J Digit Imaging (2014) 27:581–587
DOI 10.1007/s10278-014-9699-7

Creation and Implementation of Department-Wide Structured Reports: An Analysis of the Impact on Error Rate in Radiology Reports

C. Matthew Hawkins · Seth Hall · Bin Zhang · Alexander J. Towbin

Published online: 24 May 2014

Abstract The purpose of this study was to evaluate and compare textual error rates and subtypes in radiology reports before and after implementation of department-wide structured reports. Randomly selected radiology reports that were generated following the implementation of department-wide structured reports were evaluated for textual errors by two radiologists. For each report, the text was compared to the corresponding audio file. Errors in each report were tabulated and classified. Error rates were compared to results from a prior study performed prior to implementation of structured reports. Calculated error rates included the average number of errors per report, average number of nongrammatical errors per report, the percentage of reports with an error, and the percentage of reports with a nongrammatical error. Identical versions of voice-recognition software were used for both studies. A total of 644 radiology reports were randomly evaluated as part of this study. There was a statistically significant reduction in the percentage of reports with nongrammatical errors (33 to 26 %; $p=0.024$). The likelihood of at least one missense omission error (omission errors that changed the meaning of a phrase or sentence) occurring in a report was significantly reduced from 3.5 to 1.2 % ($p=0.0175$). A statistically significant reduction in the likelihood of at least one comission error (retained statements from a standardized report that contradict the dictated findings or impression) occurring in a report was also observed (3.9 to 0.8 %; $p=0.0007$). Carefully constructed structured reports can help to reduce certain error types in radiology reports.

Keywords Radiology · Structured reports · Errors

C. M. Hawkins · S. Hall · A. J. Towbin (✉)
Department of Radiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue ML 5031, Cincinnati, OH 45229, USA
e-mail: alexander.towbin@cchmc.org

B. Zhang
Department of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA

C. M. Hawkins
Department of Radiology, University of Washington School of Medicine, 1959 Pacific St SS-202, Seattle, WA 98195-7117, USA

Introduction

Textual errors are common in radiology reports. Previous studies have reported errors in 4.8 to 22 % of radiology reports when using speech recognition software [1–3]. However, a recent, thorough evaluation of all error types within our department found textual errors in approximately 40–60 % of all radiology reports [4].

Recently, there has been a push towards structured reporting in radiology [5, 6]. Proponents of structured reporting tout the many potential benefits including the possibility of decreasing the number of textual errors in final reports [4–7]. Studies evaluating the ability of structured reports to decrease the number of textual errors have not been performed.

Since our initial study which established a baseline error rate, our department has created and implemented department-wide standardized, structured reports [8]. During the design of each report, special attention was placed on avoiding specific errors such as double periods and retained structured elements. The primary objective of this study was to evaluate the impact our revamped reports had on the error rate within radiology reports. We hypothesized that the use of structured reports would decrease the overall error rate in comparison to less structured but standardized templates.

Materials and Methods

After obtaining a waiver from the institutional review board, all radiology reports and digital audio files dictated over a 3-day period were logged into a Microsoft Access database (Microsoft, Redmond, WA). All reports were dictated using speech recognition software (RadWhere v3.0.24, Nuance, Boston, MA). The speech recognition software was identical (including version) to what was used in our prior study [4].

Each report was dictated using one of 228 distinct standardized, structured reports in use in our department. The structured reports are pre-populated in the dictation field based on the radiology information system procedure code for each examination. Thus, for every study performed in the department, the radiologist starts with the department standard, structured report in the dictation field. The radiologist is then able to modify portions of the report via structured pick-lists, fill-in fields, and/or prose dictation to answer the clinical question or to report findings that are not structured. The combination and frequency of use of pick-lists, fill-in-fields, and/or prose dictation is varied for each report and is determined by collaborative consensus of members from each imaging section of the department [8].

