J Digit Imaging (2011) 24:724–728
DOI 10.1007/s10278-010-9344-z

Non-Clinical Errors Using Voice Recognition Dictation Software for Radiology Reports: A Retrospective Audit

Chian A. Chang · Rodney Strahan · Damien Jolley

Published online: 26 October 2010

Abstract The purpose of this study is to ascertain the error rates of using a voice recognition (VR) dictation system. We compared our results with several other articles and discussed the pros and cons of using such a system. The study was performed at the Southern Health Department of Diagnostic Imaging, Melbourne, Victoria using the GE RIS with Powerscribe 3.5 VR system. Fifty random finalized reports from 19 radiologists obtained between June 2008 and November 2008 were scrutinized for errors in six categories namely, wrong word substitution, deletion, punctuation, other, and nonsense phrase. Reports were also divided into two categories: computer radiography (CR= plain film) and non-CR (ultrasound, computed tomography, magnetic resonance imaging, nuclear medicine, and angiographic examinations). Errors were divided into two categories, significant but not likely to alter patient management and very significant with the meaning of the report affected, thus potentially affecting patient management (nonsense phrase). Three hundred seventy-nine finalized CR reports and 631 non-CR finalized reports were examined. Eleven percent of the reports in the CR group had errors. Two percent of these reports contained nonsense phrases. Thirty-six percent of the reports in the non-CR group had errors and out of these, 5% contained nonsense phrases. VR dictation system is like a double-edged sword. Whilst there are many benefits, there are also many pitfalls. We hope that raising the awareness of the error rates will help in our efforts to reduce error rates and strike a balance between quality and speed of reports generated.

Keywords Voice recognition · Reporting · Productivity · Speech recognition · Workflow · Radiology reporting

C. A. Chang (✉) · R. Strahan · D. Jolley
Southern Health Department of Diagnostic Imaging,
246 Clayton Road,
Clayton, VIC 3168, Australia
e-mail: chianaun@gmail.com

Introduction

Radiologists play a crucial role in patient care. It is the responsibility of the radiologist to convey findings and reports to the referring clinician in a timely and accurate manner. Before voice recognition (VR) dictation, the radiologist would dictate a report which is then typed by a medical typist and sent back to the radiologist for editing and report finalization. Only then can the report be sent or viewed by the referring clinician. As radiology practices grow, they are under increasing pressure to increase the number of examinations and decrease the turnaround time of reports generated. This is precisely why VR dictation is introduced, to reduce the time between report dictation and finalization.

It is not without its drawbacks. It is known that one of the major disadvantages of such a system is transcription errors. Errors range from deletion, wrong word substitution, or reports containing confusing and inaccurate sentences. Pezzullo et al. reported [1] that 35% of VR reports after sign-off contained errors. Quint and colleagues [2] have reported an error rate of 22% using voice recognition dictation. McGurk and colleagues [3] had an error rate of only 4.8%, but half of them contained errors that affected understanding. Careful dictating and proofreading is important to reduce such errors as they have dire consequences in patient management. Examples of such errors include dictating left instead of right or the dictation program

mistakes like sounding words, e.g., renal and adrenal, lateral and bilateral, or hyperintense and hypointense.

In this study, we describe our experience and error rates relating to VR dictation and briefly discuss its impact on radiologists, referring clinicians, costs, and ultimately patient care.

Method

The study was performed in the Southern Health Department of Diagnostic Imaging, Melbourne, VIC. The dictation system is a GE RIS with Powerscribe 3.5 VR. It has been used at our institution for 4 years prior to us conducting this study. All the radiologists in the sample population have had several years of experience using a VR dictating system.

At least 50 random finalized reports were obtained from 19 radiologists in the department from June 2008 until November 2008. The sample size was determined by the statistical model to gain sufficient statistical power. These reports were divided into two groups, computer radiography (CR=plain films) and non-CR (which encompass ultrasound, computed tomography, magnetic resonance imaging, nuclear medicine, and angiographic examination reports). These reports were checked by two independent individuals (one radiologist and one radiology trainee) who first generated a protocol and standardized a scoring system. A blinded double-read of several reports was then performed by the two individuals to establish the reliability of the measurement instrument. Both the reviewers were not specifically trained for this task but took the position of a referrer, i.e., medically trained with an average English comprehension level.

