Reduction of Voice Recognition Errors in Radiological Dictation:
Effects of Systematic Individual Feedback

Rina Patel, Brent Greenberg, Steven Montner, Alexandra Funaki, Christopher Straus, Steven Zangan, and Heber MacMahon

Department of Radiology, University of Chicago Hospitals & Clinics, Chicago, IL

INTRODUCTION

- **Voice Recognition Software in Radiology**
  - Before voice recognition software, radiologists would dictate reports that were then transcribed and sent back to the radiologist for review and finalization.
  - With the advent of voice recognition, turnaround times for final reports have decreased, with associated increases in productivity.
  - A recent study showed improvement in report turnaround time from 28 hours to 12.7 hours, with an associated 5% increase in volume of reports.

- **Errors with Voice Recognition Software**
  - Recognition errors are frequent and require correction by the dictating radiologist.
  - Errors are a result of the technology and can include grammatical, typographic, or other errors.
  - Some errors are minor (e.g., grammatical errors) or easily recognized as an error, while others can be more impactful and require review.

RESULTS

- **Table 1. Scores per attending (percentage of 25) and average score for every 2 month interval.**

<table>
<thead>
<tr>
<th>Radiologist</th>
<th>Before Intervention</th>
<th>After Intervention 1</th>
<th>After Intervention 2</th>
<th>After Intervention 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sept and Oct</td>
<td>Nov and Dec</td>
<td>Jan and Feb</td>
<td>Mar and Apr</td>
</tr>
<tr>
<td>Radiologist 1</td>
<td>93%</td>
<td>94%</td>
<td>94%</td>
<td>93%</td>
</tr>
<tr>
<td>Radiologist 2</td>
<td>85%</td>
<td>91%</td>
<td>94%</td>
<td>89%</td>
</tr>
<tr>
<td>Radiologist 3</td>
<td>65%</td>
<td>75%</td>
<td>92%</td>
<td>89%</td>
</tr>
<tr>
<td>Radiologist 4</td>
<td>97%</td>
<td>98%</td>
<td>98%</td>
<td>95%</td>
</tr>
<tr>
<td>Radiologist 5</td>
<td>87%</td>
<td>95%</td>
<td>92%</td>
<td>94%</td>
</tr>
<tr>
<td>Radiologist 6</td>
<td>95%</td>
<td>100%</td>
<td>98%</td>
<td>100%</td>
</tr>
<tr>
<td>Average</td>
<td>88%</td>
<td>92%</td>
<td>94%</td>
<td>94%</td>
</tr>
</tbody>
</table>

**Figure 1.** Scores per attending (percentage of 25) and average score for every 2 months. Scores are improved from September/October to November/December (after the 1st intervention). Scores plateau after the 2nd and 3rd interventions.

- **Before and After the First Intervention**
  - The scores from after the first intervention (November and December) were significantly improved compared to before the intervention (September and October), 89% to 93%, p-value of 0.02
  - The scores stabilized after the first intervention.
  - No significant improvement was noted between the subsequent interventions.
  - Feedback was provided for specific reports on a daily as needed basis.

PURPOSE

- **Implement a quality improvement initiative in the Chest Section of the University of Chicago Radiology Department to address the frequency of voice recognition errors and reduce the number of errors in the final report.**

MATERIALS AND METHODS

- **Quality Improvement Initiative**
  - The project began with a quarterly review of 10 reports from each attending radiologist.
  - Reports were randomly collected and reviewed by another attending radiologist.
  - The report was scored for frequency of unrecognized voice recognition errors.
  - Results were tabulated and periodically presented and distributed to the faculty.
  - Based on these results, a more intensive feedback program was initiated in November 2010.

- **Data Collection**
  - The project was limited to the Chest Section of the University of Chicago Radiology Department.
  - Reports were collected by a single attending chest radiologist.
  - 20 sequential chest radiology reports and 5 sequential CT reports were collected for each radiologist in the chest section.
  - All of the reports were collected from a randomly selected day.

- **Scoring**
  - The reports were printed and distributed to other members of the Chest Section for review and scoring.
  - Reports were scored each month.
  - A single radiologist reviewed the reports of another radiologist.
  - The radiologists reviewed reports of different individuals each month in order to limit scoring bias.

- **Scoring System**
  - Each report had an initial value of 1 point.
  - Grammatical, typographical, or spelling errors resulted in a deduction of 0.25 points.
  - Insignificant word substitutions by the voice recognition system resulted in a deduction of 0.5 points.
  - An error that was confusing or potentially altered the meaning of the report incurred deduction of 1 full point.
  - No more than 1 point could be deducted per report.
  - Due to small sample size, the scores were aggregated for every two months in order to reduce random variation in scores.

- **Quality Improvement Intervention**
  - Every two months, the dictating radiologist was given his or her reports with the errors marked.
  - Individual error rates and suggestions for improvements, including microphone positioning, use of macros/templates and careful proof reading, were discussed at monthly section meetings.
  - Feedback was provided for specific reports on a daily as needed basis.

- **Analysis**
  - Scores were given as a percentage (of 25) and as an error rate (25 minus the score).
  - Individual and average percentages were charted to demonstrate improvements or worsening.
  - A two-tailed paired t-test was performed between each intervention (i.e., between every 2 month group) and for the first 2 months compared to the last 2 months.

- **Turnaround Time**
  - Turnaround time was based on the length of time between the exam end time and report end time for each case.
  - Turnaround times were averaged for every 2 month interval and compared using a two-tailed paired t-test.

DISCUSSION

- **A significant improvement in scores was noted after the first intervention (in November/December, 4 months after the project began).**
  - Scores stabilized after the first intervention (between 4 and 9 months).
  - Some radiologists had consistently higher scores. These radiologists frequently used different methods of dictation and review.
  - One of the radiologists dictated the initial report, but then used the keypad for reviewing and editing.
  - Subjectively, it appeared that radiologists with fewer words or a "telegraphic" style of dictation had fewer errors.
  - One of the radiologists began using a headset for dictation, and he noticed a subjective decrease in his frequency of errors after using the headset.
  - "A significant finding was that certain words or phrases were noted to be due to limitation in improvement (i.e., those with 100% accuracy are unable to improve)."
  - "We define trend was noted in the types of errors (i.e., whether ‘1 point’ errors were more common before or after the intervention)."
  - The frequency of each error type for three of the radiologists from October (before the intervention) were compared with those from April (after the intervention).
  - Although there appeared to be a trend towards fewer 1 point errors for two of the radiologists, no significant difference was noted among the types of errors from October and April.

- **Examples of Error Types**
  - **1 point:**
    - Confusing or potentially altering the meaning of the report
    - "This is to say that central venous catheter is unchanged." intended sentence: "NG tube, ET tube, and central venous catheter are unchanged."
    - Insignificant word substitution
    - "Asymmetric soft tissue in the right breast (4096) which is nonspecific" intended sentence: "Asymmetric soft tissue in the right breast (4096) which is nonspecific"
    - Grammatical or typographic error
  - **No more than 1 point could be deducted per report.**
  - "Focal nodular right upper lobe opacity as previously characterized by CT scan."
  - "Focal nodular right upper lobe opacity as previously characterized by CT scan."

**References**