The same two radiologists that classified errors in the baseline study acted as reviewers for this study. Working independently, each radiologist reviewed a sample of radiology reports from the study period. The reports were presented to the reviewers in a random fashion as determined by the Microsoft Access database. The reviewers each compared the text report to its corresponding audio file. All audio files contained only words that were dictated by the radiologist. The reviewers used their knowledge of the structured reports to identify when these dictated words triggered the translation of a structured element. Errors in the text reports were tabulated and categorized into five main classes: nonsense errors, missense errors, spelling/grammatical errors, translation errors, and errors of omission/comission. Each class was further subdivided into the same 11 error subtypes as described in the prior study (Table 1).

Several error rates were calculated including the average number of errors per report, average number of nongrammatical errors per report, the percentage of reports containing an error, and the percentage of reports containing a nongrammatical error. The formula for each error rate calculation is included in Table 2. The calculated error rates were then compared to the results from our prior study, prior to implementation of structured reports.

In our previous study, there was a discrepancy between how reviewers graded grammatical errors, with only one reviewer grading sentence fragments as errors. In the current study, both reviewers used the same definition of a grammatical error, and sentence fragments were not graded as errors. Because there was a discrepancy in the way the two reviewers graded grammatical errors in the prior study, only the results for the reviewer who did not grade sentence fragments as errors were used for retrospective comparison in the current study.

During the random review, a small percentage of reports were graded by both reviewers. These reports were used to assess inter-reader agreement for each error type by comparing the likelihood (%) of at least one error occurring in a report for each reviewer. In addition, the number of errors per dictated word was compared between reviewers for this subset of reports. Because the purpose of this study was to determine the frequency of errors in structured reports, the duplicated reviews were counted for Reviewer 1 only and were excluded from the study statistics of Reviewer 2. This was done so that errors were not double counted for the overall study.

All error rates were statistically compared prior to and after implementation of standardized, structured reports. The proportion difference of at least one error subtype being present in each report was evaluated using the Chi-square test or Fisher's exact test. This type of analysis was performed for all error subtypes (other than Spelling/Grammar—grammatical errors) because the number of these error subtypes in reports is most commonly zero, and rarely is there more than one of a specific error subtype in a report. The differences in the mean grammatical errors per report, average number of errors per report, and average number of nongrammatical errors per report were assessed by Poisson regression. Finally, the percentage of reports with an error, and the percentage of reports with a nongrammatical error were evaluated using the Chi-square test. All statistical analyses were performed using SAS® (Version 9.3, Cary, NC). *P* values of less than 0.05 were considered to indicate statistical significance.

Results

There were 2,117 reports dictated during the 3 days data was collected. Of these, 644 unique randomly distributed reports were analyzed by one of the two reviewers. The two radiologists graded a different number of unique reports; Reviewer 1 graded 441 reports while Reviewer 2 graded 264 reports. There were 61 reports graded by both reviewers. The 61 reports were used to assess inter-reader agreement. The error data from Reviewer 1 for these 61 reports is included in the overall error analysis. These reports are excluded from Reviewer 2's data leaving Reviewer 2 with 203 included reports.

Overall, the two reviewers graded reports similarly. The likelihood of a report containing at least one error was not significantly different for any error type although it bordered on significant for missense translational errors ($p=0.05$) and spelling/grammar—grammatical errors ($p=0.09$). The data comparing inter-reader agreement is shown in Table 3. In

Table 1 From Hawkins CM et al. [4] used with permission from Springer Science+Business Media B.V.

