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

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

Introduction

Voice Recognition Software in Radiology

- Versions of voice recognition software have been used to generate radiology reports since the late 1980s[1]; though this early software was limited in its uses, accuracy, and report generation time
- Advances in voice recognition software have decreased turnaround times for final reports and increased 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 dictated[2]
- Although turnaround time is important, accuracy of the report is a primary concern
- Thus, error rates with voice recognition software are still a major issue
  - A recent study of MRI reports found that 30 to 42% of voice recognition reports contained errors, compared to 6 to 8% of transcriptionist reports[3]

Work Flow and Radiology Reports Prior to Voice Recognition Software

- Prior to voice recognition software, radiologists used transcriptionists or word processers to generate reports
- Radiologists dictated reports that were sent to a medically trained transcriptionist or staff typist, the transcribed report was then sent back to the radiologist for review and finalization
- Turnaround times were up to 24 hours or greater for finalization of reports
  - Long turnaround times are a detriment to the ordering physician, who often relies on immediate image interpretations
  - Long turnaround times also make it difficult for the radiologist to accurately review and revise the final report
    - The delay of 24 hours, makes it difficult for the radiologist to remember the images and original dictation when finalizing the transcription
    - Thus, simple word substitutions by the transcriptionist may easily be missed
- The development of voice recognition technology was of great interest to radiologists, as it could potentially improve turnaround time and increase work flow efficiency

Evolution of Voice Recognition Software in Radiology

- Reports of the use of voice recognition software in radiology are published as early as the late 1980s[1]
- Earlier versions of voice recognition software were limited to Navigation and Discrete Dictation programs
  - Speech recognition software used for "Navigation" is limited to commands that control an application
  - "Discrete dictation" systems identify each individual word that is spoken; thus requiring the speaker to pause between each word so that the computer can identify that word[4]
- Later software used Continuous Dictation systems
  - With continuous systems, users can speak at a natural pace
  - When spoken at a natural pace, words are blurred together and the acoustics of each word changes depending on the preceding and subsequent words[4]

Errors in Voice Recognition Software

- Understanding how voice recognition software works is important to understanding the common errors associated with this technology
  - Speech is converted to text based on vocabulary and language models
    - Vocabulary models match the acoustics of the spoken word with a word in a pre-defined dictionary
    - Language models assist the vocabulary models by picking words that are more likely to occur in that part of the sentence[4]
- Errors with voice recognition software occur when the acoustics of the spoken word do not match the vocabulary dictionary
  - For example, ambient noise or mispronunciation of a word changes the acoustics of a word
  - As mentioned, continuous speech changes the acoustics of individual words due to effects from the preceding and subsequent words
    - Thus, vocabulary dictionaries must recognize multiple acoustic versions of a single word

Impact of Errors in Radiology Reports

- Voice recognition software is unable to recognize every potential pronunciation of a particular word; thus creating potential errors in reports
- Reported frequency of voice recognition error rates in the radiology literature ranges from 4.8-42%[3,5-8]
- Uncorrected errors have potentially serious consequences
  - Some errors are minor (e.g. grammatical errors) or are easily recognized by the ordering physician 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 medicolegal consequences

Voice Recognition Software at University of Chicago

- Voice recognition software was implemented at University of Chicago in September 10, 2007
- The software currently in use is Nuance RadWhere for Radiology, which includes Dragon NaturallySpeaking by Nuance Communications
- Although this technology has been used for many years at our institution, the frequency of errors associated with voice recognition software has not been formally documented
- Furthermore, programs to address voice recognition errors are not commonly employed

Purpose

- The purpose of this study was to 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
  - The 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

Methods: 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

Methods: 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

Methods: Scoring

- 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 fluctuation

Quality Improvement Intervention

- Each month, the dictating radiologist was given his or her reports with the errors marked
  - This allowed the radiologist to identify the frequency of his or her errors and common types of errors
- Individual error rates and suggestions for improvements, including microphone positioning, use of macros and careful proof reading, were discussed at monthly section meetings
- Some radiologists also provided immediate feedback of errors on a day-to-day basis

