Automating a Severity Score Guideline for Community-Acquired Pneumonia Employing Medical Language Processing of Discharge Summaries

Carol Friedman, Ph.D.[1,2], Charles Knirsch, M.D. [3,2]
Lyudmila Shagina, MA., [2]George Hripcsak, M.D.[2]

[1]Department of Computer Science, Queens College CUNY
[2]Department of Medical Informatics, Columbia University
[3] U.S. Pharmaceuticals, Pfizer Inc.

Obtaining encoded variables is often a key obstacle to automating clinical guidelines. Frequently the pertinent information occurs as text in patient reports, but text is inadequate for the task. This paper describes a retrospective study that automates determination of severity classes for patients with community-acquired pneumonia (i.e. classifies patients into risk classes 1-5), a common and costly clinical problem. Most of the variables for the automated application were obtained by writing queries based on output generated by MedLEE[1], a natural language processor that encodes clinical information in text. Comorbidities, vital signs, and symptoms from discharge summaries as well as information from chest x-ray reports were used. The results were very good because when compared with a reference standard obtained manually by an independent expert, the automated application demonstrated an accuracy, sensitivity, and specificity of 93%, 92%, and 93% respectively for processing discharge summaries, and 96%, 87%, and 98% respectively for chest x-rays. The accuracy for vital sign values was 85%, and the accuracy for determining the exact risk class was 80%. The remaining 20% that did not match exactly differed by only one class.

INTRODUCTION

The electronic patient record contains a vast amount of patient information that is accessible electronically. This information would be invaluable if it were available for automated applications, such as decision support or outcomes analysis, but those applications require encoded data. However, a large portion of the information in the patient record is not encoded because it occurs as text. Medical language processing (MLP) techniques have been developed[1-5] to extract and encode clinical information in text.

Several studies of MLP systems were performed in the specialized domain of radiological reports that demonstrated satisfactory performance[4,6-9]. An important source of clinical information occurs in broader clinical domains, such as discharge summaries, and admission notes. These reports are much more challenging for language processing systems because they contain complex sentences and relationships, and encompass a very large vocabulary.

In this paper, we present a feasibility study pertaining to an MLP extraction and encoding system, called MedLEE[1], for an application associated with community acquired pneumonia (CAP). A key reason this application was selected was because it has the potential to impact patient care. Another reason was that it used variables obtained by processing the *History, Chief Complaint,* and *Physical Examination* sections of text reports, and therefore was well-suited for evaluating performance of an MLP system in a broad domain. For this study, discharge summaries were used; for a real clinical decision support CAP application, emergency room (ER) reports or office visit notes would be required. However, these types of reports are not currently available in electronic form at New York Presbyterian Hospital (NYPH), and therefore the discharge summary was used as substitute for the sake of estimating performance.

Community-acquired pneumonia (CAP) is a common clinical problem having a significant affect on health care. Over 4 million cases are diagnosed each year in the United States, more than 600,000 patients a year are hospitalized, and the total cost for inpatient care approaches $4 billion per year[10,11]. An application that automatically and accurately assigns risk categories to patients could help physicians identify low-risk patients and help lower hospitalization rates for pneumonia. It could also provide objective measures for treatment.

BACKGROUND

The CIS system at NYPH provides the infrastructure for this study. It consists of a clinical repository that includes a combination of textual and structured encoded information. Laboratory test findings and cer-

1091—8280/99/$5.00

tain demographic information are stored in the repository in encoded form whereas radiology reports and discharge summaries are stored in textual form.

MedLEE was integrated with the CIS system at NYPH and is used routinely to encode information in radiological reports of the chest (CXR). Studies of MedLEE and other MLP systems[4,6-9] measured performance for applications associated with radiology reports, a limited domain. Other evaluations of MLP systems[12,13] were reported for applications associated with admission diagnoses and discharge summaries, both broad domains, but these were not concerned with automation of guidelines.

Hospital admissions for pneumonia vary markedly from region to region and physicians depend on subjective impressions in order to decide whether or not to hospitalize a patient. Guidelines to help physicians manage patients with CAP have been developed[10] that require manual entry of a substantial amount of data. A study related to automating a guideline for pneumonia was undertaken[14], but the study was not concerned with MLP issues.

In this application we automate a prediction rule for the prognosis of CAP that was developed by Fine[10]. The rule assigns patients with CAP to one of five risk categories so that they could be treated according to their risk class. In addition, we also collected other variables from the discharge summary and chest x-ray that were relevant to CAP. In all, a total of 18 variables were collected using MedLEE.

