Title

Between Always and Never: Evaluating Uncertainty in Radiology Reports Using Natural Language Processing.

Permalink

https://escholarship.org/uc/item/3f22329f

Journal

Journal of Digital Imaging, 33(5)

Authors

Callen, Andrew

Dupont, Sara

Price, Adi

et al.

Publication Date

2020-10-01

DOI

10.1007/s10278-020-00379-1

Peer reviewed

Journal of Digital Imaging (2020) 33:1194–1201
https://doi.org/10.1007/s10278-020-00379-1

ORIGINAL PAPER

Between Always and Never: Evaluating Uncertainty in Radiology Reports Using Natural Language Processing

Andrew L. Callen[1] · Sara M. Dupont[2] · Adi Price[3] · Ben Laguna[3] · David McCoy[3] · Bao Do[4] · Jason Talbott[3] · Marc Kohli[3] · Jared Narvid[3]

Received: 12 December 2019 / Revised: 10 June 2020 / Accepted: 23 July 2020 / Published online: 19 August 2020

Abstract

The ideal radiology report reduces diagnostic uncertainty, while avoiding ambiguity whenever possible. The purpose of this study was to characterize the use of uncertainty terms in radiology reports at a single institution and compare the use of these terms across imaging modalities, anatomic sections, patient characteristics, and radiologist characteristics. We hypothesized that there would be variability among radiologists and between subspecialities within radiology regarding the use of uncertainty terms and that the length of the impression of a report would be a predictor of use of uncertainty terms. Finally, we hypothesized that use of uncertainty terms would often be interpreted by human readers as "hedging." To test these hypotheses, we applied a natural language processing (NLP) algorithm to assess and count the number of uncertainty terms within radiology reports. An algorithm was created to detect usage of a published set of uncertainty terms. All 642,569 radiology report impressions from 171 reporting radiologists were collected from 2011 through 2015. For validation, two radiologists without knowledge of the software algorithm reviewed report impressions and were asked to determine whether the report was "uncertain" or "hedging." The relationship between the presence of 1 or more uncertainty terms and the human readers' assessment was compared. There were significant differences in the proportion of reports containing uncertainty terms across patient admission status and across anatomic imaging subsections. Reports with uncertainty were significantly longer than those without, although report length was not significantly different between subspecialities or modalities. There were no significant differences in rates of uncertainty when comparing the experience of the attending radiologist. When compared with reader 1 as a gold standard, accuracy was 0.91, sensitivity was 0.92, specificity was 0.9, and precision was 0.88, with an F1-score of 0.9. When compared with reader 2, accuracy was 0.84, sensitivity was 0.88, specificity was 0.82, and precision was 0.68, with an F1-score of 0.77. Substantial variability exists among radiologists and subspecialities regarding the use of uncertainty terms, and this variability cannot be explained by years of radiologist experience or differences in proportions of specific modalities. Furthermore, detection of uncertainty terms demonstrates good test characteristics for predicting human readers' assessment of uncertainty.

Keywords Diagnostic uncertainty · Natural language processing

Background

Despite continued advances in medical knowledge, risk, uncertainty, and ambiguity remain a lasting part of medical practice [1]. Risk refers to events with known probability [2], while ambiguity deals with those events with unknown probability. Uncertainty usually refers to both risk and ambiguity [3]. The radiology report communicates the certainty or doubtfulness of particular diagnoses pertaining to the clinical question. Radiologists must effectively communicate risk and ambiguity as an essential

✉ Andrew L. Callen
andrew.callen@cuanschutz.edu

[1] Department of Radiology, University of Colorado Anschutz Medical Campus, Denver, CO, USA

[2] Sublte Medical Inc, Menlo Park, CA, USA

[3] Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, CA, USA

[4] Department of Radiology, Stanford University Medical Center, Stanford, CA, USA

part of their work. Yet in practice, the manner in which uncertainty (risk and ambiguity) is communicated in reports varies significantly among radiologists [4, 5]. Radiologists' reports have been shown to reflect their comfort/discomfort with uncertainty [6].

On the other hand, studies have shown that clinicians prefer reports which are certain and unambiguous [7–10]. Furthermore, clinicians often interpret the language of uncertainty in different ways [4]. Ambiguous language in radiology reports can lead to poor patient care [11, 12]. These data have buttressed warnings against excessive hedging [13] and have led institutions (and specialties) to create standardized lexicons [8, 14, 15].

However, there are inherent limitations to all diagnostic tests which must be conveyed. Failure to effectively communicate uncertainty (including differential diagnoses) can also lead to poor care and litigation [16].