Error category	Error type	Definition	Intended/spoken phrase	Transcribed/reported phrase
Nonsense	Nonsense	Passages/words/phrases that make no sense or have no sensible meaning.	The lungs are clear.	The lungs nuclear.
Missense	Missense—translational	Translation error that changes the meaning of a phrase/sentence.	There is no opacity in the left lung.	There is an opacity in the left lung.
	Missense—omission	Words not transcribed (omitted) that subsequently change the meaning of a phrase/sentence.	There is no pneumonia.	There is pneumonia
	Missense—human	Human error that changes the meaning of a sentence.	Right lower lobe pneumonia.	Left lower lobe pneumonia.
Spelling/ Grammar	Spelling/grammar—typographical error	Grammatical/spelling error resulting in a typographical error that is not a nonsense/missense error.	The right lung is clear.	The rgiht lung is clear.
	Spelling/grammar—homonym error	Misuse of words or phrases that sound the same, but are semantically distinct.	The right lung is clear.	The write lung is clear.
	Spelling/grammar—grammatical error	Standard grammatical errors including the use of sentence fragments.	The lungs are clear.	The lungs is clear.
	Improper period use error	Duplicate periods or lack of a period at the end of a sentence.	The lungs are clear.	The lungs are clear..
Omission/ comission	Omission—omission error (other)	Omitted word/phrase that does not result in a missense or nonsense error.	The lungs are clear.	Lungs are clear.
	Comission error	Retained statement from a standardized template that contradicts the dictated findings or impression.	There is a hazy opacity obscuring the right hemidiaphragm.	The lungs are clear. There is a hazy opacity obscuring the right hemidiaphragm.
Translational	Translational (other)	Translation error that does not result in a nonsense/missense error. The resulting sentence still has sensible meaning, as opposed to nonsense errors.	The lungs are clear.	The lungs are clean.

addition to the individual error rates, there was no significant difference in the mean number of errors per dictated word (mean [standard deviation]=0.008 [0.029] for Reviewer 1 and 0.014 [0.030] for Reviewer 2; $p=0.34$).

The error rates generated from this review were compared to the error rates generated from 311 reports that were evaluated prior to implementation of structured reports. Radiologists used structured reports 100 % of the time in the current study.

There were 461 errors identified and classified in the current study. At least one error was present in 37 % of reports and there were 0.72 errors per report. Neither value is significantly different compared to the prior study when 41 % of reports had at least one error and 0.75 errors per report were present ($p=0.27$ for the percent of reports with at least one error; $p=0.43$ for the number of errors per report). Grammatical errors were again the most common type of error, accounting for approximately 46 % of all errors compared to 41 % in the prior study. This difference was not statistically significant ($p=0.7$). Excluding grammatical errors, 26 % of reports contained nongrammatical errors. This is a statistically significant decrease from the prior study where 33 % of reports had nongrammatical errors ($p=0.024$). The results comparing the general error rates of the two studies can be found in Table 2 and the results comparing the specific error types can be found in Table 4.

Table 2 Error rate formulas and comparison before and after implementation of department-wide standardized reports

Error rate	Formula	Prior to implementation of structured reports[a]	Following implementation of structured reports	*p* value
Errors per report	Total number of errors/total number of reports	234/311=0.75	461/643=0.72	0.431
Nongrammatical errors per report	Total number of nongrammatical errors/total number of reports	136/311=0.44	243/643=0.38	0.176
Percentage of reports with errors	(Number of reports with errors/total number of reports)×100 %	(128/311)×100 %=41 %	(241/643)×100 %=37 %	0.267
Percentage of reports with nongrammatical errors	(Number of reports with nongrammatical errors/total number of reports)×100 %	(103/311)×100 %=33 %	(168/643)×100 %=26 %	0.024

[a] Data from Hawkins CM et al. [4]

Table 3 Individual error rates and inter-reader comparison for each error type

	Likelihood (%) of at least one error in a report		*p* value
	Reviewer 1	Reviewer 2	
Nonsense error	3.3	1.7	1.00
Missense—translational error	13.3	3.3	0.05
Missense—omission error	1.7	0.0	0.50
Missense—human error	0.0	0.0	–
Spelling/grammar—typographical error	0.0	0.0	–
Spelling/grammar—homonym error	0.0	0.0	–
Spelling/grammar—grammatical error	23.3	11.7	0.09
Improper period use error	3.3	10.0	0.27
Omission—omission error (other)	3.3	3.3	1.00
Omission—comission error	3.3	0.0	0.50
Translational (other)	6.7	3.3	0.68

Table 4 Distribution of error types and comparison before and after implementation of revamped, department-wide standard reports