The CAP prediction rule is a two step process. The first step determines whether a patient with CAP belongs to risk class I (lowest-risk). The following conditions must all be true:
i). Age is less than 50, ii) No history of neoplastic disease, congestive heart failure, cerebrovascular disease, renal disease or liver disease, and iii). The physical examination findings do not include any of the following abnormalities: altered mental status, pulse ≥ 125/minute, respiratory rate ≥ 30/minute, systolic blood pressure < 90 mm Hg and temperature < 35° C or > 40 ° C.

If a patient does not belong to class I, step 2 is followed to assign the patient to a class II-V based on a score. The scoring system is shown in Table 1. It lists the type of information, the number of points contributing to the score, and the source of the data used for the present study. A patient is assigned a risk class based on the following scores: class II (< 71), class III (71-90), class IV (91-130), and class V (> 130).

Data	Points	Srce	Data	Points	Srce
male	age	C	ab sys bp	20	D
female	age-10	C	ab temp.	15	D
nurse h.	10	D	ab pulse	10	D
neoplastic	30	D	ab resp. r	20	D
liver dis.	20	D	ab bun	20	C
chf	10	D	ab sod.	20	C
crbrv. dis.	10	D	ab gluc.	10	C
renal dis.	10	D	ab hemat	10	C
ch. m. st.	20	D,X	ab art pH	30	C
pleural eff	10	X	ab ppO	10	C

Table 1. Scores for assignment to risk classes II-V. The source of the data for this study is also shown: C represents the CIS and consists of encoded data. The other two data sources, D and X, represent data obtained from textual reports using MedLEE. D represents discharge summary, and X represents chest x-ray. The abbreviation **ab** symbolizes *abnormal value.*

METHODS

Cases were selected by automatically identifying patients who were discharged during a certain year and who were assigned an ICD9 diagnostic code of 486 (typically given for CAP). The first 100 patients from that group were chosen who had an admission chest x-ray. Discharge summaries were then obtained for those patients. Cases were eliminated from the study if there was no discharge summary or if the patient was HIV positive, leaving 79 cases. Demographic information needed for scoring was obtained from the CIS along with the necessary laboratory test findings.

A medical expert (who had no knowledge of the MLP system) was enlisted to read the reports in the test set and manually establish the reference standard. A previous study[15] demonstrated that one expert is sufficient for aggregate measures. We measured the expert's performance with respect to identifying 6 CXR conditions from a previous study, and found his performance above average compared to the 12 physicians who participated in that study.

In addition to the variables shown in Table 1 (needed for risk assessment), 6 other variables associated with CAP (cough, dyspnea, sputum production, fever, rule out pneumonia, pneumonia) were also collected. The expert was given instructions describing the source and criteria for extracting the variables. For example, numeric values for vital signs were to be obtained from the *Physical Examination* or *History of Present Illness* sections of the discharge summary; if the value of a vital sign was missing an **A** was to be assigned. Values for the other variables were **Y** (if positively asserted), **N** (if negated as in *no liver disease*), or **A** (if the variable was not mentioned). Values for comorbidities, such as **neoplastic disease** were to be obtained from the *Past Medical History*, *History of*

Present Illness, and *Review of Systems* sections of the discharge summary. Values for symptoms such as **cough** and **dyspnea** were collected as two distinct groups; one group was obtained from the *Clinical Information* section of the chest x-ray and the second from the *Chief Complaint* or *History of Present Illness* sections of the discharge summary.

Since the instructions were complex, the expert was first shown a few reports from the training set and asked to read the reports and determine the values for the variables according to the instructions. This was done to ensure that the task was clear and that ambiguities were clarified beforehand. The expert was then given the CXR and discharge summary reports from the test set and extracted the required information manually.

The chest x-ray and discharge summary reports in the test set were processed using MedLEE. A version of MedLEE was used that was developed prior to the start of work on this application so that the performance could be evaluated prior to any specialized training for this application. MedLEE was previously extended to the domain of discharge summaries[16] but the extension was not evaluated.

Although MedLEE itself was not modified for this application, queries had to be developed in order to retrieve the appropriate values for the variables from the output generated by MedLEE. In order to train the queries, a set of 40 cases were collected for patients admitted in a different year using the same criteria used for collecting the validation set. A second medical expert (also independent from the developers) helped configure the queries for the variables based on the output MedLEE generated.

Queries associated with the vital signs were straightforward to write because they just involved retrieving the values. The queries corresponding to the co-morbidities and symptoms were more complex; some corresponded to broad categories that could be asserted or negated many different ways. For example, for *change in mental status*, target output terms such as **confusion**, **lethargy**, **comatose**, and **decreased mental status** would also be applicable. These type of queries were developed by showing the expert a list of target output terms generated by MedLEE and having him choose the terms that were associated with each of the co-morbidities and symptoms. In addition, he also looked over the reports in the training set and highlighted relevant terms.

