Current Problems in Diagnostic Radiology 48 (2019) 524–530

ELSEVIER

Current Problems in Diagnostic Radiology

journal homepage: www.cpdrjournal.com

Using a Natural Language Processing and Machine Learning Algorithm Program to Analyze Inter-Radiologist Report Style Variation and Compare Variation Between Radiologists When Using Highly Structured Versus More Free Text Reporting

Lane F. Donnelly, MD[a,1,2,*], Robert Grzeszczuk, PhD[b], Carolina V. Guimaraes, MD[a,1], Wei Zhang, PhD[a], George S. Bisset III, MD[a,3]

[a] *Department of Radiology, Texas Children's Hospital, Houston, TX*
[b] *InContext, Houston, TX*

ABSTRACT

Purpose: To use a natural language processing and machine learning algorithm to evaluate inter-radiologist report variation and compare variation between radiologists using highly structured versus more free text reporting.
Materials and Methods: 28,615 radiology reports were analyzed for 4 metrics: verbosity, observational terms only, unwarranted negative findings, and repeated language in different sections. Radiology reports for two imaging examinations were analyzed and compared – one which was more templated (ultrasound – appendicitis) and one which relied on more free text (chest radiograph – single view). For each metric, the mean and standard deviation for defined outlier results for all dictations (individual and group mean) was calculated. The mean number of outlier metrics per reader per study was calculated and compared between radiologists and between the two report types. Wilcoxon rank test and paired Wilcoxon signed rank test were applied. The radiologists were also ranked based on the number of outlier metrics identified per study.
Results: There was great variability in radiologist dictation styles – outlier metrics per report varied greatly between radiologists with the maximum 10 times higher than the minimum score. Metric values were greater ($P < 0.0001$) on the standardized reports using free text than the more structured reports.
Conclusions: The algorithm successfully evaluated metrics showing variability in reporting profiles particularly when there is free text. This variability can be an obstacle to providing effective communication and reliability of care.

Introduction

The radiology report is the documented product of radiologists' efforts and the formal and primary means by which radiologists communicate with referring care providers.[1] Multiple previous studies have shown: free dictation styles vary greatly between radiologists; these variations adversely affect report clarity; and standardization of report structure and language improves communication.[2-6] Unfortunately, most of the data supporting the use of standardized and structured reports is based on limited audits and subjective grading by reviewers.[7-11]

Many radiology departments have moved to department sanctioned standardized, structured reports.[1-12] Some of these reports are very structured with the majority of entered data coming from pick lists of prechosen options with very little opportunity for free text entry, such as would be the case for interpretation of mammographic studies using BIRADS.[13,14] However, most commonly, structured radiology reports have structured headers, and possibly subheaders, with defined language for normal studies that rely on free text dictation in the appropriate section when abnormalities are present.[2, 3]

Natural Language Processing (NLP) is an area of Artificial Intelligence (AI) that uses computer algorithms to extract meaning from text. While speech recognition uses transcription software to take human voice as input and transform it into strings of characters, NLP takes the resulting text and tries to make sense of it. The power of NLP lies in its ability to analyze the vast amounts of, what is commonly referred to as, "unstructured data" and make decisions based on the content. Common NLP tasks can be as diverse as automatic text classification (e.g., distinguishing valid emails from spam), sentiment analysis (e.g., making stock trading decisions based on positive or negative aspect of a business news wire), machine translation (e.g., translating Web pages and street signs from Mandarin to English),

Disclosures: Robert Grzeszczuk is the CEO of InContext. All other authors have no disclosures.

*Reprint requests: Lane F. Donnelly MD, Stanford University School of Medicine, Stanford Children's Health, 180 El Camino Real Ste M384, Mail Code: 5885, Palo Alto, CA 94304.

E-mail address: lanefd@stanford.edu (L.F. Donnelly).

[1] Current address: Department of Radiology, Stanford University, Stanford, CA.

[2] Current adress: Center for Pediatric and Maternal Health, Lucile Packard Children's Hospital at Stanford.

[3] Children's Hospital, New Orleans, LA 70118.

https://doi.org/10.1067/j.cpradiol.2018.09.005
0363-0188/© 2018 Elsevier Inc. All rights reserved.

automatic triage of emergency cases based on a clinical report content, to customer support conversational agents (or chatbots) that can answer customer's questions and resolve common problems.

NLP has rich and storied history dating back to 1950, when it relied mostly on rule-based linguistic approaches, but has since evolved toward statistical methods and methodology borrowed from Machine Learning (ML). Traditionally, text undergoes lexical, morphological, and grammatical analysis, which is then followed by semantic analysis, all of which yield numerous features that can then be used by various ML algorithms to extract meaning. In recent years, artificial neural networks (ANNs) have been gaining steady popularity for many common NLP tasks, like Named Entity Recognition (NER), Semantic Role Labeling (SRL), summarization, word embeddings, and others. Deep Learning (DL) refers to use of computationally intensive ANNs with many hidden layers.

