CONTRIBUTION OF A SPEECH RECOGNITION SYSTEM TO A COMPUTERIZED PNEUMONIA GUIDELINE IN THE EMERGENCY DEPARTMENT

WENDY W. CHAPMAN, DOMINIK ARONSKY MD, MARCELO FISZMAN MD, PETER J. HAUG MD
DEPT. OF MEDICAL INFORMATICS, LDS HOSPITAL/UNIVERSITY OF UTAH, SALT LAKE CITY, UTAH

ABSTRACT

Objective: Evaluate the effect of a radiology speech recognition system on a real-time computerized guideline in the emergency department.

Methods: We collected all chest x-ray reports (n=727) generated for patients in the emergency department during a six-week period. We divided the concurrently generated reports into those generated with speech recognition and those generated by traditional dictation. We compared the two sets of reports for availability during the patient’s emergency department encounter and for readability.

Results: Reports generated by speech recognition were available seven times more often during the patients’ encounters than reports generated by traditional dictation. Using speech recognition reduced the turnover time of reports from 12 hours 33 minutes to 2 hours 13 minutes. Readability scores were identical for both kinds of reports.

Conclusion: Using speech recognition to generate chest x-ray reports reduces turnover time so reports are available while patients are in the emergency department.

INTRODUCTION

Although clinical guidelines have been produced for a variety of diseases, implementing those guidelines is difficult. Barriers to guideline implementation include antagonism from physicians and time required to complete guidelines [1]. Computerizing guidelines can decrease the time required to complete a guideline by collecting the data automatically and using that data to drive the guidelines [2]. However, problems still exist with computerizing guidelines. Often the clinically relevant data available in hospital information systems is not sufficient or not represented in a computable (i.e., coded) format.

A pneumonia guideline implemented in the emergency department of LDS Hospital in Salt Lake City, Utah is data driven and doesn’t require additional data entry by physicians [3]. Some of the data elements required by the guideline are contained in chest x-ray reports. Gathering information from the chest x-ray report is difficult for two reasons: 1) Reports are usually not available during a patient’s encounter in the emergency room; and 2) Reports available during a patient’s encounter are in free text format and cannot be easily used by an automated guideline.

Researchers have addressed problems of report timeliness and of coding free text reports. Attempts to increase timeliness of radiology reports include requiring radiologists to generate structured reports, accelerating transcription, allowing radiologists to sign reports from home, and implementing a speech recognition system [4]. Methods for coding reports include structured reporting, keyword searches, and natural language processing (NLP) [5].

In this paper we examine whether implementation of a radiology speech recognition system contributes to computerized implementation of a guideline in an emergency department. We measured differences between speech recognition (SR) and traditional dictation (TD) reports relating to turnaround time of reports and to aspects of writing style that could affect how accurately the reports are parsed by an NLP system.

BACKGROUND

Speech recognition systems have been tried in radiology practices for over a decade. Large vocabulary continuous speech recognition systems allow radiologists to speak naturally and are currently accurate enough to be successfully implemented in a radiology department [6]. Figure 1 illustrates the typical report generation process that occurs when an emergency department patient is referred to the radiology department for a chest x-ray. SR systems eliminate the final two steps of the process by recording the text of the dictation as the radiologist speaks and by allowing the radiologist to finalize the report immediately after dictating.

Eliminating the need for transcriptions lowers costs. LDS Hospital spends approximately $21,000 per month on transcription. Eliminating the need for the transcription process also potentially improves the timeliness of capturing a report on a hospital information system (HIS).
In September 1999 LDS Hospital implemented a SR system called Dragon Naturally Speaking (Voice Activated Technologies Systems, Inc., Santa Rosa, CA). The system was made available to the entire radiology department but was not mandatory. To assess the SR system's potential effect on a real-time computerized guideline in the emergency department we compared reports generated by speech recognition against reports generated by traditional dictation.

METHODS

We collected all chest x-ray reports (n=727) dictated for the emergency department during a six-week period (Jan. 8 – Feb. 18, 2000). Radiologists were not required to use the SR system, so we measured the proportion of reports dictated during the study period by speech recognition and by traditional dictation. We compared SR and TD reports for timeliness and for differences in writing style that could affect our NLP system's ability to accurately parse the reports.

Timeliness

To evaluate the timeliness of each set of reports we examined turnover time and availability of the reports during patients’ encounters in the emergency department. Turnover time was defined as the number of hours from the order of an x-ray to the report’s availability on the HELP hospital information system [7]. Because reports are available on the system before the radiologist finalizes the report, we used the time the report was transcribed as the time of report completion. From the perspective of a real-time emergency department guideline, turnover time is only relevant if it falls within the time of the patient encounter. We measured the proportion of SR and TD reports that were on the HIS during the patient’s stay in the emergency department.