RESULTS

A summary of the results comparing the co-morbidities and symptoms automatically obtained from the chest x-ray and discharge summaries with the reference standard are shown in Table 2. The results showing the accuracy for the variables that have numeric values (vital signs, risk category, and scores) are presented in Table 3.

	cxr findings	dsum findings
accuracy	.96 (.94-.97)	.93 (.91-.94)
sensitivity	.87 (.81-.92)	.92 (.89-.95)
specificity	.98 (.96-.99)	.93 (.91-.95)

Table 2. Performance measures and 95% confidence intervals comparing findings (associated with co-morbidities and symptoms) obtained automatically by processing chest x-ray reports and discharge summaries using MedLEE against a reference standard determined manually by a clinical expert.

The accuracy for the vital signs was 85%. The accuracy corresponding to an exact match of the risk class was 80%, and the accuracy for obtaining a match differing by at most one category was 100%. There was no tendency to consistently overestimate or underestimate the classes because there were roughly the same number of classes that differed by plus and minus one. The accuracy for the exact score was 59%. Matching a score exactly is a very strict criteria. If we relax the criteria and allow for differences of at most 10 points, the accuracy was 81%. The cumulative accuracy was 94% allowing for a difference of at most 20 points

vital sign values	risk class (exact class)	risk class (differ by 1)	exact score
.85 (.81-.89)	.80(.69-.88)	1 (.96-1)	.59 (.47-.70)

Table 3. Accuracy measures comparing data obtained using MLP with values from the reference standard. The values consisted of vital signs, exact match of risk class category, risk categories differing by at most 1 category, and match of exact score.

DISCUSSION

The automated system performed very well (average accuracy, sensitivity, and specificity were 93%, 92%, and 93%) in identifying co-morbidities and symptoms in discharge summaries. The accuracy and specificity were only slightly lower than those found for chest x-rays. This is noteworthy considering that discharge summaries are much more complex than radiological reports because discharge summaries contain many more types of information and the sentence structures are typically longer and more complicated. The automated system also performed well in retrieving vital signs from discharge summaries. The accuracy in computing an exact risk class was somewhat lower

(80%), but still quite reasonable. All 20% of cases that did not have exactly the same class differed by at most 1 class all of the time. These results are very encouraging and demonstrate that is feasible to use MLP to automate the computation of severity classes for patients with CAP, providing that the appropriate clinical reports are available in a timely fashion.

An analysis of the errors revealed that they could be grouped into four broad categories: MedLEE parsing errors, query errors, errors in the reference standard, and errors due to conflicting information in the discharge summaries. It should be noted that the analysis of errors was performed by the system developer.

Parsing errors were due to incorrect parses and to information not yet captured by the system. For example, several errors occurred because MedLEE did not adequately capture family information. This caused co-morbidities to be attributed to patients instead of family members. Another cause of error stemmed from incorrect handling of certain temporal expressions. Because the variety of temporal expressions is very large, it will require a substantial amount of effort to comprehensively address this problem. Fortunately the performance was satisfactory in spite of this problem.

Query errors were mainly due to terms missing from the query. For example **difficulty breathing** was missing from the query testing for shortness of breath. However, although correcting these errors by adding additional terms will be easy, they may cause other errors in future studies. In some cases the errors were attributable to errors in logic. An example of this type of error occurred in the query for **fever** because fever could have been inferred from the output generated by MedLEE (i.e. the term **temperature** and a value indicative of fever was in the output) using a simple rule specifying the appropriate upper and lower bounds for fever.

Another source of error was the reference standard itself. One cause for discrepancy appeared to stem from physician disagreement. In the current study one expert determined the reference standard while a second one determined the criteria for the variables. In a number of cases, the automated system adhered to the criteria established by the second expert but the expert who determined the reference standard differed. For example, in one case the automated system found that **changes in mental status** was present because the discharge summary stated that the patient had *seizures* (a condition for **changes in mental status** according to the criteria). However, the reference standard recorded that **changes in mental status** was not mentioned in the report. Other errors attributable to the reference standard were due to human error on the part of the expert. For example, there were a few cases where the expert missed negations.

Other errors were due to inconsistencies in the discharge summary itself. Sometimes a report asserted that the patient had a condition in one section but negated the condition in another; these discrepancies were not resolved in a definitive way.

Future studies concerning the reference standard should be addressed by enlisting more experts to read the reports. Such a study will provide us with a more powerful evaluation of the automated system. An interesting aspect of such a study will be the measure of inter-rater error for this particular task. The error rate may be higher than ascertained in previous studies because the manual extraction task appears to be more complex. In the previous studies, the subjects were asked to read chest x-ray reports and provide *Yes/No* answers to a list of 6-9 variables. In the current study, the expert had to choose an answer (*Yes /Negated /Absent*) for 14 of the variables and supply a numeric value for another 4.