It is now possible to measure the use of uncertainty terms among radiologists. The digitization of radiology reporting in the last two decades has allowed for the evaluation of large databases of archived reports. Algorithms for analyzing unstructured radiology reports, also known as natural language processing (NLP), have been used to analyze large numbers of reports for increasingly abstract concepts. An NLP algorithm was used in 2002 to analyze a large number of chest radiograph reports and acquire statistics such as sidedness of pathology [17]. Several studies have sought to develop search engines for radiology reports that can return reports with certain keywords and build teaching file databases automatically [18]. Extensive work has also been done to detect the presence and type of follow-up imaging recommendations [19, 20].

As a first step toward understanding how radiologists report uncertainty, whether variability in reporting does indeed exist, and the context of such variability, we wrote and validated an algorithm to automatically detect the use of uncertainty terms in unstructured radiology reports.

We hypothesized that there would be substantial variability among radiologists and between subspecialities within radiology regarding the use of uncertainty terms. Additionally, we hypothesized that there would be differences in the use of hedging terms between patient admission status (inpatient, emergency, and outpatient). We additionally hypothesized that the length of impression section of a report would be a predictor of use of uncertainty terms. Finally, we hypothesized that use of uncertainty terms would often be interpreted by human readers as "hedging." To test these hypotheses, we applied a natural language processing (NLP) algorithm to assess and count the number of uncertainty terms within radiology reports and to evaluate whether various features predict use of uncertainty terms.

Methods

Data Collection

This retrospective analysis of radiology reports was approved by the Committee on Human Research at our institution. No patient identifiers were included in the analyzed reports.

In total, 642,569 radiology reports were collected from a 5-year long period from 2011 through 2015 at a single center at our institution. This center is an academic radiology training site, with resident and fellow trainees previewing studies, followed by a formal review with an attending radiologist, who subsequently edits and signs the final reports. Procedural dictations from all subspecialties were excluded given their primary non-diagnostic context. Breast imaging dictations were excluded given their standardized lexicon. There were no nuclear medicine studies included, as the site does not have a dedicated nuclear medicine service.

Software Development

Uncertainty is defined as a fact or condition that lacks firm predictability or a condition of lacking certainty about a matter or circumstance [21]. We define uncertainty detection as identification of words or phrases which confer uncertainty to observations made by the radiologist. For example, in the sentence "findings could represent pneumonia," the adverb (and expression of uncertainty) "could" modify the verb "represent," conferring uncertainty to the subject, "findings."

Expressions of uncertainty can take the form of single words or expressions, such as "cannot" or "can not," providing a signal of uncertainty as in "cannot exclude pneumonia." Uncertainty can also manifest as a diagnostic differential, as in "soft tissue stranding may be due to infection or inflammation." Our program is implemented in both R and Python 2.7 and uses a regular expression approach to search report text and return matches to a database of words and expressions that convey uncertainty (Appendix Table 1). This database was generated via a process whereby terms within published lexicons were entered into a keywords dictation search engine to confirm their use by radiologists [22, 23].

Our NLP accepts unstructured report "Impressions" sections as input. The Heuristics are as follows: (1) eliminate all statements not included explicitly in the "Findings" or "Impression" section of the report (e.g., "I have personally read the above report"), (2) tokenize the remaining text, and (3) match and count words against the database of uncertainty terms.

Uncertainty Assessment—Statistical Analysis

We evaluated the proportion of reports including 1 or more uncertainty terms, compared with those which contained no uncertainty terms—our NLP output was thus dichotomized. Other variables included imaging modality (CT, MRI, ultrasound, radiography), the years of experience of the radiologist reading the study, the admission status of the patient (inpatient vs outpatient vs emergency), the subspecialty of the radiologist (abdominal, chest, musculoskeletal, neuroradiology, and non-specialist/diagnostic), and the length (word count) of the impression.

To evaluate the effect of each variable on uncertainty term usage, chi-square and Wilcoxon rank sum tests were used to assess group differences. Multivariate logistic regression was also performed including variables for subspecialty, admission status, and word count. For each variable, the proportion of uncertainty was represented on a bar plot along with the standard error and 95% confidence interval. The level of significance was adjusted according to the Bonferroni correction for multiple comparisons based on a type I risk alpha = 0.05 (level of significance = alpha/number of comparisons). Spearman's correlation was performed to assess the relationship of the frequency of hedging and years of reader experience.