	Total number of errors		Percentage (%) of total errors		Errors per report		Number of reports per one error		Likelihood (%) of at least one error in a report		*p* value[c]
	Trial 1[a]	Trial 2[b]	Trial 1[a]	Trial 2[b]	Trial 1[a]	Trial 2[b]	Trial 1[a]	Trial 2[b]	Trial 1[a]	Trial 2[b]	
General error types											
Translational errors[d]	88	175	37.6	38	0.283	0.272	3.53	3.68	10.3	9.6	0.75
Missense errors[e]	50	82	21.4	17.8	0.178	0.161	6.22	7.85	14.5	10.6	0.08
Grammatical errors[f]	98	218	41.8	47.3	0.315	0.339	3.173	2.95	21.5	19.4	0.45
Specific error types											
Nonsense error	16	29	6.8	6.3	0.051	0.045	19.4	22.2	5.1	6.2	0.541
Missense—translational error	34	72	14.5	15.6	0.109	0.112	9.1	8.9	10.3	9.6	0.753
Missense—omission error	11	8	4.7	1.7	0.035	0.012	28.3	50.5	3.5	1.2	0.0175
Missense—human error	5	2	2.1	0.4	0.016	0.003	62.2	322	1.6	0.3	0.041
Spelling/grammar—typographical error	3	3	1.3	0.7	0.010	0.005	104	215	1	0.3	0.337
Spelling/grammar—homonym error	0	2	0	0.4	0	0.003	NA	322	0	0.3	1
Spelling/grammar—grammatical error	95	213	40.6	46.2	0.305	0.331	3.3	3	21.5	19.4	0.448
Improper period use error	31	57	13.2	12.4	0.1	0.09	10	11.3	9	7.3	0.362
Omission—omission error (other)	15	30	6.4	6.5	0.048	0.047	20.7	21.5	4.5	4.4	0.917
Omission—comission error	12	9	5.1	2	0.039	0.014	25.9	71.6	3.9	0.8	0.0007
Translational (other)	12	36	5.1	7.8	0.039	0.056	25.9	17.9	3.6	5	0.320

Sample calculation nonsense error, trial 1: (16 nonsense errors/234 total errors)×100%=6.8 % of all errors; 16 nonsense errors/311 total reports=0.051 errors per report; 1/0.051 errors per report=1 error per 19.4 reports

[a] Trial 1=prior to implementing department-wide structured reports (need reference). Total errors (Trial 1)=234; total reports (trial 1)=311

[b] Trial 2=after implementation of department-wide structured reports. Total errors (trial 2)=461; total reports (trial 2)=643

[c] The *p* value for grammatical errors compared the average number of grammatical errors (spelling/grammar) per report prior to and after implementing structured reports using the two sample *t* test with comparison of the means. The remaining *p* values represent the evaluation of the proportion difference of at least one error subtype being present in each report using the Chi-square test or Fisher's exact test. *p* values less than 0.05 were considered significant

[d] General translational errors represent the sum of nonsense errors, missense translational errors, missense omission errors, omission omission errors, and translational (other) errors

[e] General missense errors represent the sum of missense translational errors, missense omission errors, and missense human errors

[f] General grammatical errors represent the sum of spelling/grammar typographical errors, spelling/grammar homonym errors, and spelling/grammar grammatical errors. Note that spelling/grammar improper period is not counted as a grammatical error in this analysis

Missense errors (errors that change the meaning of a phrase or sentence) were again the most common nongrammatical error type, accounting for 17.8 % of all errors. The likelihood of at least one missense error occurring in a report decreased from 14.5 % in the prior study to 10.6 % in the current study (p=0.08). The most common subtype was again missense translational errors (errors of translation that changed the meaning of the phrase or sentence), which accounted for 16 % of all errors. There were 0.11 missense translational errors per report in the current study. The overall likelihood of a report having at least one missense translational error was unchanged from the prior study (10.3 % in the first study and 9.6 % in the current study; p=0.75).

Overall, there were 0.012 missense omission errors per report (omission errors that changed the meaning of a phrase or sentence). This was decreased from the 0.035 errors of this type per report in the prior study. The likelihood of a report containing at least one missense omission error was significantly reduced from 3.5 to 1.2 %. (p=0.0175).

The likelihood of a report containing at least one comission error (retained statements from a standardized report that contradict the dictated findings or impression) was also significantly reduced from 3.9 to 0.8 % (p=0.0007).

The remaining error subtype error rates are summarized in Table 4.