Another reason this task appears more involved is that more detailed instructions had to be followed to determine values for the variables. For example, the instructions specified which sections should be looked at (different sections were used for different types of information), and also specified rules for choosing a single value if a variable was mentioned multiple times in a report. For example, vital signs often occurred in the *Physical examination* section (sometimes more than once), in *History of Present Illness*, and in the *Hospital Course*. The reason it was important to specify particular sections was because we wanted to use sections that closely corresponded to the information that would be available in a real application. In a real automated assessment system for CAP, information would most likely be obtained from the admission *Physical examination*, *History*, and *Chief Complaint*, but information from the *Hospital Course* would not be appropriate.

Another factor concerning manual extraction performance may involve properties of the discharge summaries in contrast to radiology reports. The lengths of discharge summaries are considerably longer, and discharge summaries contain more comphrensive clinical information than radiology reports.

While this study demonstrated that MedLEE performed very well in extracting information from discharge summaries, and that the automated system was

effective for the CAP application, more evaluations have to be performed before we employ a practical clinical application. In addition, we cannot make claims about the effectiveness of the methodology. Although the discharge summaries were selected at random and the application was chosen by an independent clinician for clinical purposes, there may be something about the application that makes the automated extraction task easy regardless of the methodology. It would be interesting to measure the performance of a keyword search technique for this application.

This study had several limitations. One limitation was that one expert was used to determine the reference standard. A second limitation was that the developer analyzed the causes of error. A third limitation was that inpatient reports were used. Since inpatients have more complications than outpatients do, different results may have been obtained using outpatient records.

CONCLUSION

This study demonstrated that it is feasible to automate determination of risk classes for patients with CAP by using natural language processing of patient reports, particularly information from the history, physical exam, chief complaint, and chest x-ray. The performance associated with the processing of discharge summaries were very good (accuracy, sensitivity, and specificity of 93%, 92%, 93%), in spite of the broadness and complexity of the domain. Because our study used one clinical expert to determine the reference standard, further studies would be desirable using several experts in order to establish a better reference standard and assess inter-rater error for this task.

Acknowledgements

This publication was supported in part by grants LM0624 and LM05627 from the National Library of Medicine. We would like to thank Alan Gerstel for his participation in this study.

References

1. Friedman C, Alderson P, Austin J, Cimino J, Johnson S. A general natural language text processor for clinical radiology. JAMIA 1994;1(2) 161-174.
2. Sager N, Lyman M, Buchnall C, Nhan N, Tick L. Natural language processing and the representation of clinical data. JAMIA 1994;1(2) 142-160.
3. Baud R, Rassinoux A, Scherrer J. Natural language processing and semantical representation of medical texts. Meth Inform Med 1992;31(2) 117-125.
4. Haug P, Ranum D, Frederick P. Computerized extraction of coded findings from free-text radiologic reports. Radiology 1990;174 543-548.
5. Zweigenbaum P, Bouaud B, Bachimont J et al. From text to knowledge: a unifying document-oriented view of analyzed medical language. Meth Inform Med 1998;311-575.
6. Hripcsak G, Friedman C, Alderson P, DuMouchel W, Johnson S, Clayton P. Unlocking clinical data from narrative reports. Ann of Int Med 1995;122(9) 681-688.
7. Zingmond D, Lenert L. Monitoring free-text data using medical language processing. Comp Biomed Res 1993;26 467-481.
8. Jain N and Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In Masys D.ed., *Proceedings of the Fall 1997 AMIA Annual Symposium.* Phil. Hanley & Belfus. 1997; 829-833.
9. Knirsch C, Jain N, Pablos-Mendez A, Friedman C, Hripcsak G. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated decision support system. Infection Control and Hospital Epidemiology 1998;19(2) 94-100.
10. Fine M, Auble T, and Yealy D. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med 1997;336:243-250.
11. Medicare and Medicaid statistical supplement. 1995; Health Care Finance.
12. Gundersen M, Haug P, Pryor T, et al. Development and evaluation of a computerized admission diagnoses encoding system. Comput Biomed Res 1996;29:351-72.
13. 13 Sager N, Lyman M, Nhan NT, Tick L, Borst F, and Scherrer J. Clinical knowledge bases from natural language patient documents. In Lun KE, ed. *MEDINFO 92*;1375-1389.
14. Aronsky D and Haug P. Diagnosing community acquired pneumonia with a bayesian network.In Chute C.ed., *Proceedings AMIA 98 Annual Symposium.* Philadelphia. Hanley&Belfus. 1998; 632-636.
15. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. JAMIA 1999;6(2) 143-150.
16. Friedman C. Towards a comprehensive medical language processing system: methods and issues.In Masys D.ed., *Proceedings of the Fall 1997 AMIA Conference.* Phil. Hanley & Belfus. 1997; 595-599.