The purpose of this manuscript is to use a NLP and machine learning algorithm to evaluate inter-radiologist report variation and compare variation between radiologists using highly structured versus more free text reporting.

Methods

IRB approval was obtained for this study. A study was performed to evaluate differences in radiology reports between radiologists, comparing a highly structured radiology report with predominantly “pick list” choices to a templated radiology report where default language describing the normal state is replaced by free text language when abnormalities are present. The department is an academic department with both radiology residents and fellows. However, given department workflow, the majority of studies are dictated directly by faculty.

The department uses Powerscribe 360 dictation software (Nuance Communications, Burlington, MA). System-wide standardized reports are deployed in the department using the features of the dictation software, including pick-lists. Standardization includes defined sections and headers (*Exam, Clinical History, Technique, Comparison, Findings, Impression*), standardized subsections (where appropriate), standardized language for normal studies, and standardized language for common abnormal exams. There are approximately 250 such standardized reports for various imaging studies. For atypical and complex abnormal examinations, free dictation of text into the appropriate report section or subsection occurs. Depending upon the nature of the imaging examination, some of the reports are highly structured with almost all data being entered by pick-list and very little free text. Other imaging examinations have standardized headers with default normal language which is most commonly replaced by free text when abnormalities are present.

The department partnered with a software developer (*InContext*, Houston, Texas) to develop a program to evaluate various aspects of radiology reports. All reports generated during the month of October 2015 were de-identified and loaded into the developed software program. This included 28,615 radiology reports. The developed software program used NLP and machine learning algorithms in order to identify and classify statements within the reports.

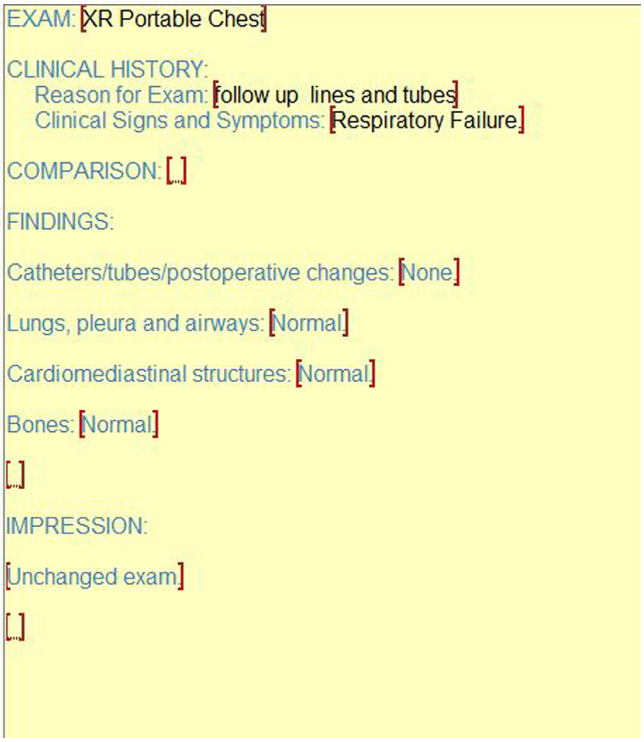
Data analytic algorithms were applied to characterize elements of reporting style. Specifically, text of each report was pre-processed using conventional low-level lexical tools, that included splitting it into individual tokens and stemming, followed by syntactical feature tagging, including Parts of Speech (POS) labeling, as well as semantic tagging using publically available ontologies (RadLex, RSNA) to identify clinically relevant terms (e.g., anatomical structures, observations, findings, etc.) to be used as features. Finally, a Long Short Term Memory Recurrent Neural Network (LSTM RNN) was used as a shallow labeler to group tokens into semantically related chunks, which were then fed into a hybrid Dynamic Programming/Rule-Based Sematic Role Labeling (SRL) algorithm. Open source toolkits were used throughout the process, including Apache OpenNLP, Apache Lucene, Stanford NLP, Keras, and Google's TensorFlow. Classifiers were trained on a 20% subset of the entire set of reports. The resulting standalone solution was applied offline to a set of deidentified historical reports producing numerical values for each of the metrics described per each report analyzed.

Characteristics analyzed included verbosity, observational terms only, unwarranted negative findings, and repeated language in different sections. Verbosity was defined as the percentage of sentences for a particular reader that were greater than 30 words in length. Observations-only reports were defined as reports where the *Impression* section contained only sentences with descriptive terms and lacked any interpretive terms, for example “*Impression*: There is a small amount of fluid.” Unwarranted negative findings were defined as reports containing statements like “*Findings*: Visualized portions of the bladder, gallbladder, and kidneys are within normal limits.” which were not in a direct response to a clinical question posed by the referring physician (e.g., “Rule out pneumonia.”). Although not always true, it can be argued that at times such statements are superfluous and they should not be included in the report if they refer to unremarkable structures thus providing little to no diagnostic value to the referring physician. Repeats were defined as identical or very similar language included in two sections, in the *Findings* and *Impression*, for example. Repeats were subcategorized as verbatim or fuzzy. Verbatim repeats were defined as word-for-word reiterations, where the same sentence was re-used in multiple places in the same report (e.g., potentially copied and pasted). Fuzzy repeats were defined as sentences that were paraphrased with only minor, stylistic modifications that did not affect the clinical significance of the statement. For example “A fairly long appendix is noted with its tip in the subhepatic region” restated as “An elongated appendix is noted with its tip in the subhepatic region.” elsewhere in the report.