Each report was scrutinized for grammatical errors in six categories, wrong word substitution (A), deletion (C), insertion (D), punctuation (E), other (F), and the most significant error, nonsense phrases (B) which include sentences which are meaningless or contain words completely irrelevant to the report. If more than one error was present, the worst error was documented, e.g., if a series of deletions and insertions were present, it would be classified as a nonsense phrase. Images from the examinations (CR, CT, MRI, etc.) performed were not reviewed during this process thus, errors of radiological interpretation are not included in this study. This is a retrospective study so that the 19 radiologists involved were not aware that their reports were being scrutinized, so the results are blinded and anonymous. VR dictation software allows the use of canned texts as a method of reducing errors and to improve turnaround time. Less than 2% of the reports scrutinized were canned.

For statistical analysis, errors are divided into two groups, significant but not likely to alter patient management and very significant with the meaning of the report affected, thus potentially affecting patient management. Some would argue that any error reflects badly on the author, with the radiology report being the primary means of communicating to the referring clinician the results of an imaging examination or procedure. Very significant errors were nonsense phrases. This is the most serious type of error as the referring clinician is unable to understand or is at risk of misinterpreting the report. The sentences are so fragmented they have to deduce their own conclusion from the radiologist's report. This vastly increases the risk of patient mismanagement. Significant errors are errors such as wrong word substitution, missing/additional words or punctuation, where the referring clinicians can still understand the report with a lower risk of misinterpretation. Examples of such errors include:

Type A: Wrong word substitution.
"Haemodynamics performed by urologist" instead of Urodynamics.
"There is mild prominence of the leaf renal pelvis..."
Type B: Nonsense phrase.
Finger X-ray: "Present no and after the metacarpal bones."
MRI head: "No flow was demonstrated within the coil right hysterectomy occurring artery aneurysm."
Abdominal ultrasound: "The spleen is markedly enlarged with length there is the ureter suggestive of portal hypertension."
Type C: Deletion.
"The ventricles are normal in size morphology."
Missing an and.
Type D: Insertion.
"The right osteomeatal unit is obscure is obstructed by mucosal thickening."
"…located within the right lower lobe the middle lobe."
Type E: Punctuation.
"No separate mass is identified in neck the thyroid is mildly enlarged and slightly heterogeneous particularly towards the upper pole of the right lobe the right lobe has a length of 5 cm compare to the left lung with a length of 2.7 cm."
What happened to punctuation?
Type F: Other, including spelling.
"…there is minor irregulatiy to the posterioir cord at the T4, T5 and T6 levels."

The total number of CR and non-CR reports in the entire database for each radiologist during this time period comprised the study population. There were 52,573 reports in total (39,416 CR and 13,157 non-CR reports). Out of this study population, at least 50 reports from each radiologist were randomly selected giving a total of 990 reports (379 CR and 631 non-CR reports). As our center is a training hospital, reports which have been dictated by radiology trainees and

Table 1 CR crude error rates and weightings

Radiologist	Type	Pop	Weight (%)	Number	Error	Rate (%)
A	CR	2,397	6	20	1	5
B	CR	520	1	20	6	30
C	CR	1,965	5	15	1	7
D	CR	385	1	6	0	0
E	CR	709	2	14	2	14
F	CR	7,988	20	50	0	0
G	CR	443	1	40	17	43
H	CR	30	0	3	1	33
I	CR	519	1	12	1	8
J	CR	4,892	12	11	1	9
K	CR	3,764	10	50	3	6
L	CR	2,531	6	19	2	11
N	CR	1,107	3	20	0	0
O	CR	1,048	3	20	0	0
P	CR	1,604	4	35	4	11
Q	CR	7,892	20	24	1	4
R	CR	1,622	4	20	2	10
Crude	CR	39,416	100	379	42	11

Pop number of reports in the study population, *Number* number of reports reviewed, *Error* number of reports containing errors

sent to the radiologists for checking and verification were excluded. Reports randomly selected are those that were dictated and finalized by the respective radiologists.