Analysis

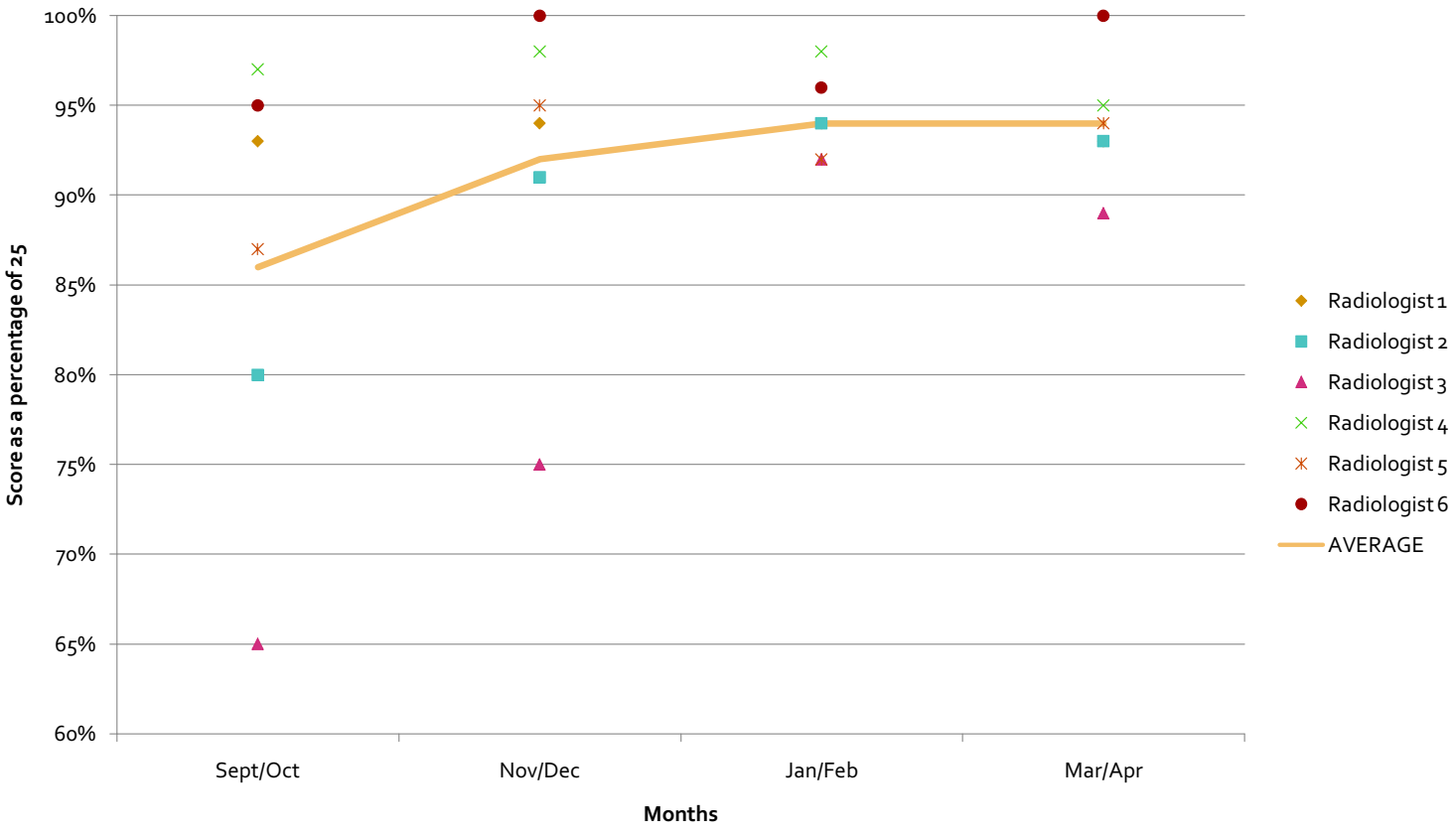
- Scores were given as a percentage (of 25) and as an error rate (25 minus the score)
  - Error rates were charted to demonstrate worsening or improvement
- A two tailed paired t-test was reported between each intervention (i.e. between every 2 month group) and for the first 2 months compared to the last 2 months
- Turnaround time was based on the length of time between the exam end time and report final time for each case