Radiology reports from two imaging examinations were compared: those for interpretation of *single view chest radiography* and those for *ultrasound abdomen limited: evaluate for appendicitis*. The number of examinations included in the analysis included: *single view chest radiography* (3261) and *ultrasound abdomen limited: evaluate for appendicitis* (967). These two examinations were chosen because one represents a highly structured template report with little free text and one is structured but more reliant on free text. The single view chest radiography report is structured into a section with standardized language for normal, but it relies on free text when abnormalities are present (Fig 1). As most of these examinations are portable chest radiographs on patients in the intensive care units, most exams are abnormal. The ultrasound appendicitis report represents a very standardized structured report with mostly pick lists and very little free text (Fig 2).

The two reports were evaluated for all of the previously described metrics: Verbosity, observational terms only, unwarranted negative findings, and repeated language in different sections. When a metric was deviated from the expectation, it was considered an outlier. Each metric had a specific definition of an outlier (verbosity: outlier = sentence greater than 30 words, observational terms only: Impression had observational [not interpretive] terms only, etc. as described above). The mean of outlier results for each metric per study was calculated for each radiologist who dictated that particular exam.

For each metric, the mean and standard deviation of all dictations were calculated for ultrasound appendicitis reports and single view chest radiograph reports. The mean number of outlier metrics per reader per study was calculated and compared between radiologists and between the two report types. A Wilcoxon rank test was applied to evaluate for statistically significant differences. In addition, for the radiologists who had dictated reports for both *single view chest radiography* and those for *ultrasound abdomen limited: evaluate for appendicitis*, a paired test – Wilcoxon signed rank test, was applied to take account of the dependence in the data. This paired test allowed each radiologist to serve as their own control.



EXAM: [XR Portable Chest]

CLINICAL HISTORY:
Reason for Exam: [follow up lines and tubes]
Clinical Signs and Symptoms: [Respiratory Failure]

COMPARISON: [...]

FINDINGS:

Catheters/tubes/postoperative changes: [None]

Lungs, pleura and airways: [Normal]

Cardiomediastinal structures: [Normal]

Bones: [Normal]

[...]

IMPRESSION:

[Unchanged exam]

[...]

FIG 1. Image of standardized report for *Single View Chest Radiography*. The red boxes denote changeable fields. Note that "Normal" is the default for the subsections within FINDINGS. However, as most of the exams are performed on inpatients, and most commonly patients in critical care with life support apparatus, the findings are uncommonly normal. When there are abnormalities, free text dictation is used to describe the abnormalities. (Color version of figure is available online.)

In addition, to create a quantitative depiction of the potential variability between radiologists' dictation styles, for each of the 4 evaluated parameters, each radiologist was ranked from 1 (given to the radiologist with the least number of outlier metrics) up to the number of dictating radiologists (for the radiologist with the highest number of outlier metrics). For each radiologist, a score was created by adding the rankings together for each of the 4 evaluated metrics. This was done separately for both *single view chest radiography* and those for *ultrasound abdomen limited: evaluate for appendicitis* reports and displayed in bar graph format to compare the height of those bars between radiologists.

Results

There were 28 radiologists who created reports for *single view chest radiography* evaluated in the study. There were 23 radiologists who created reports for *ultrasound abdomen limited – appendicitis* evaluated in the study. There were 20 radiologists who had dictated both a *single view chest radiography* and an *ultrasound abdomen limited – appendicitis*.

The analysis for statistically significant differences in means for the evaluated parameters is shown in Table 1. The results of the paired Wilcoxon signed rank test performed on the 20 radiologists who had dictated both a *single view chest radiography* and an *ultrasound abdomen limited – appendicitis* is shown in Table 2. Both analyses show that there were statistically significant differences in the total number of normalized measures between the 2 imaging studies as well as between the groups for the quality metrics *observations only* and *repetitions*. Fig 3 illustrates the differences in *repetition* between the two imaging study reports. Results were not significant for the metrics *verbosity* or *unwarranted negative results*. Despite the negative statistical significance between verbosity between the two types of reports, Fig 4 illustrates the variability between radiologists in the frequency of verbosity.