Results

Tables 1 and 2 display the results for CR and non-CR, respectively. Three hundred seventy-nine finalized CR reports and 631 non-CR finalized reports were examined. Forty-two (11%) reports in the CR group contained errors. Six (2%) of these reports contained nonsense phrases (type B error). Two hundred thirty (36%) reports in the non-CR group contained errors and out of these, 34 (5%) contained type B errors. These are termed "crude error rates" but do not take into account the varying number of reports generated by each radiologist. To allow for this, weighted proportion error rates were computed using each radiologist's

Table 2 Non-CR crude error rates and weightings

Radioliologist	Type	Pop	Weight (%)	Number	Error	Rate (%)
A	Non-CR	1,695	13	30	8	27
B	Non-CR	197	1	30	14	47
C	Non-CR	586	4	35	20	57
D	Non-CR	490	4	44	21	48
E	Non-CR	1,395	11	36	33	92
G	Non-CR	501	4	10	5	50
H	Non-CR	462	4	47	29	62
I	Non-CR	487	4	38	11	29
J	Non-CR	585	4	39	11	28
L	Non-CR	332	3	31	8	26
M	Non-CR	605	5	50	12	24
N	Non-CR	717	5	50	21	42
O	Non-CR	1,622	12	50	10	20
P	Non-CR	734	6	15	8	53
Q	Non-CR	919	7	26	1	4
R	Non-CR	1,422	11	50	15	30
S	Non-CR	408	3	50	3	6
Crude	Non-CR	13,157	1	631	230	36

Pop number of reports in the study population, *Number* number of reports reviewed, *Error* number of reports containing errors

total number of reports in the study population. Some reports contain multiple errors. Table 3 displays the total number of errors.

For CR reports, the overall weighted error rate is 6% (95% CI from 2.9% to 9.1%). This was much less than the crude rate of 11%. The reason for this is that in the population, 40% of the CR reports were provided by two radiologists who had low error rates (0–4%). The crude type B error rate for CR reports was 2% and the weighted error rate is only 0.5% (95% CI from 0.1% to 1%).

For non-CR reports, the weight is spread more evenly and the weighted error rate was close to the crude. The weighted error rate for non-CR reports is 38% (95% CI from 34% to 42%). The crude type B error rate for non-CR reports was 5%, but the weighted error rate is marginally greater at 8% (95% CI from 6% to 10%).

Statistical Analysis

Table 4 and Fig. 1 show the results of a binary regression analysis which estimates the relative risk of errors between CR and non-CR reports and between radiologists. The regression models the relative risk of errors between groups using a Bernoulli error and a log link. The numbers in the column headed "exp (coef)" can be interpreted as relative risk estimates with confidence intervals in the far right columns.

This shows that the non-CR reports have 3.5 times the risk of error (on average) than CR reports. Radiologist A was selected as the baseline (purely by alphabetical order). Each of the other radiologists is rated against radiologist A by a relative risk estimate. For example, radiologist B is about 2.3 times more error prone compared to radiologist A while radiologist F has a relative risk of 0.0 compared to radiologist A because of an error rate of 0 (who also did not report any non-CR examinations). Radiologist T rated well at less than one fourth the error rate of radiologist A.

Table 3 Total number of errors

Error type	Number of errors	Percentage of total errors
A	181	37.6
B (CR)	6	1.2
B (non-CR)	34	7
C	69	14.3
D	133	27.6
E	30	6.2
F	29	6
Total errors	482	

Table 4 Binary regression analysis estimating risk of errors

Covariate	Relative risk	95% Confidence interval
Type of report		
CR[a]	1	
Non-CR	3.6	3.5 to 3.6
Radiologist		
A[a]	1.0	
B	2.3	2.276 to 2.386
C	2.0	1.989 to 2.063
D	1.9	1.830 to 1.903
E	3.6	3.589 to 3.691
F	0.0	0.000 to 1.322
G	3.0	2.911 to 3.028
H	2.5	2.416 to 2.501
I	1.2	1.125 to 1.185
J	1.2	1.155 to 1.205
K	0.9	0.831 to 0.872
L	1.3	1.228 to 1.290
M	1.0	0.933 to 0.979
N	1.4	1.401 to 1.462
O	0.7	0.705 to 0.738
P	1.9	1.853 to 1.929
Q	0.4	0.436 to 0.458
R	1.2	1.204 to 1.250
S	0.2	0.226 to 0.253

[a] Baseline category

Discussion

Potential benefits of VR dictation systems are decreased report turnaround time by providing almost instantaneous report availability on the PACS system and cost savings in