Discussion

Following the creation and implementation of department-wide standardized, structured reports, we observed a statistically significant decrease (p=0.024) in the percentage of reports with nongrammatical errors; decreasing from 33 to 26 % in the current study. The reduction in nongrammatical errors observed in our study is largely due to a reduction in comission errors and missense omission errors. Reduction in these two error types accounted for 80 % of the overall reduction in nongrammatical error rate (errors per report) observed in this study.

The likelihood of at least one comission error occurring in a report was significantly decreased from the prior study (p=0.0007). We believe that this decreased error rate is directly related to the creation of structured reports particularly because as new reports were created, special attention was paid to avoid constructing reports that would contain elements that would be likely to be erroneously retained when dictating. An example of this was in constructing the report for a CT of the abdomen. Our original standard report for a CT of the abdomen/pelvis contained the phrase: "The appendix is visualized and is normal." This phrase was part of a large grouping of pertinent negatives and often was left in the report, even when the appendix was abnormal or not identified. In the new structured report, there is a specific section for the appendix. Instead of prepopulating the normal appendix statement in this section, it is left blank and the radiologist must choose the appropriate selection from a structured pick-list or dictate the abnormal findings. While we believe that careful construction of the structured reports is the cause of the decrease in comission errors, it is possible that this finding is secondary to the radiologists' increased familiarity with the structured reports.

There was a 69 % decrease in the likelihood of at least one missense omission error occurring in a report. This change was also significant (p=0.0175). There are several potential reasons why this type of error decreased. First, this may represent an improved familiarity of the types of words that get omitted when dictated by the radiologists. If the radiologist is more in tune with this type of error, he or she will be more likely to correct it before it reaches the final report. Second, the improvement in the missense omission error rate may be due to the structured reports. It is possible that the reports were constructed in such a way that phrases are being dictated more commonly than sentences. It is unlikely that the decrease in this type of error is a result of improved speech recognition by the software as the number of translation errors per report is unchanged from the prior study and identical dictation software was used.

Overall, there was no change in the number of errors per report between the two studies (p=0.43). It is not surprising that the total error rate in radiology reports was not significantly changed as uncommon abnormal findings are still freely dictated in our department. We allow free-dictation for uncommon abnormal findings for two reasons: first, with current technology, it is impossible to create structured language for all variations of abnormal; and second, prior studies have shown a lack of improvement in, and sometimes worsening, clarity of reports generated by purely structured reporting systems [9].

Grammatical errors (not including sentence fragments) remain the most common type of error in radiology reports at our institution. While the grammatical error rate increased in the current study, this difference was not significant. Because free-prose dictation will be necessary to describe uncommon abnormal findings for the foreseeable future, this type of error will remain problematic until an error cue for grammatical errors is added to speech recognition software. Even though the grammatical error rate is high, it should be noted that no grammatical errors were identified in structured content; all grammatical errors occurred in freely dictated abnormal findings.

Translational errors of all types (missense, nonsense, and other) are usually due to the speech recognition software. It is not surprising that this type of error was unchanged as there was no change in the software during our study. As newer versions of the medical speech engines are released, we anticipate that this type of error will decrease. Errors cues could

also be used as a method to decrease the frequency of translational errors. In this scenario, the speech recognition software would identify and underline words that were more likely to be incorrectly translated.

Overall, the percentage of reports containing at least one error was not significantly changed (p=0.176). The 37 % of reports containing an error in the current study remains higher than the 4.8–22 % reported in prior studies [1–3]. Several factors likely contribute to the higher error rate found in this study. First, we broadly defined and categorized errors into one of eleven different error types. This represents a more granular classification of errors and allows us to identify types of errors that were not considered on prior studies. This is particularly true with grammatical errors. While counting grammatical errors as errors in a report may be controversial, we believe evaluating this error type is important as there has been an increased emphasis placed on the radiologist's ability to communicate not only with the ordering healthcare provider but also with the patient.

Directly comparing the text report to the audio file is a potential second reason that our calculated error rate is higher than prior studies. Listening to the audio files allows us to identify and classify errors that cannot be identified by only reading the text report. These error types include many translational errors such as errors of omission (errors that occurred when a spoken word was not transcribed by the speech recognition system) and typographical errors.