Concerning creating a quantitative depiction of the potential variability between radiologists' dictation styles, for the *single view chest radiography* as there were 28 reading radiologists, for each of the 4 parameters, the radiologist with the lowest number of identified items was scored a 1 and the highest was scored a 28 with those in between being numerically ranked 2 through 27. Therefore, the range of potential scores was between 4 (if same radiologist having the lowest ranking in all 4 items) and 112 (if same radiologist having the highest ranking in all 4 items). Results of the ranking are shown in Fig 5. There is a large variability in scores with the lowest scoring radiologist having been ranked a 9 and the highest being ranked 81. This is approximately a 10-fold difference in scores.

For the *ultrasound abdomen limited – appendicitis* there were 23 reading radiologists. For each of the 4 parameters, the radiologist with the lowest number of identified items was scored a 1 and the highest was scored a 23 with those in between being numerically ranked 2 through 22. Therefore, the range of potential scores was between 4 (if same radiologist having the lowest ranking in all 4 items) and 92 (if same radiologist having the highest ranking in all 4 items). Results of the ranking are shown in Fig 6. There is a large variability in scores with the lowest scoring radiologist having been ranked a 12 and the highest being ranked 80. This is approximately a 6-fold difference in scores.

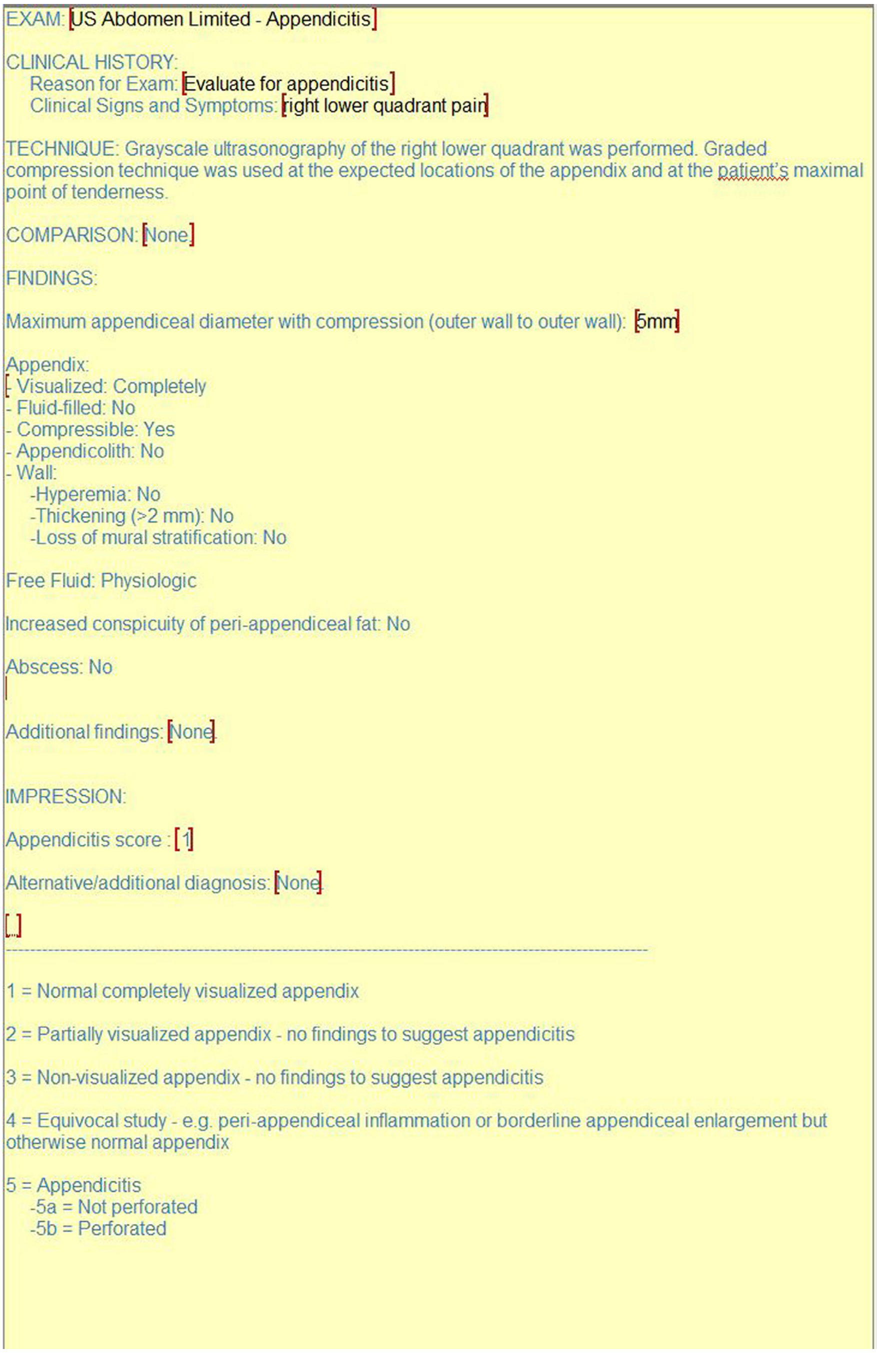
Discussion

A number of previous studies have demonstrated that implementation of standardized and structured reports improves communication with referring physicians.[1-12] This improvement is thought to be related to the decrease in variability in dictation style between radiologists in the free text environment.[1-12] This lack of data uniformity and structure related to nonstandardized lengthy narratives can hinder clear communication.[6] Valuable information is trapped in the prose of unstructured reports.[6] This study demonstrates that NLP and machine learning algorithms can be used to evaluate significant volumes of radiology reports for metrics which could be used for tasks such as quality control, teaching, and as feedback and learning materials for practicing radiologists.