Software Validation

Two attending radiologists at our institution (JT, 6 years of experience, and MK, 10 years of experience) who had no knowledge of the NLP algorithm and were given no details of how the NLP worked reviewed de-identified report impressions. These "gold-standard" readers (one neuroradiologist, one abdominal radiologist) were given very broad and basic instructions—that is, to determine whether the reporting radiologist was communicating uncertainty or hedging. Standard classification metrics between the presence of 1 or more uncertainty terms and the human readers' assessment was calculated via the confusion matrix including accuracy, sensitivity (or recall), specificity, precision, and F1-score.

Results

All 642,569 radiology report impressions from 171 reporting attending radiologists were collected from the 5-year period from 2011 through 2015. Of the 642,569 reports, 216,938 (33.8%) contained at least one uncertainty term. Of the 642,569 reports, 435,392 (67.8%) were from the general radiography/diagnostic service, 76,407 (11.9%) neuroradiology, 54,881 (8.5%) ultrasound, 42,883 (6.7%) abdominal, 24,258 (3.8%) chest, and 7752 (1.2%) musculoskeletal radiology. The most frequent uncertainty terms were "likely," "could," "versus," "non-specific," "suggestive", and "concerning" Fig. 1.

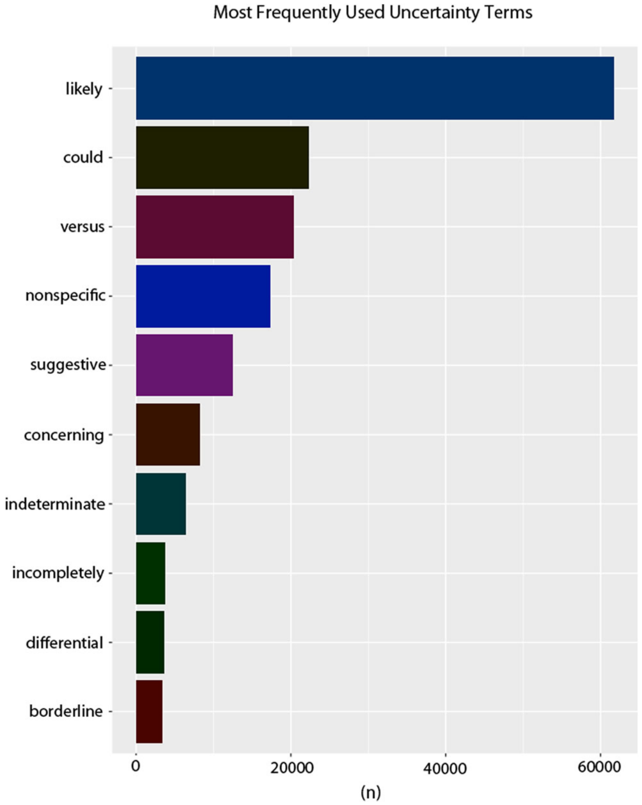

Fig. 1 Most frequently used uncertainty terms across all subspecialties, modalities, and patient admission status. X-axis, number of times individual uncertainty term appeared in the corpus of reports

Proportion of Uncertain Reports by Subspecialty/Section

The proportion of radiology reports utilizing uncertainty terms across anatomic subspecialties was evaluated. Thoracic and abdominal imaging reports demonstrated the greatest proportions of uncertainty, with 55.0% and 54.1% of reports containing uncertainty terms, respectively. Conversely, the reports of general radiologists contained the fewest uncertainty terms, with 29.5% of reports containing uncertainty terms. The differences in rates of expressed uncertainty between each subspecialty were statistically significant ($p < 0.003$), with the exception of neuroradiology versus musculoskeletal imaging (34.3% vs 34.8%, $p = 0.42$) and chest versus abdominal radiology (55.0% vs 54.1%, $p =$ 0.13). Multivariate logistic regression demonstrated lowest significant odds ratios for general diagnostic radiology (OR .78, CI .76–.79, $p < 2e\text{-}16$), neuroradiology (OR .83, CI .82–.86, $p < 2e\text{-}16$), ultrasound (OR .95, CI .93–.98, $p < .0006$), as well as musculoskeletal (OR .89, CI .84–.93, $p = 3.4e\text{-}6$).

Proportion of Uncertain Reports by Patient Admission Status

The proportion of radiology reports utilizing uncertainty terms was analyzed based on the admission status of the patient (i.e., inpatient, outpatient, or emergency). Inpatient radiology reports demonstrated the greatest proportion of expressed uncertainty, with 40.3% of reports containing uncertainty terms. Emergency room reports demonstrated uncertainty in 34.6% of reports. Outpatient reports demonstrated the least uncertainty, with 28.0% of reports containing uncertainty terms. The differences in these proportions were statistically significant ($p < .017$). However, in logistic regression, inpatient status (odds ratio of 1.1) was not significant (Fig. 2).