Results

Scores Per Radiologist

	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%

Individual and Averaged Scores



Before and After the Quality Improvement 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 94%, p value 0.53
  - Jan/Feb to Mar/Apr: 94% to 94%, p value 0.84

Types of errors

- 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)
- No significant difference was noted between the types of errors from October and April

	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

*1 = major error, changing the meaning of the report, 0.5 = insignificant word substitution, 0.25 = grammatical or typographic error

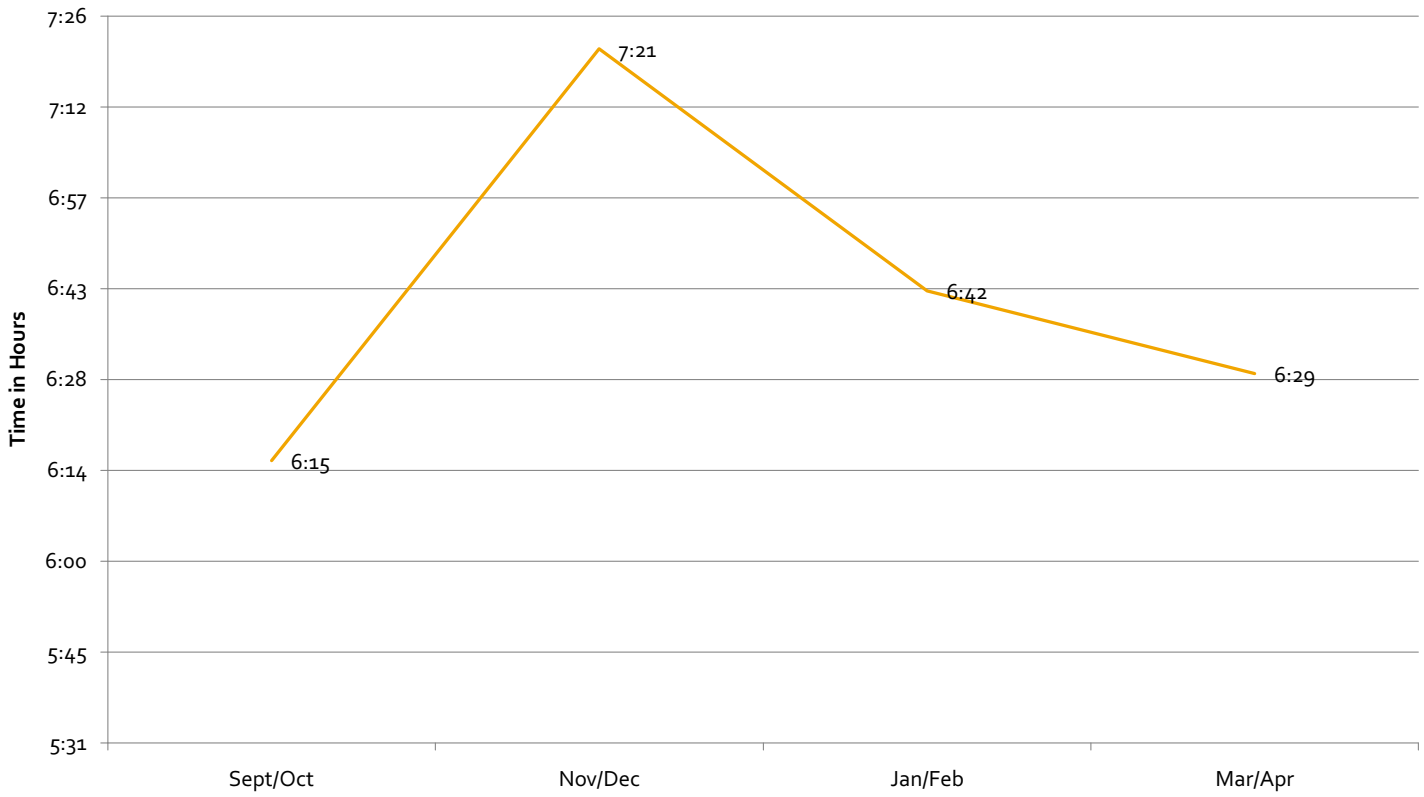
Examples of error types

- 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"