Our study design may have also led to an increased error rate compared to other studies. In prior studies, emphasis was placed only on "significant" error types, which were defined as errors that could potentially lead to the conveyance of incorrect or confusing information [1, 3]. In this study, all error types were evaluated, regardless of their potential significance. We believe that every effort should be made to eliminate all error types [10] in order to foster better communication, and therefore better patient care. Like others, we believe the radiology report is the most important basis on which radiologists are judged by referring clinicians and patients [7].

There are several limitations to this study. The first limitation is manual error logging, which introduces the possibility of both over- or undercounting errors. While an automated error detection system may be able to more accurately detect certain types of errors, we believe that a system of this type would have difficulty classifying errors of omission and comission [4].

A second limitation is the reliance on data from a prior study to drive comparative statistics. A more elegant study would have employed a direct comparison of reporting styles over a shorter time period and have reviewers review the different reports in a completely random (with regard to time points) manner. We attempted to account for this limitation as much as possible by having the same reviewers score each radiology report using the same technology. In addition, radiologists used the same version of the software to dictate reports. Because it took several months to roll out the department-wide standardized, structured reports [8], we could not perform a true A versus B comparison.

Finally, this study is limited in that we were not able to compare structured reports to freely dictated prose reports. In our initial, baseline study, radiologists used and preferred the department standard, nonstructured report templates. These reports contained normal findings described in prose. This comparison can be seen as both detrimental and beneficial depending on the error type. We would expect certain errors (such as grammatical errors and translational errors) to decrease in frequency as the predetermined content increases. At the same time, we would anticipate that certain errors (such as errors of comission) increase as the predetermined content increases. In the end, we believe that our study demonstrates a method that can be used to test different types of reports over time. This potentially iterative process can help us to identify the reporting process and report structure that leads to the fewest number of errors.

Conclusions

A systematic approach to create and implement department-wide structured reports resulted in a statistically significant decrease in the percentage of reports containing a nongrammatical error, missense omission error, and comission error. While our results may or may not be generalized to other radiology departments, they do show that a concerted effort to decrease these error types in radiology reports by increasing structured report content can be successful.

Disclosures None.

References

1. Quint LE, Quint DJ, Myles JD: Frequency and spectrum of errors in final radiology report generated with automatic speech recognition technology. J Am Coll Radiol 5:1196–1199, 2008
2. McGurk S, Brauer K, Macfarlane TV, Duncan KA: The effect of voice recognition software on comparative error rates in radiology reports. Br J Radiol 81:767–770, 2008
3. Kanal KM, Hangiandreou NJ, Sykes AMG, Eklund HE, Araoz PA, Leon JA, Erickson BJ: Evaluation of the accuracy of continuous speech recognition software system in radiology. J Digit Imaging 13:211–212, 2000
4. Hawkins CM, Hall S, Hardin J, Salisbury S, Towbin AJ: Prepopulated radiology report templates: a prospective analysis of error rate and dictation time. J Digit Imaging 25(4):504–511, 2012

5. Kahn CE, Heilbrun ME, Applegate KE: From guidelines to practice: how reporting templates promote the use of radiology practice guidelines. J Am Coll Radiol 10:268–273, 2013
6. Schwartz LH, Panicek DM, Berk AR, Li Y, Hricak H: Improving communication of diagnostic radiology findings through structured reporting. Radiology 260(1):174–181, 2011
7. Reiner BI: The challenges, opportunities, and imperative of structured reporting in medical imaging. J Digit Imaging 22(6): 562–568, 2009
8. Larson DB, Towbin AJ, Pryor RM, Donnelly LF: Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. Radiology 267(1):240–250, 2013
9. Johnson AJ, Chen MY, Zapadka ME, Lyders EM, Littenberg B: Radiology report clarity: a cohort study of structured reporting compared with conventional dictation. J Am Coll Radiol 7:501–506, 2010
10. Reiner BI, Knight N, Siegel EL: Radiology reporting, past, present, and future: the radiologist's perspective. J Am Coll Radiol 4:313–319, 2007