Proportion of Uncertain Reports by Reader Experience

The proportion of radiology reports utilizing uncertainty terms was analyzed according to years of experience of the attending radiologist who signed the report. At our hospital, there is a substantial number of per diem and volunteer clinical faculty who work infrequently. Therefore, in order to be most representative of the practice of our center, readers who read less than 1% of the total volume of reports were excluded from the analysis: 20 readers remained. Within those readers, experience ranged from < 1 to 41 years. There was no statistically significant correlation between rates of uncertainty and years of experience (Spearman's $R = 0.19$, $p = 0.44$).

Proportion of Uncertain Reports by Modality

In our hospital, all ultrasounds are read by our ultrasound section. Radiographs are read by a mixture of fellows and attending radiologists from all sections representing a non-specialized/diagnostic service. For this reason, proportionate differences in uncertainty among these services are represented in the above data on subspecialty/section. No significant differences in expressed uncertainty were present between MRI and CT.

Relationship Between Report Length and Uncertainty

Rates of uncertainty were compared by word count of the impression. The median word count of reports labeled as containing uncertainty terms was 36 (IQR 30), compared with 28 (IQR 26) in reports that were not identified as expressing uncertainty ($p < 0.001$). No differences in report length were found for subspecialty or patient class. Odds ratio for report length was significant, at 1.01 (CI 1.01–1.02, $p < 2\text{e-}16$).