This study also demonstrates and confirms that there is high variability in radiologist dictation styles based on the parameters evaluated. For reports of *single view chest radiograph*, the style variability rankings of radiologist related to the number of outlier metrics per evaluated parameter ranged highly with a 10-fold difference between the lowest scoring radiologist at 9 and the highest at 81 (Fig 5). Both the low and high actual values were near the respective possibly obtainable low and high values, demonstrating marked variation in the styles of those respective radiologists. Findings were similar for the dictated reports for *ultrasound abdomen limited – appendicitis* (Fig 6), despite that report being more templated.

This study also demonstrates that a more structured templated report with less free text, (*ultrasound abdomen limited: evaluate for appendicitis)* had statistically significantly less mean values for metrics which met outlier criteria, as compared to a templated report that relied on more free text dictation (*single view chest radiography).*

The previous studies demonstrating that implementation of standardized and structured reports improves communication with referring physicians were all set up similarly and have common limitations.[1-12] First, in each of these studies, the analysis was conducted by limited audit,[7-11] often with numbers of evaluated reports constituting not a large sample size. The audits were done manually and labor intensive process is probably what was responsible for the relatively small sample sizes used. With the NLP and machine learning program, all radiology reports can be evaluated, eliminating the



EXAM: [US Abdomen Limited - Appendicitis]

CLINICAL HISTORY:
Reason for Exam: [Evaluate for appendicitis]
Clinical Signs and Symptoms: [right lower quadrant pain]

TECHNIQUE: Grayscale ultrasonography of the right lower quadrant was performed. Graded compression technique was used at the expected locations of the appendix and at the patient's maximal point of tenderness.

COMPARISON: [None]

FINDINGS:

Maximum appendiceal diameter with compression (outer wall to outer wall): [5mm]

Appendix:
[- Visualized: Completely
- Fluid-filled: No
- Compressible: Yes
- Appendicolith: No
- Wall:
-Hyperemia: No
-Thickening (>2 mm): No
-Loss of mural stratification: No

Free Fluid: Physiologic

Increased conspicuity of peri-appendiceal fat: No

Abscess: No
]

Additional findings: [None]

IMPRESSION:

Appendicitis score : [1]

Alternative/additional diagnosis: [None]

[...]

1 = Normal completely visualized appendix

2 = Partially visualized appendix - no findings to suggest appendicitis

3 = Non-visualized appendix - no findings to suggest appendicitis

4 = Equivocal study - e.g. peri-appendiceal inflammation or borderline appendiceal enlargement but otherwise normal appendix

5 = Appendicitis
-5a = Not perforated
-5b = Perforated

FIG 2. Image of standardized report for *Ultrasound Limited – Appendicitis*. The red boxes denote changeable fields. Pick list feature is utilized for the findings. The radiologist picks from score of 1-5 and the findings auto-populate. The only real opportunities for free text dictation are in field box for "Additional Findings" and in the field box for "Alternative/ additional diagnosis" in the impression. (Color version of figure is available online.)

TABLE 1
Comparison of the mean for outlier metrics per study between reports for ultrasound for appendicitis and portable chest radiography for all reading radiologists

	Ultrasound – Appy		Chest X-ray		
Normalized measure	N	Mean ± SD	N	M ean ± SD	P value
Observations only	23	0.0178 ± 0.0143	28	0.0505 ± 0.0304	0.0002
Unwarranted negative results	23	0.0132 ± 0.0207	28	0.0119 ± 0.0230	0.61
Repetitions	23	0.0074 ± 0.0111	28	0.0270 ± 0.0207	0.0007
Verbosity	23	0.0015 ± 0.0024	28	0.0016 ± 0.0027	0.34
Total	23	0.0399 ± 0.0283	28	0.0910 ± 0.0417	<0.0001

TABLE 2
Comparison of the mean for outlier metrics per study between reports for ultrasound for appendicitis and portable chest radiography paired for radiologists who had dictated both types of studies

