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 increase 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[1]
- **Errors with Voice Recognition Software**
  - Recognition errors are frequent and require correction by the dictating radiologist
  - Reported frequency of voice recognition error rates in the radiology literature ranges from 4.8-42%[2-4]
  - Uncorrected errors have potentially serious consequences
    - Some errors are minor (e.g. grammatical errors) or easily recognized as an error
    - Other errors can be confusing or misleading and alter the meaning of the report
  - Whether errors are major or inconsequential, they can be embarrassing evidence of careless proofreading and have potential medico-legal consequences.

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

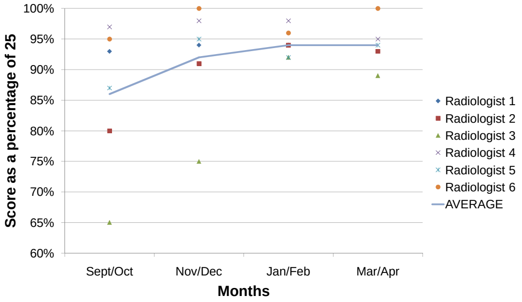
- **Quality Improvement Initiative**
  - The project began with a quarterly review of 10 reports from each attending radiologist
    - The reports were randomly collected and were reviewed by another attending radiologist
    - The report was scored for frequency of unrecognized voice recognition errors.
    - 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 radiograph 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
    - For example, Radiologist 1 reviewed Radiologist 2 for the month of September, but reviewed Radiologist 3 for October
  - 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 worsening or improvement
  - A two tailed paired t-test was reported between each intervention (i.e. between every 2 month group) and for the first 2 months compared to the last 2 months
  - Turnaround time was based on the length of time between the exam end time and report final time for each case; turnaround times were averaged for every 2 month interval and compared using a two tailed paired t-test

RESULTS

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

	Before Intervention	After Intervention 1	After Intervention 2	After Intervention 3
	Sept and Oct	Nov and Dec	Jan and Feb	Mar and Apr
Radiologist				
1	93%	94%	94%	93%
2	80%	91%	94%	93%
3	65%	75%	92%	89%
4	97%	98%	98%	95%
5	87%	95%	92%	94%
6	95%	100%	96%	100%
Average	86%	92%	94%	94%

- **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.



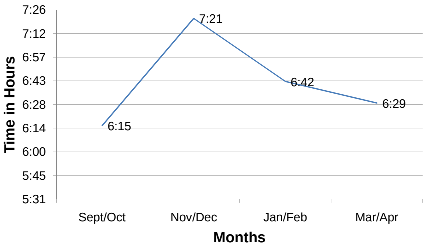
- **Before and After the First Intervention**
  - The scores from after the first intervention (November and December) were significantly improved compared to before the intervention (September and October), 86% to 92%, p-value of 0.02
  - The scores stabilized after the first intervention
  - No significant improvement was noted between the subsequent interventions
    - Nov/Dec to Jan/Feb: 92% to 86%, p value = 0.56
    - Jan/Feb to Mar/Apr: 86% to 94%, p value = 0.40
- **Types of Errors Made**
  - The frequency of each error type for three of the radiologists from October (before the intervention) were compared with those from April(after the 3rd intervention)
  - Although there appeared to be a trend towards fewer 1-point errors for two of the radiologists, no significant difference was noted between the types of errors from October and April
- **Examples of Error Types**
  - 1 point:
    - Confusing or potentially alters the meaning of the report
    - "**Thickness and two. A the knee to**, and central venous catheter are unchanged."
    - Intended sentence: "**NG tube, ET tube,** and central venous catheter are unchanged"
  - 0.5 point:
    - Insignificant word substitution
    - "Asymmetric soft tissue **and me** right breast (40/96) which is nonspecific"
    - Intended sentence: "Asymmetric soft tissue **in the** right breast (40/96) which is nonspecific"
  - 0.25 point
    - Grammatical or typographic error
    - "Focal nodular right upper lobe opacity as previously **characterize** by CT scan"
    - Intended sentence: "Focal nodular right upper lobe opacity as previously **characterized** by CT scan"

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

	Radiologist 1		Radiologist 2		Radiologist 6	
Error Type	Oct	Apr	Oct	Apr	Oct	Apr
1	0	1	2	1	0	0
0.5	0	2	6	1	0	0
0.25	4	1	1	1	0	0

- **Turnaround Time**
  - The average turnaround time (i.e. length of time between the exam end time and report final time) was significantly increased after the first intervention
    - 6 hours, 15 minutes for September/October and 7 hours, 21 minutes for November/December, p = 0.009
  - However, the average turnaround decreased from November/December to January/February and March/April.
  - No significant difference was noted between the average turnaround time in September/October and the average time in March/April
    - 6 hours, 15 minutes for September/October and 6 hours, 29 minutes for March/April, p = 0.55

- **Figure 2.** Average turnaround time for the Chest Section for every 2 months. Increase in turnaround time between September/October and November/December. Decrease in turnaround time after the 3rd and 4th interventions.



DISCUSSION

- A significant improvement in scores was noted after the first intervention (in November/December, 4 months after the project began)
- Scores stabilized between the subsequent interventions (between 4 to 9 months)
  - The plateau in scores may be due to limitation in improvement (i.e. those with 100% accuracy are unable to improve)
- No definite trend was noted in the types of errors (i.e. whether "1 point" errors were more common before or after the intervention)
  - However, evaluation of word type was limited to three radiologists and there was a limited number of data points
- Some radiologists had consistently higher scores. These radiologists frequently had different methods of dictation and revision
    - One of the radiologists dictated the initial report, but then used the keyboard 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

- **Conclusion**
  - Use of intensive individual feedback within a peer group provided insight into patterns of errors that tended to be unique for each radiologist. The use of peer review also provided additional motivation for careful proof reading of reports, and resulted in a substantial reduction in the final error rate

References

1Krishnaraj A, Lee JKT, Laws SA, Crawford TJ. Voice recognition software: effect on radiology report turnaround time at an academic center. *Am. J. Roentgenol* 2010; 195:194-197
2McGurk S, Brauer K, MacFarlane TV, Duncan KA. The effect of voice recognition software on comparative error rates in radiology reports. *BJR* 2008; 80:767-770
3Chang CA, Strahan R, Jolley D. Non-clinical errors using voice recognition dictation software for radiology reports: a retrospective audit. *J Digit Imaging* 2010; 24:724-728