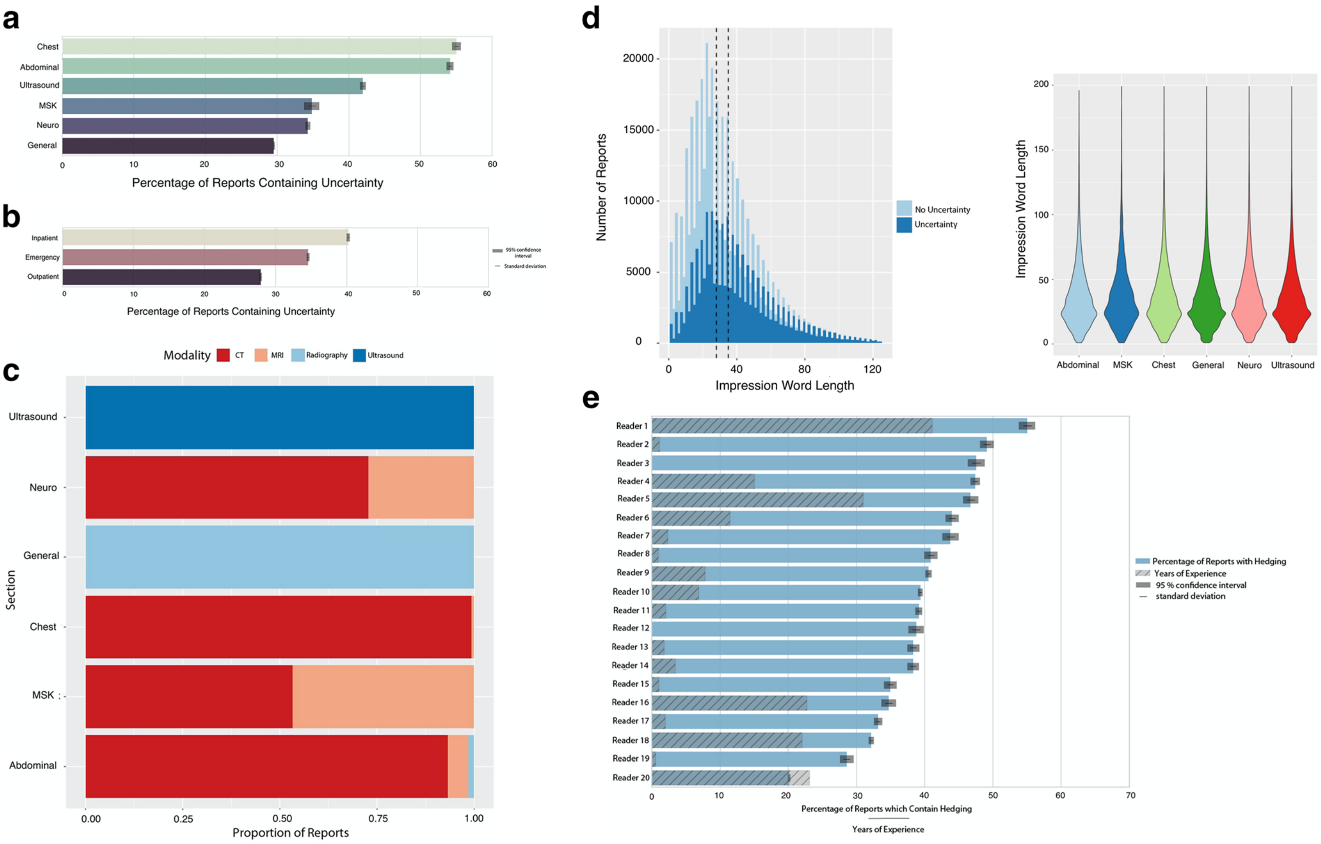

Fig. 2 (**a**) Proportions of reports using uncertainty terms by imaging subspecialty. (**b**) Proportion of reports using uncertainty terms by patient admission status. (**c**) Proportions of imaging modalities present for a given imaging subspecialty. Note that ultrasound and general radiology (radiography) only read a single modality. (**d**) Left, histograms of impression word length in reports deemed as expressing uncertainty versus those not deemed uncertain. Right, violin plots illustrating similar histograms of impression word length for each subspecialty

Software Validity with Radiologist Evaluation

Two attending radiologists reviewed report impressions of 453 and 931 reports, respectively (a randomly acquired sample of reports from the database), with 452 studies in common. For reader 1, of 453 reports analyzed, 193 were identified as expressing uncertainty. For reader 2, of 931 reports analyzed, 287 were identified as expressing uncertainty. Spearman's correlation coefficient was calculated for evaluation of inter-reader variability, with $R = 0.68$ ($p = 8.1e\text{-}62$). Joint probability was calculated to be 0.84. Cohen's kappa was calculated, $k = 0.67$.

The performance of the software was compared with each reader independently. When compared with reader 1 as a gold standard, accuracy was 0.91, sensitivity (or recall) was 0.92, specificity was 0.9, and precision was 0.88, with an F1-score of 0.9. When compared with reader 2, accuracy was 0.84, sensitivity (or recall) was 0.88, specificity was 0.82, and precision was 0.68, with an F1-score of 0.77.

Discussion

Radiologists must effectively communicate uncertainty as an essential part of their work. Unfortunately, terms used in radiology reports to express diagnostic uncertainty have poor agreement among physicians [5, 8]. In view of the association between uncertainty and diagnostic variation, over-testing, unnecessary surgery, increased hospitalization, and cost, some authors have called for standardization of radiology terminology [8, 24–26]. In this study, we sought to evaluate whether variability in the use of uncertainty terms exists and the related features associated with this differential use.

We find marked differences in the frequency of use of uncertainty terms both at the level of subspecialty and the individual radiologist. For example, musculoskeletal imaging demonstrated the lowest rate of uncertainty, 34%, significantly lower than chest, 55%. The odds ratios in multivariate regression were less than 1 and significant for musculoskeletal, neuroradiology, ultrasound, and general diagnostic radiology as opposed to chest and abdominal radiology (Fig. 3). And while true that longer reports are more likely to contain uncertainty terms, no subspecialty showed significant differences in report length. However, rather than viewing uncertainty as a malfeasance variably committed by some subspecialties, uncertainty may reflect a number of factors ranging from the included pathologies, the nature of anatomic imaging, and the nearness and frequency of findings in adjacent organs. In this light, uncertainty represents a substantial opportunity to identify pathologies among radiology subspecialties that may warrant dedicated lexica. This notion is supported by the measured improvements in diagnostic accuracy subsequent to structured lexica such as BI-RADS, TI-RADs, and LI-RADs [27, 28]. Additionally, some standardized lexica and follow-up guidelines inherently contain terms of uncertainty (e.g., the management of incidental pulmonary nodules) [29].

At the level of the individual, radiologists demonstrate marked variability in individual use of uncertainty terms, ranging from 20 to 55% of reports. This variability is not fully explained by subspecialty affiliation given the variability even within subspecialty.

Previous work has shown that natural language processing algorithms can be used to analyze a large number of reports for increasingly ambiguous and context-dependent concepts [30, 31]. We developed the current NLP algorithm to primarily address the lack of datasets including the use of uncertainty terms in radiology [22]. However, we also hypothesized that human readers use these terms to assess for uncertainty or "hedging." And while the words and phrases used by radiologists to communicate the likelihood of a diagnosis are often misinterpreted by clinicians [4, 5, 7], we hypothesized that the presence of uncertainty terms would predict humans readers' assessment that uncertainty is being expressed. In this way, we validated this program by assessing performance against human readers. We showed that the program detects uncertainty with high levels of accuracy, sensitivity, and specificity. These findings also lend support for simplified "uncertainty" lexicons [8]. Additionally, the possibility remains that radiologists can agree on the meaning and intent of ambiguous language more than their clinician readers.