Readability

NLP systems are vulnerable to linguistic complexity and perform less accurately on reports that are ambiguous [8]. One characteristic of ambiguous reports is long sentences. To evaluate potential differences in dictation style that might affect our NLP system's ability to accurately parse reports we measured the readability of both sets of reports. Readability is a measure of linguistic complexity that assigns a score to a document based on the length of the sentences and the length of the words in the sentences. Studies in readability of health related texts have focused on printed and web-based documentation for patient education [9,10]. A recent study compared the readability scores of journal articles accepted to *Annals of Internal Medicine* before and after the review process [11]. The readability of different types of radiology reports has also been analyzed by Sierra, et al [12].

Two common and well-validated readability formulas are those of Flesch [13] and Gunning [14]. Both formulas are based on word and sentence length. The Flesch Reading Ease Index is a 100-point scale in which 100 represents the easiest texts and 0 the most difficult. The Flesch-Kinkaid Grade Level translates the reading ease score into the U.S. grade level required by the reader in order to easily read and comprehend the document. Another grade-level rating system is the Gunning Fog Index that emphasizes word length as a measure of the document’s "fog level.” Table 1 shows the relationship among the indexes. The Gunning Fog Index is usually higher than the Flesch-Kinkaid.

<table>
<thead>
<tr>
<th>Table 1. Interpretation of Readability Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Flesch Reading Ease Index</strong></td>
</tr>
<tr>
<td>90-100</td>
</tr>
<tr>
<td>80-89</td>
</tr>
<tr>
<td>70-79</td>
</tr>
<tr>
<td>60-69</td>
</tr>
<tr>
<td>50-59</td>
</tr>
<tr>
<td>30-49</td>
</tr>
<tr>
<td>0-29</td>
</tr>
</tbody>
</table>

*Adapted from [11] and [15]*

We removed all headers and footers from the 727 reports then measured descriptive statistics and readability scores. We used t-tests to compare SR and TD reports for statistical difference (α=.05).
RESULTS

Within the 40-day study period 727 chest x-ray reports were dictated for patients in the emergency department. Of the 727 reports concurrently generated in the study period 55% (400/727) were generated by traditional dictation, and the remaining 45% (327/727) were generated by speech recognition. As shown in Figure 2, six of the 17 radiologists who dictated reports for the emergency department during the study period used speech recognition for more than 60% of their reports.

![Image of bar chart showing percent of reports dictated with VR]

Figure 2. Percentage of reports dictated with speech recognition (SR). Radiologists are ordered from the left by the number of reports they generated for the emergency department during the study period, ranging from n=92 (Rad 1) to n=ten (Rad 17).

Timeliness

The average turnover time for SR reports was 2 hours 13 minutes (CI: 1:44-2:42) compared to 12 hours 33 minutes (CI: 11:28-13:37) for TD reports. During the study period the average length of stay for emergency department patients who received a chest x-ray was 3.5 hours. No differences between one-view and two-view reports existed for SR or TD. Of the SR reports 64% (210/327) were available during the patient's emergency room encounter, whereas only 8% (31/400) of the TD reports were. Histograms of turnover time for SR and TD reports are shown in Figure 3. A histogram of study patients’ length of stay is in the background of the figure. Length of stay was identical for patients whose reports were generated by SR or by TD.

To compare the timeliness of radiologists who used SR to those who used TD we separated the radiologists by what dictation method they used the majority (>60%) of the time. Figure 4 presents the timeliness of the two groups of radiologists.

![Image of histograms showing timeliness of radiologists]

Figure 3. Histogram of report turnover times. The white bar histogram represents the length of stay for emergency department patients for whom chest x-rays were ordered during the study period.

Figure 4. Timeliness of individual radiologists. (a) represents radiologists who primarily used speech recognition; (b) represents radiologists who primarily used conventional dictation.
Percentages of timely SR reports for radiologists who used SR ranged from 58% (53/81) to 76% (38/50). Timeliness of TD reports for radiologists who used TD ranged from 0% (0/10) to 31% (9/29).

Readability
The median length of all chest x-ray reports generated during the study period was 368 words. The reports averaged 50 words per sentence, and 1.75 syllables per word. Table 2 contains readability scores for SR and TD reports. No statistically significant differences (α = .05) existed between SR and TD reports on any of the readability measures.

