Automatic detection of actionable findings and communication mentions in radiology reports using natural language processing

Abstract

Objectives. To develop and validate classifiers for automatic detection of actionable findings and documentation of nonroutine communication in routinely delivered radiology reports.

Methods. Two radiologists annotated all actionable findings and communication mentions in a training set of 1,306 radiology reports and a test set of 1,000 reports randomly selected from the electronic health record system of a large tertiary hospital. Various feature sets were constructed based on the impression section of the reports using different preprocessing steps (stemming, removal of stopwords, negations, and previously known or stable findings) and n-grams. Random forest classifiers were trained to detect actionable findings, and a decision-rule classifier was trained to find communication mentions. Classifier performance was evaluated by the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity.

Results. On the training set, the actionable finding classifier with the highest cross-validated performance was obtained for a feature set of unigrams, after stemming and removal of negated, known and stable findings. On the test set, this classifier achieved an AUC of 0.876 (95% CI 0.854-0.898). The classifier for communication detection was trained after negation removal, using unigrams as features. The resultant decision rule had a sensitivity of 0.841 (95% CI 0.706-0.921) and specificity of 0.990 (95% CI 0.981-0.994) on the test set.

Conclusions. Automatic detection of actionable findings and subsequent communication in routinely delivered radiology reports is possible. This can serve quality control purposes and may alert radiologists to the presence of actionable findings during reporting.

Keywords: Radiology information systems, Natural language processing, Machine learning, Personal Communication, Quality Control

Key points:

- Classifiers were developed for automatic detection of the broad spectrum of actionable findings and subsequent communication mentions in routinely delivered radiology reports.
- Straightforward report preprocessing and simple feature sets can produce well-performing classifiers.
- The resultant classifiers show good performance for detection of actionable findings and excellent performance for detection of communication mentions.

Abbreviations:

ACR: American College of Radiology

EHR: electronic health record

Introduction

The radiology report records the radiologist's interpretation of an imaging examination, and is the basis for the reporting and communication of imaging findings in routine clinical practice. However, some findings may require additional nonroutine communication with the referring physician because of their urgency or unexpectedness. These findings that necessitate special communication are called actionable findings [1]. Communication of actionable findings should be documented in the radiology report or otherwise indicated in the electronic health record (EHR). Monitoring the detection and communication of actionable findings in reports delivered in routine clinical care is important for quality control and auditing purposes, but requires laborious manual review that may not be feasible at scale.

A considerable number of studies have used natural language processing to automate the detection of actionable findings in radiology reports [2-11]. However, almost all of these studies focused on the detection of a limited set of findings (often related to a specific modality or organ system) or even a single finding. Moreover, most studies used their own list of actionable findings, and it is not clear how well their results would generalize to available guidelines, such as proposed by the American College of Radiology (ACR) Actionable Reporting Work Group [1]. Finally, most of these studies did not try to detect documentation of communication with the referring physician if the report contained an actionable finding.

The aim of this study is to develop and validate algorithms for automatic detection of the broad spectrum of actionable findings and documentation of communication in routinely delivered radiology reports.

Materials and methods

This study was approved by the Medical Ethics Review Committee of our institution (registration number MEC-2017-200). Since all data were anonymized and retrospectively collected, informed consent of the subjects was not required according to Dutch legislation.

Data

The study population consisted of two sets of radiology reports: a training set to train and validate the classifiers, and a test set to assess their performance. All reports were taken from the EHR system of the Department of Radiology and Nuclear Medicine of our institution, one of the largest tertiary hospitals in The Netherlands. The reports cover all radiological imaging modalities and specialties, and were all written in Dutch. All reports are based on a semi-structured report template consisting of four sections: indication, clinical history, findings, and impression. The sections contain free text provided by the interpreting radiologist. The impression section could be left empty, if the radiologist gave his impression in the findings section. There is a standard operating procedure to include a remark in the report if a finding has been communicated nonroutinely. This procedure applies for all divisions in our department. During the study period, there were no major changes in the report template.