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 time decreased from in January/February and March/April
- No significant difference was noted between the average turnaround time in September/October and the average turnaround time in March/April
  - 6 hours, 15 minutes for September/October and 6 hours, 29 minutes for March/April, p = 0.55

Turnaround Time of Reports



Discussion

Outcomes of the Quality Improvement Initiative

- A significant improvement in scores was noted after the first intervention (4 months after the project began)
  - This finding suggests that feedback from peers increases awareness of voice recognition errors
  - Furthermore, a peer review program may increase motivation to provide error-free reports
- Scores appeared to plateau between the subsequent interventions (between 4 to 9 months)
  - This plateau may be due to limits of improvement (i.e. those with 100% accuracy do not have further room for improvement)
  - The "intervention" or discussion of errors at the department meetings did not change from month to month
- 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 types of errors was limited to three of the six radiologists
  - Furthermore, evaluation was limited to two of the study months (the first and the last months)
  - These limitations could mask any potential trends in frequency of error types

Differences in Radiologists' Scores

- Some radiologists had consistently higher scores than other radiologists
- These radiologists had different methods of dictation and revision
  - One of the radiologists, who frequently had 100% scores, dictated the initial report, but then used the keyboard for editing
    - Using the keyboard for editing may have limited additional voice recognition errors
  - Subjectively, it appeared that radiologists with fewer words or a "telegraphic" style of dictation had fewer errors
    - Fewer words or phrases allows for fewer opportunities for error

Effect on Turnaround Times

- After the first intervention (between September/October and November/December), there was a significant increase in report turnaround time
- However, turnaround time returned to baseline during the subsequent months
- This increase in turnaround time may have been a result of multiple factors
  - The focus on voice recognition errors and reduction of errors may have increased the time spent reviewing reports and thus increased the report turnaround time
  - The months of November and December, when the turnaround time increased, was a time of decreased staffing due to the holidays and/or RSNA
    - With decreased staff, the workload increases and thus the time between finalization of the images and review of images by the attending radiologist increases

Strategies for future intervention

- Future directions of this quality improvement project include focus on specific changes in dictation behavior
- Encouraging the use of a certain type of dictation style (i.e. the telegraphic style of dictation used by several of the radiologists) may help reduce error rates
- The development of standard "macros" (or preset dictations) may help reduce variability in reports and frequent errors
- Implementation of microphone headsets or changing microphone position may reduce the effects of ambient noise and thus reduce error rates

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

- 1. Robbins AH, Horowitz DM, Srinivasan MK, et al. Speech controlled generation of radiology reports. Radiology 1987; 164(2):569-573
- 2. Krishnaraj A, Lee JKT, Laws SA, and Crawford TJ. Voice recognition software: Effect on radiology report turnaround time at an academic medical center. AJR Am J Roentgenol 2010; 195(1):194-197
- 3. Strahan RH and Schneider-Kolsky ME. Voice recognition versus transcriptionist: Error rates and productivity in MRI reporting. J Med Imaging Radiat Oncol 2010; 54(5):411-414
- 4. Lai J and Vergo J. MedSpeak: report creation with continuous speech recognition. American Association for Computing Machinery Special Interest Group-Computers 97 conference proceedings 1997; 431-438
- 5. Chang CA, Strahan R, Jolley D. Non-clinical errors using voice recognition dictation software for radiology reports: a retrospective audit. J Digit Imaging 2011; 24(4):724-728
- 6. Pezzullo JA, Tung GA, Rogg JM, Davis LM, Brody JM, and Mayo-Smith WW. Voice recognition dictation: radiologist as transcriptionist. J Digit Imaging 2008; 21(4):384-389
- 7. McGurk S, Brauer K, MacFarlane TV, Duncan KA. The effect of voice recognition software on comparative error rates in radiology reports. Br J Radiol 2008; 81(970):767-770
- 8. Quint LE, Quint DJ, Myles JD. Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. J Am Coll Radiol 2008; 5(12);1196-1199