Several aspects of our gold-standard data bear mention. First, it is not straightforward to define for all cases what represents "uncertainty" in a radiology report, nor is their universal agreement among radiologists in individual cases. However, many authors have focused on minimizing so-called ambiguous "hedging" vocabulary [13, 32] in radiology reports. While researchers have used sets of rules to define concepts such as "clinically important" reports [33], we found this rule-based approach problematic. To instruct our gold-standard readers to look for the "uncertainty" terms in our database would be equivalent to divulging the crux of the algorithm—more akin to asking our gold-standard readers to guess the output of the NLP. This would inflate our NLP's performance. Rather, we chose to define "uncertainty" more nebulously, instructing our gold-standard readers "to determine whether the radiologist is expressing uncertainty." This approach had the added benefit in revealing the inter-radiologist agreement on the concept of uncertainty. That our readers (in different subspecialties and reading across all subspecialty radiology reports) showed agreement ($k = .7$) with such vague instructions suggests that the concept of uncertainty is more concrete than expected. Moreover, the success of our approach—the use of regular expressions rather than more involved NLP algorithms—supports the notion that uncertainty vocabulary itself forms a working definition of the concept of expressed uncertainty [32].

Several limitations must be mentioned. Despite the overall large sample size, the participating radiologists represent a

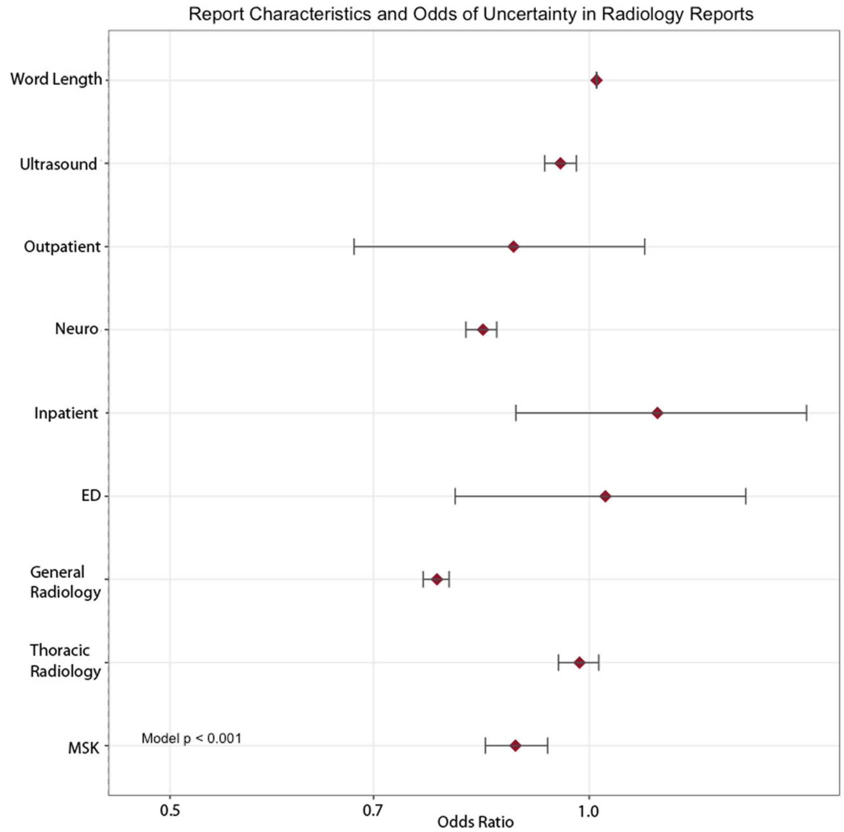

Fig. 3 Multivariate regression revealed statistically significant odds ratios greater than 1 for report word length and less than 1 for musculoskeletal, neuroradiology, ultrasound, and general diagnostic radiology as opposed to chest and abdominal radiology. Patient admission status was not statistically significant

single-institution experience. It is important to view conclusions in this context. That is to say, these findings require demonstrations of reproducibility at other sites.

A strength of the current study is the inclusion of a sufficiently large database (approximately 650,000 reports) across a 5-year period to observe broad patterns. Our analysis also demonstrates significant differences in uncertainty rates depending on the admission status of patients. Whether a patient was an inpatient, outpatient, or in the emergency room was correlated with significant differences in rates of uncertainty. Inpatients demonstrated the greatest rates of uncertainty, followed by emergency room patients. Outpatient radiology reports demonstrated the lowest rates of uncertainty. There are several possible explanations for this finding. First, inpatients in general are sicker, utilize more resources, and have more comorbidities than outpatients [34]. All of these factors might increase the likelihood that a radiologist would include uncertainty language in a report. In fact, these uncertainty data mirror imaging utilization rates comparing inpatient vs emergency room and outpatient visits [35]. Alternatively, radiologist reporting may change with the specificity of the clinical history given; it is possible that inpatients generate more nonspecific requests for imaging.