Table 2—Characteristics of reports generated by speech recognition and traditional dictation

<table>
<thead>
<tr>
<th></th>
<th>Speech Recognition (SR)</th>
<th>Traditional Dictation (TD)</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesch Reading Ease Index</td>
<td>51.7 (50.2-53.2)</td>
<td>52 (50.5-53.6)</td>
<td>0.91</td>
</tr>
<tr>
<td>Flesch-Kinkaid Grade Level</td>
<td>7.6 (7.4-7.9)</td>
<td>7.57 (7.3-7.8)</td>
<td>0.80</td>
</tr>
<tr>
<td>Gunning Fog Index</td>
<td>11.1 (10.8-11.4)</td>
<td>10.7 (10.4-11)</td>
<td>0.08</td>
</tr>
<tr>
<td>Syllables/word</td>
<td>1.75 (1.7-1.8)</td>
<td>1.76 (1.7-1.8)</td>
<td>0.40</td>
</tr>
<tr>
<td>Words/sentence</td>
<td>51.9 (48.8-54.9)</td>
<td>49.3 (46.6-52)</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Numbers in table are means (95% CI).

DISCUSSION
The SR system improved the timeliness of reports without changing the readability of the reports.

Timeliness
Implementing computerized guidelines in the emergency department requires that information from radiology reports be stored on the hospital information system before the patient leaves the emergency department. Using the SR system reduced the turnover time of x-ray reports from about twelve hours to about 2 hours and increased access to reports during patients' encounters from 8% (31/400) with TD to 64% (210/327). Despite the fact that less than half the reports were generated with the SR system, almost seven times more SR reports were available during the patients' encounters than TD reports.

Reports dictated by radiologists who mainly used SR were timely the majority of the time, whereas reports dictated by radiologists who used TD were almost never timely. Our data, however, do not rule out the possibility that radiologists using the SR system are more timely regardless of the method of dictation. However, radiologists who primarily used SR were only on time with reports they generated by traditional dictation 12% (5/41) of the time.

Readability
Readability scores did not differ for SR and TD reports. Reports dictated by the same radiologist using SR and TD were not statistically different. Thus, the SR system does not limit the radiologists' ability to speak in the same manner as when using traditional dictation. We had hypothesized that reading one's own reports during or immediately after dictation, as occurs with the SR system, might affect the dictation style of the radiologists. With respect to readability of reports our data do not support this hypothesis. A better way to study whether or not a SR system affects readability of reports would be to compare reports from radiologists who primarily used SR before and after implementation of the SR system. But because radiologists who seldom use the SR system did not receive different readability scores than those who primarily use SR, we do not anticipate that we would find a difference between pre- and post-implementation of the SR system.

We have tested our NLP system on reports that caused disagreement among physicians reading the reports and on reports that were unanimously agreed upon [7]. Our NLP system performed significantly worse on the reports that caused disagreement among physicians. Determining whether an SR system changes a radiologist's dictation style enough to affect an NLP system's ability to parse the reports requires measuring other characteristics of the report besides readability. Readability is a limited measure of document complexity and comprehensibility [16,17] that only examines sentence and word length. Sentence length is a characteristic that affected the performance of the NLP system in our previous study. Other report characteristics affecting parser performance were the amount of redundancy in the report, the amount of information provided about the implied disease in the report, and whether or not the report gave an interpretation of the findings. To identify differences in SR and TD reports that affect our NLP system's performance, we need to evaluate the characteristics that have been shown to decrease parser performance.

Future Work
Various factors can be studied when analyzing a SR system, including the ability of the system to recognize words and concepts and the contribution of
the system to an application. For this paper, we examined the SR system’s potential to contribute to a real-time computerized guideline in the emergency department. Timely access to reports is important for physicians and for computerized guidelines. Even though almost half of the chest x-ray reports gathered in our study period were generated by SR, the majority of the radiologists still use TD. Increasing the proportion of radiologists who utilize SR will improve access to radiology reports in the emergency department. Therefore, future studies should assess the usefulness of the SR system from the perspective of the radiology department with hopes of increasing the number of reports generated by speech recognition.

Conclusion

Chest x-ray reports generated by speech recognition were available sooner than reports generated by traditional dictation. Information contained in chest x-ray reports is crucial to the performance of our computerized pneumonia guideline [18]. The pneumonia guideline had access to the majority of the SR reports while the patient was still in the emergency room. SR reports received the same average readability score as TD reports, implying that parsing the reports into coded data might not be affected by the different dictation method. Implementation of a SR system for generating radiology reports not only saves money but also provides quicker access to reports for use by physicians and computerized guidelines requiring information from the reports.

Acknowledgments

This project was supported by NLM grant 1 R01 LM 06539-02.

References