	Ultrasound – Appy		Chest X-ray		
Normalized measure	N	Mean ± SD	N	Mean ± SD	P value
Observations only	20	0.0180 ± 0.0122	20	0.0586 ± 0.0237	<0.0001
Unwarranted negative results	20	0.0152 ± 0.0216	20	0.0107 ± 0.0197	0.5412
Repetitions	20	0.0060 ± 0.0062	20	0.0365 ± 0.0165	<0.0001
Verbosity	20	0.0017 ± 0.0025	20	0.0017 ± 0.0021	0.5619
Total	20	0.0409 ± 0.0238	20	0.1074 ± 0.0282	<0.0001

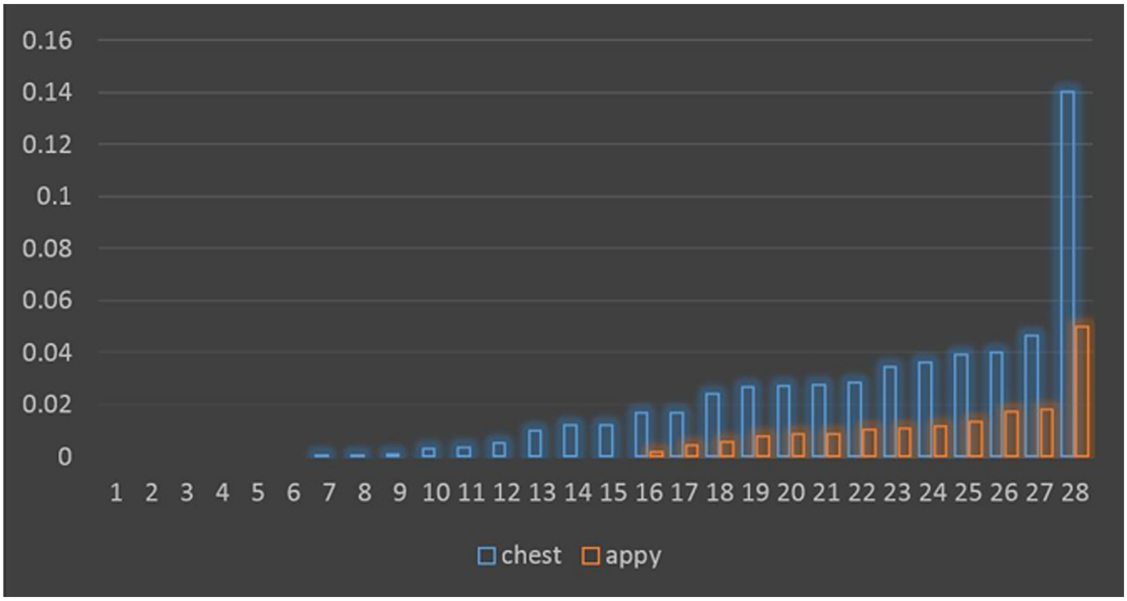


FIG 3. Comparison between mean numbers of outlier metrics related to *repetition* per study for individual radiologists. There was a statistically higher value for repetition on reports for *single view chest radiograph* as compared to on reports for *ultrasound abdomen limited – appendicitis*. Note that the values ranged from zero for some radiologists to 0.14 for the highest value on *single view chest radiograph*.

limitations of small sample sizes. Such a program also allows for monitoring and feedback generation on a continuous basis, rather than episodically related to intermittent evaluation by audit.

A second limitation is that most of the studies evaluating the effects of standardization were done by reviewers using subjective grading systems.[7-11] Although these varied in the degree of survey specificity, most were conducted by reviewers using a point-based grading. As an example, in one study in which referring physicians graded radiology reports, a 1-5 point scale was used to grade 6 elements: typographic errors, unclear language, discrepancies, too long, too short, clinical question not answered, template not uses, and relevant or appropriate recommendations were not made.[8] In another