We hypothesized that longer impression statements would be associated with greater degrees of uncertainty, and indeed, our data bears this out. In addition to increasing the likelihood of the use of an uncertainty term, a longer impression statement often reflects a careful explanation on the part of the radiologist, where the diagnostic "answer" is not clear or succinct.

The level and depth of experience a radiologist brings to a case impact the specificity of his/her reporting. Data from mammography bears this out [36, 37]. We hypothesized that years of radiologist experience would influence uncertainty rates. Specifically, we calculated individual uncertainty rates for attending radiologists (percentage of reports labeled as uncertain) and do not find an association between years of experience and reduced uncertainty rate.

In the context of clinical decision-making, aversion to risk is a well-studied cognitive bias but has not been studied with regard to individual radiologist reporting practices [38]. On whole, these data suggest substantial heterogeneity with regard to reporting uncertainty among radiologists. Simplified "uncertainty" lexicons may address this variability. More so, these data pose an important challenge and opportunity for medical educators to articulate best practices. Future work may measure the downstream effects of the variable use of these terms.

Conclusion

We developed and validated an algorithm to detect the presence of uncertainty in radiology reports. Using this validated tool, we analyzed large numbers of reports to better understand the variability in use of uncertainty terms among radiologists. Future work may identify the factors that influence individual variability in the expression of uncertainty and develop curricula to improve standardization.

Take-Home Points

- A validated NLP algorithm can accurately classify whether a radiology report expresses uncertainty.
- Substantial variability exists among radiologists, between subspecialties, and across imaging modalities within radiology regarding the communication of uncertainty.
- This data represents an approach to the study of when, why, and how radiologists communicate uncertainty.

Funding Information Dr. Callen was supported by the National Institutes of Health (NIBIB) T32 Training Grant, T32EB001631.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Appendix

Table 1 Uncertainty Terms

Alternatively
Ambiguous
Atelectasis + consolidation
Atelectasis + infiltrate
Borderline
Bosniak 2f
Cannot
Cannot
Concern
Consideration
Consolidation atelectasis could
Differential
Doubt
Equivocal
Etiology
Excluded
Incompletely evaluated
Incompletely characterized
Incompletely imaged
Indeterminate
Infection + inflammation
Likely
May
Might
No frank
Non-specific
Not known
Not absolute
No clear
No definite
No gross
No obvious
Other cause
Possible
Presume-
Probabl-
Question
Suggestive
Suspicio-
Too small to
Uncertain
Unclear
Unknown age
Unknown chronicity
Vague
Versus
Without absolute
Without clear
Without definite
Without gross
Without obvious
Worrisome