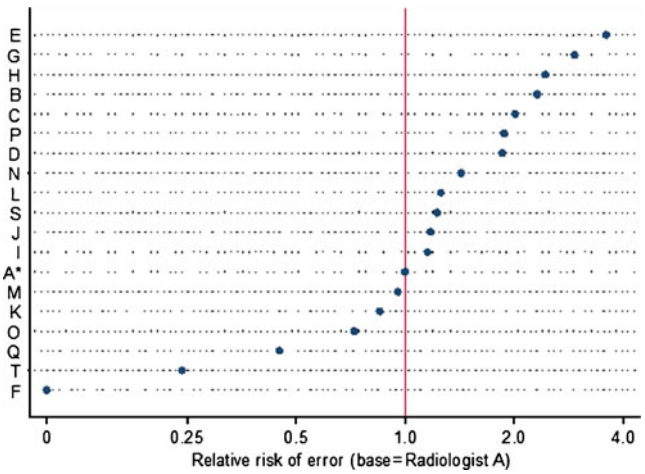

Fig. 1 Relative risk estimates for each radiologist, relative to radiologist A, taken from a binomial regression model

terms of transcriptionist wages. What is difficult to measure and compare is the additional task for the radiologist as a transcriptionist. This results in increased stress and leads to more errors, time required to dictate a report, and detail/length of reports.

A total of 6% of CR reports and 38% of non-CR reports contained errors. Our CR report error rates are relatively low compared to the literature [1, 2] but are relatively high for non-CR reports. Although the type B error rate was low for CR reports (0.5%), it is relatively high for non-CR reports (8%). There was also a wide range of error rates between each radiologist.

There are many causes for these error rates. This study also indicates that error rates are related to the type of report (CR vs. non-CR). The shorter and less complex the examination, the less likely errors are made. Other causes range from pronunciation, clarity and speed of the radiologist dictating the report, to failure to proofread the reports accurately. Carelessness of the reporter may also be a possibility.

It has been established [3] that disruptions (e.g., phone calls, noise, interruptions from clinicians) greatly contribute to the error rates. A radiologist constantly interrupted while trying to proofread a complex report will inevitably have trouble concentrating. As a result, errors which would have normally been corrected have the potential to be overlooked. This probably results in a type F error occurring, particularly spelling errors. Another cause for this type of error may be user frustration with the VR system which arises when the radiologist reverts back to typing their reports. Voice recognition dictation is also widely believed to reduce the radiologist's productivity [4]. Radiologists under pressure to complete the reporting worklist may try to retrieve this lost time by dictating shorter reports with fragmented and incomplete sentences.

Another interesting finding presented by Pezzullo et al. [1] is the issue of cost. There is an increase in dictation cost using a voice recognition system compared to conventional transcription. This is due to a decrease in the productivity of the radiologists. Additional time is required for the radiologist to transcribe and proofread their reports. Another valid point is, although the institution saves cost by having a VR system (not having to employ medical transcriptionists), the saved revenue is not necessarily transferred back to the radiologist. Their workload has increased as they are now also effectively the transcriptionist.

Limitations of the study include underestimating the number of errors as some potential errors that could have been overlooked include incorrect right or left substitution, error in measurements, and measurement units (e.g., cm and mm) and incorrect dates.

Finally, VR systems may pose a medicolegal dilemma. A report with numerous spelling and grammatical errors may be viewed as carelessness on the radiologist's part. Imagine what the public, the patient, and the malpractice attorney's view is on reports with confusing and nonsensical sentences! Thus, every error is significant.

Conclusion

A relatively low rate of 6% of our CR reports had errors. What is more significant is the high (38%) rate for non-CR reports with errors. We have also found a considerable variation between radiologists in their error rates. It is also likely that some radiologists are unaware of the relatively high error rates that occur when using a VR dictating system. At our institution, we have introduced a "gate keeper role" where a radiologist or radiology trainee is the point of contact for clinicians seeking requests or results to reduce the number of interruptions. We are also in the process of employing more staff so radiologists are less pressured to complete the workload. We hope that these findings result in an increase in awareness and reduced error rates (especially type B errors) in our efforts to find a balance between quality and speed of reports generated at our institution.

References

1. Pezzullo JA, Tung GA, Jefferey MR, Lawrence MD, Jefferey MB, William WM: Voice recognition dictation: radiologist vs transcriptionist. J Digital Imaging 21:384–389, 2008
2. Quint LE, Quint DJ, Myles JD: Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. J Am Coll of Radiol 5:1196–1199, 2008
3. McGurk S, Brauer K, MacFarlane TV, Duncan KA: The effect of voice recognition software on comparative error rates in radiology reports. BJR 80(970):767–770, 2008
4. Hayt DB, Alexander A: The pros and cons of implementing PACS and speech recognition systems. J Digital Imaging 14:149–157, 2001