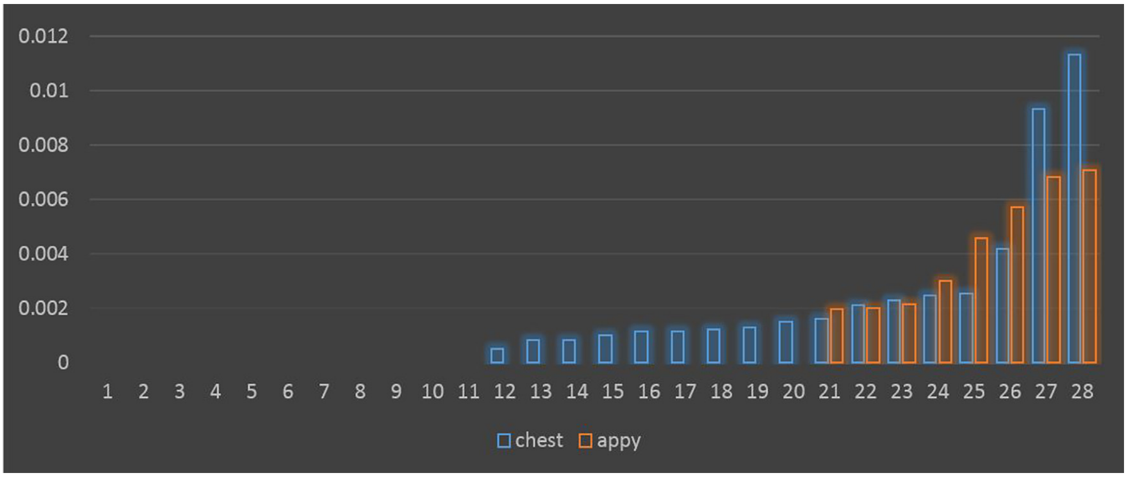


FIG 4. Comparison between mean numbers of outlier metrics related to *verbosity* per study for individual radiologists. Although there was not a statistically higher value for *verbosity* on reports for *single view chest radiograph* as compared to on reports for *ultrasound abdomen limited – appendicitis*, note variation between individual radiologists. Numerous radiologists had zero incidence of *verbosity*, as compared to 0.015 for the highest value on *single view chest radiograph*.

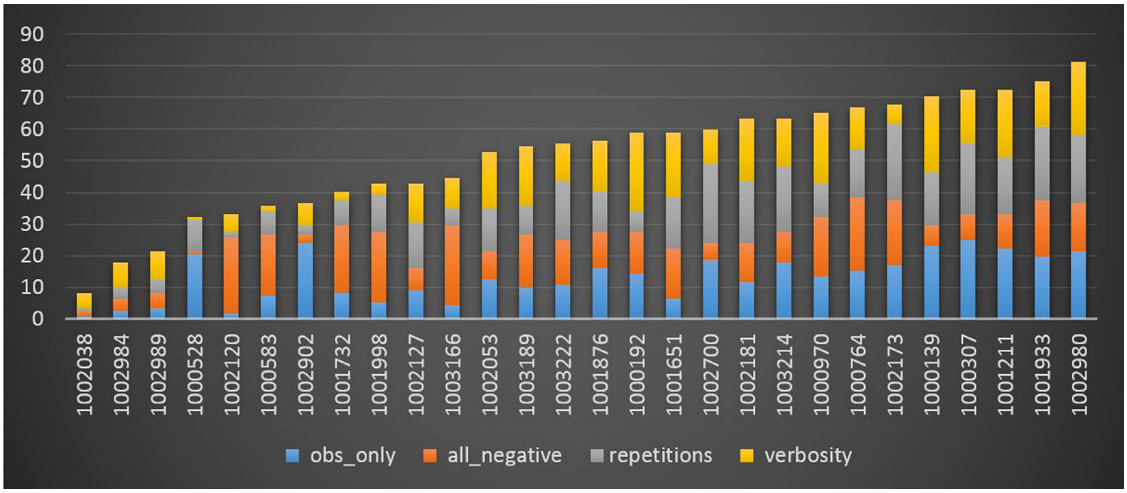


FIG 5. Depiction of variability between radiologists dictation styles on *single view chest radiograph*. Rankings for all 4 measured variables were ranked (1 for least, 28 for highest) and then added together for each individual radiologist. Numbers along x-axis = code for each radiologist. Note that the variation was great ranging from the lowest radiologist score at 9 to the highest score at 81. Obs_only = observations only, all_negative = unwarranted negative results.

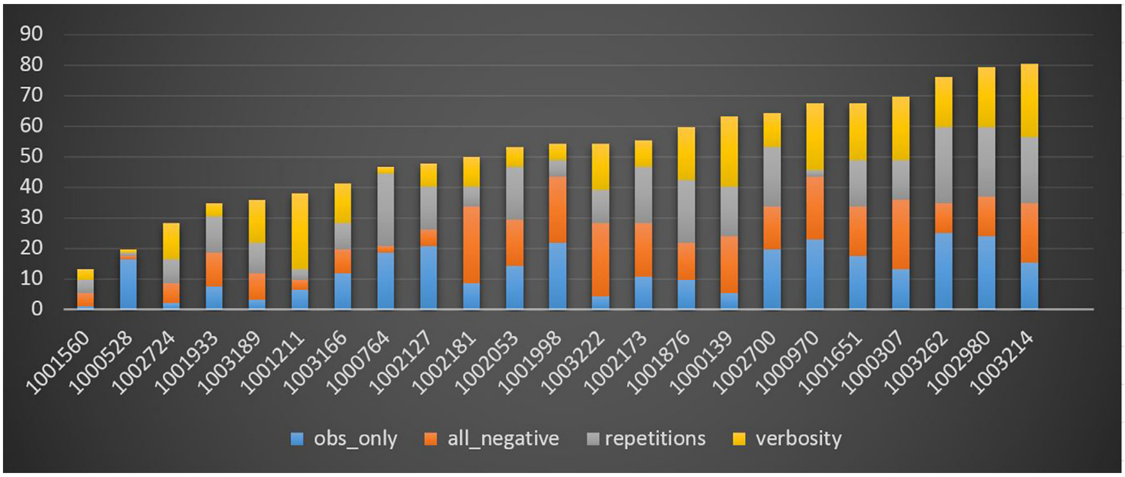


FIG 6. Depiction of variability between radiologists dictation styles on *ultrasound abdomen limited – appendicitis*. Rankings for all 4 measured variables were ranked (1 for least, 23 for highest) and then added together for each individual radiologist. Numbers along x-axis = code for each radiologist. Note that the variation was great ranging from the lowest radiologist score at 12 to the highest score at 80. Obs_only = observations only, all_negative = unwarranted negative results.