References

1. Schoenfeld AJ, Harris MB, Davis M. Clinical uncertainty at the intersection of advancing technology, evidence-based medicine, and health care policy. JAMA Surg. 2014;149(12):1221–2.
2. Platt ML, Huettel SA. Risky business: the neuroeconomics of decision making under uncertainty. Nat Neurosci. 2008;11(4):398–403.
3. Saposnik G, Sempere AP, Prefasi D, Selchen D, Ruff CC, Maurino J, et al. Decision-making in Multiple Sclerosis: The Role of Aversion to Ambiguity for Therapeutic Inertia among Neurologists (DIScUTIR MS). Front Neurol. 2017;8:65.
4. Rosenkrantz AB, Kiritsy M, Kim S. How "consistent" is "consistent"? A clinician-based assessment of the reliability of expressions used by radiologists to communicate diagnostic confidence. Clin Radiol. 2014;69(7):745–9.
5. Khorasani R, Bates DW, Teeger S, Rothschild JM, Adams DF, Seltzer SE. Is terminology used effectively to convey diagnostic certainty in radiology reports? Acad Radiol. 2003;10(6):685–8.
6. Carney PA, Yi JP, Abraham LA, Miglioretti DL, Aiello EJ, Gerrity MS, et al. Reactions to uncertainty and the accuracy of diagnostic mammography. J Gen Intern Med. 2007;22(2):234–41.
7. Clinger NJ, Hunter TB, Hillman BJ. Radiology reporting: attitudes of referring physicians. Radiology. 1988;169(3):825–6.
8. Panicek DM, Hricak H. How Sure Are You, Doctor? A Standardized Lexicon to Describe the Radiologist's Level of Certainty. AJR Am J Roentgenol. 2016;207(1):2–3.
9. Wallis A, McCoubrie P. The radiology report—are we getting the message across? Clin Radiol. 2011;66(11):1015–22.
10. Sobel JL, Pearson ML, Gross K, Desmond KA, Harrison ER, Rubenstein LV, et al. Information content and clarity of radiologists' reports for chest radiography. Acad Radiol. 1996;3(9):709–17.
11. Valls C. Pitfalls of the vague radiology report. AJR Am J Roentgenol. 2001;176(1):253–4.
12. Levinson W. Physician-patient communication. A key to malpractice prevention. JAMA. 1994;272(20):1619–20.
13. Hoang JK. Do not hedge when there is certainty. J Am Coll Radiol. 2017;14(1):5.
14. Burnside ES, Sickles EA, Bassett LW, Rubin DL, Lee CH, Ikeda DM, et al. The ACR BI-RADS experience: learning from history. J Am Coll Radiol. 2009;6(12):851–60.
15. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol. 2017;14(5):587–95.
16. Berlin L. Radiologic errors and malpractice: a blurry distinction. AJR Am J Roentgenol. 2007;189(3):517–22.
17. Hripcsak G, Austin JHM, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology. 2002;224(1):157–63.
18. Do BH, Wu A, Biswal S, Kamaya A, Rubin DL. Informatics in Radiology: RADTF: A Semantic Search–enabled, Natural Language Processor–generated Radiology Teaching File. Radiographics. 2010;30(7):2039–48.
19. Dutta S, Long WJ, Brown DFM, Reisner AT. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. Ann Emerg Med. 2013;62(2):162–9.
20. Dang PA, Kalra MK, Blake MA, Schultz TJ, Stout M, Lemay PR, et al. Natural language processing using online analytic processing for assessing recommendations in radiology reports. J Am Coll Radiol. 2008;5(3):197–204.

21. Dogra N, Giordano J, France N. Cultural diversity teaching and issues of uncertainty: the findings of a qualitative study. BMC Med Educ. 2007;7(1):8.
22. Wu AS, Do BH, Kim J, Rubin DL. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. J Digit Imaging. 2011;24(2):234–42.
23. Hanauer DA, Liu Y, Mei Q, Manion FJ, Balis UJ, Zheng K. Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. AMIA Annu Symp Proc. 2012;2012:321–30.
24. Bhise V, Rajan SS, Sittig DF, Morgan RO, Chaudhary P, Singh H. Defining and measuring diagnostic uncertainty in medicine: a systematic review. J Gen Intern Med. 2018;33(1):103–15.
25. Zwaan L, Singh H. The challenges in defining and measuring diagnostic error. Diagnosis (Berl). 2015;2(2):97–103.
26. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med. 2013;173(6):418–25.
27. Hoang JK, Middleton WD, Farjat AE, Langer JE, Reading CC, Teefey SA, et al. Reduction in thyroid nodule biopsies and improved accuracy with American college of radiology thyroid imaging reporting and data system. Radiology. 2018;287(1):185–93.
28. Tang A, Bashir MR, Corwin MT, Cruite I, Dietrich CF, Do RKG, et al. Evidence supporting LI-RADS major features for CT- and MR imaging-based diagnosis of hepatocellular carcinoma: a systematic review. Radiology. 2018;286(1):29–48.
29. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. Radiology. 2017;284(1):228–43.
30. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology. 2016;279(2):329–43.
31. Lacson R, Khorasani R. Practical examples of natural language processing in radiology. J Am Coll Radiol. 2011;8(12):872–4.
32. Srinivasa Babu A, Brooks ML. The malpractice liability of radiology reports: minimizing the risk. Radiographics. 2015;35(2):547–54.
33. Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology. 2005;234(2):323–9.
34. Abbass IM, Krause TM, Virani SS, Swint JM, Chan W, Franzini L. Revisiting the economic efficiencies of observation units. Manag Care. 2015;24(3):46–52.
35. Prabhakar AM, Misono AS, Harvey HB, Yun BJ, Saini S, Oklu R. Imaging utilization from the ED: no difference between observation and admitted patients. Am J Emerg Med. 2015;33(8):1076–9.
36. Molins E, Macià F, Ferrer F, Maristany M-T, Castells X. Association between radiologists' experience and accuracy in interpreting screening mammograms. BMC Health Serv Res. 2008;8:91.
37. Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DSM, Sickles EA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology. 2009;253(3):632–40.
38. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak. 2016;16(1):138.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.