et al*. Prospective validation of the pediatric appendicitis score. *J Pediatr* 2008;153:278–82.

10 Bhatt M, Joseph L, Ducharme FM, *et al*. Prospective validation of the pediatric appendicitis score in a Canadian pediatric emergency department. *Acad Emerg Med* 2009;16:591–6.

11 Escriba A, Gamell AM, Fernandez Y, *et al*. Prospective validation of two systems of classification for the diagnosis of acute appendicitis. *Pediatr Emerg Care* 2011;27:165–9.

12 Kulik DM, Uleryk EM, Maguire JL. Does this child have appendicitis? A systematic review of clinical prediction rules for children with acute abdominal pain. *J Clin Epidemiol* 2013;66:95–104.

13 Chaudhry B, Wang J, Wu S, *et al*. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Ann Intern Med.* 2006;144:742–52.

14 Friedman C, Knirsch C, Shagina L, *et al*. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. *AMIA Annu Symp Proc* 1999:256–60.

15 Aronsky D, Fiszman M, Chapman WW, *et al*. Combining decision support methodologies to diagnose pneumonia. *AMIA Annu Symp Proc* 2001:12–16.

16 Pakhomov SV, Buntrock J, Chute CG. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. *J Biomed Inform* 2005;38:145–53.

17 Cheng LT, Zheng J, Savova GK, *et al*. Discerning tumor status from unstructured MRI reports—completeness of information in existing reports and utility of automated natural language processing. *J Digit Imaging* 2010;23:119–32.

18 Wagholikar KB, MacLaughlin KL, Henry MR, *et al*. Clinical decision support with automated text processing for cervical cancer screening. *J Am Med Inform Assoc* 2012;19:833–9.

19 Chapman WW, Dowling JN, Wagner MM. Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients. *Ann Emerg Med* 2005;46:445–55.

20 Grouin C, Deléger L, Rosier A, *et al*. Automatic computation of CHA2DS2-VASc score: information extraction from clinical texts for thromboembolism risk assessment. *AMIA Annu Symp Proc* 2011;2011:501–10.

21 Synnergren J, Olsson B, Gamalielsson J. Classification of information fusion methods in systems biology. *In Silico Biol* 2009;9:65–76.

22 Boström H, Andler SF, Brohede M, *et al*. On the definition of information fusion as a field of research. Technical report at the University of Skövde, HS-IKI-TR-07-006. 2007.

23 Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning. Morgan Kaufmann. 2001:282–9.

24 McCallum A. *MALLET: a machine learning for language toolkit*. 2002. http://mallet.cs.umass.edu/ (last accessed on 10/10/2013)

25 Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.

26 UMLS. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

27 Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. *Methods Inf Med.* 1998;37:334–44.

28 Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc.* 2005;12:296–8.

29 Noreen EW. Computer-intensive methods for testing hypotheses: an introduction. New-York: Wiley, 1989.

30 Kharbanda AB, Stevenson MD, Macias CG,, *et al* Interrater reliability of clinical findings in children with possible appendicitis. *Pediatrics.* 2012;129:695–700.

31 Kharbanda AB, Fishman SJ, Bachur RG. Comparison of pediatric emergency physicians' and surgeons' evaluation and diagnosis of appendicitis. *Acad Emerg Med* 2008;15:119–25.