study, reports were graded based on a report completeness scale (incomplete, partially complete, or very complete) and on a report effectiveness scale (ineffective, partially effective, and very effective).[11] These studies have been helpful in showing the usefulness of standardized reports, as opposed to free text reports, and used the best tools available to them at the time. However, the machine learning program can evaluate reports using both more metrics and a more consistent and less subjective algorithm, as it does not rely on inconsistencies between human reviewers or inconsistent application of a point grading system. Also, as stated previously, analysis by the machine learning program is not limited by sample size.

The radiology community has begun to move to department sanctioned standardized reports because of the data showing the associated improved communication.[1-14] However, for many standardized, templated standardized reports – headers, subheaders, and language for normal studies are often defined and complex abnormalities are addressed through free text.[2, 3] This study shows that although those standardizations may have improved communication as compared to complete individual customization and free text dictation, there is still much variability in radiologists' dictations even when using the standardized templates. An argument can be made that movement to more templated reports with pick lists for abnormalities should be implemented whenever possible to minimize free text dictation and help decrease this variability. However, even in such templated reports where free text is minimal, there still lies opportunity for radiologists to demonstrate variability.

Future studies should evaluate whether the providing of this type of data, comparing an individual radiology provider to their peers would be helpful to altering outlier behavior and driving further standardization. For example, would the radiologist who is an outlier in terms of verbosity compared to the remainder of the group change their dictation style if they were given the feedback on an ongoing basis showing that their reports were different than their peers? Would it help trainees develop more effective communication habits in their radiology reports?

There are a number of limitations to this study. The metrics which were chosen had arbitrary cut offs defined as to what was considered an outlier. Also, the comparison of reports from the chest radiographs as compare to those of the ultrasound report were chosen because each represented the extreme in standardized template versus standardized headers with free text but each represents very different clinical scenarios which could affect what types of language is appropriate. That being said, this study does show that an algorithm can successfully be used to evaluated language based report metrics and that those metrics do document variability in reporting profiles, particularly when there is free text.

References

1. Sistrom CL, Langlotz CP. A framework for improving radiology reporting. J Am Coll Radiol 2005;2:159–67.
2. Larson DB, Towbin AJ, Pryor RM, et al. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. Radiology 2013;267:240–50.
3. Guimaraes CV, DeFlorio RM, Averill LW, et al. Implementation of standardized reports within a pediatric health care system with geographically dispersed sites. J Am Coll Radiol 2015;12:1293–5.
4. Dunnick NR, Langlotz CP. The radiology report of the future: a summary of the 2007 Intersociety Conference. JACR 2008;5:626–9.
5. Langlotz CP. Structured radiology reporting: Are we there yet? Radiology 2009;253:23–5.
6. Durack JC. The value proposition of structured reporting in interventional radiology. AJR 2014;203:734–8.
7. Schwartz LH, Panicek DM, Berk AR, et al. Improving communication of diagnostic radiology findings through structured reporting. Radiology 2011;260:174–81.
8. Gunn AJ, Alabre CI, Bennett SE, et al. Structured feedback from referring physicians: a novel approach to quality improvement in radiology reporting. AJR 2013;201:853–7.
9. Wildman-Tobriner B, Allen BC, Bashir MR, et al. Structured reporting of CT enterography for inflammatory bowel disease: Effect on key feature reporting, accuracy across training levels, and subjective assessment of disease by referring physicians. Abdom Radiol 2017, https://doi.org/10.1007/s00261-017-1136-1. Apr 9 [Epub ahead of print].
10. Collard MD, Tellier J, Chowdhury I, et al. Improvement in reporting skills of radiology residents with a structured reporting curriculum. Acad Radiol 2014;21:126–33.
11. Marcovici PA, Taylor GA. Structured radiology reports are more complete and more effective than unstructured reports. AJR 2014;203:1265–71.
12. Kahn CE, Langlotz CP, Burnside ES, et al. Toward best practices in radiology reporting. Radiology 2009;252:852–6.
13. Berg WA, D'Orsi CJ, Jackson PV, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? Radiology 2002;224:871–80.
14. Lehman CD, Miller L, Rutter CM, et al. Effects of training with the American College of Radiology Breast Imaging Reporting and Data System lexicon on mammographic interpretation skills in developing countries. Acad Radiol 2001;8:647–50.