The training set was taken from 58,889 consecutive radiology reports that were stored in the EHR system from December 2016 to April 2017. It consisted of 306 reports from the full set that had been labeled as containing an actionable finding by the interpreting radiologist, and another 1,000 reports that were randomly sampled from the remainder of the full set. Thus, the training set included 1,306 reports. The test set consisted of a random sample of 1,000 radiology reports from 124,909 consecutive radiology reports stored from June 2017 till March 2018, and has been described previously [12].

Data annotation

Two radiologists (with five and eight years of experience) independently annotated all actionable findings in the 2,306 reports of the training and test sets, following the guidelines and lists of actionable findings that were developed by the ACR Actionable Reporting Work Group [1]. Briefly, the ACR Work Group distinguishes three categories that need communication in addition to the radiology report. Category 1 consists of critical or urgent findings that require communication within minutes, e.g., intracranial hemorrhage. Category 2 findings are clinically significant observations that require specific medical or surgical treatment, but do not have the urgency of category 1 findings. Findings in category 2 should be communicated within hours. Examples of category 2 are pneumothorax, bone lesions at risk for pathologic fracture, and intra-abdominal infections like appendicitis or cholecystitis. Category 3 findings are incidental or unexpected, but do not require immediate treatment or other action, e.g., cirrhosis, probable malignancy on any location without acute danger to the patient, and hemodynamically significant arterial stenosis not associated with acute symptoms or otherwise immediately threatening. As these findings are incidental, there is a risk of being overlooked by the care provider who is responsible for follow-up. Category 3 findings are therefore required to be communicated within days. Before the radiologists started their annotation work, they scrutinized and discussed the ACR Work Group publication [1] and in particular the Appendix with actionable finding definitions in order to align their interpretation.

The annotators used brat, a web-based annotation tool [13]. If a report contained an actionable finding, the annotators marked the phrase that describes the finding and labeled it with the category of the finding. If the report contained multiple actionable findings, each finding was annotated separately.

After the initial round of annotations, the annotators jointly went over the reports in which their category annotations differed, and established a final category for each report. If more than one actionable finding had been annotated, the most severe category was taken as the final category.

In a separate annotation round, one of the annotators used brat to mark all phrases or sentences that mentioned communication between the interpreting radiologist and the referring physician.

Natural language processing and classifier development

Two classifiers were developed, one for detecting communication mentions and the other for detecting actionable findings. As an input we used the impression section of the reports. If the impression section was missing, we used the description section (this occurred in 23.7% (309/1306) of the reports in the training set and in 27.5% (275/1000) of the reports in the test set). The input text was converted to lowercase and split into sentences.

We experimented with different preprocessing steps: stopword removal (using the Dutch stopword list in the R package *stopwords*), stemming (reducing words to their base form, using the Dutch stemmer in R package *SnowballC*), and removal of phrases that indicated negated, previously known or stable findings. Phrases were removed by searching the report for any of a small set of trigger words (the Dutch equivalents of “not”, “no”, “without”, “known”, and “unchanged”). If a trigger word was found, that word and all words following it until the end of the sentence were discarded. The remaining text was then split in n-grams (sequences of n words, we varied n from 1 to 3), which were used as features to train the classifiers. In order not to bias the classifier for actionable findings, we removed all communication mentions that had been annotated in the training set before training the classifier.

Each report was assigned a binary label for actionable findings, where a value of 1 indicated presence of a final annotation of category 1, 2 or 3, and a value of 0 indicated absence. Similarly, binary labels were assigned for the presence or absence of communication mentions in the reports. Classifiers were trained on the training set using 10-fold cross-validation. For the actionable findings, we trained a random forest classifier (method *ranger* in the R package *caret*) [14]; for the communication detection, we trained a rule-based classifier, JRip (Weka tool kit) [15], an implementation of the RIPPER algorithm [16]. The performance of the final models was tested on the test set.

Evaluation

The annotations of actionable findings by the two radiologists were collected in a 2x2 confusion matrix indicating presence (category 1, 2, or 3) or absence of an actionable finding in the reports. From this matrix, interannotator agreement scores were derived: accuracy (proportion of agreement) and the kappa coefficient (proportion agreement corrected for chance agreement). The final annotations were taken as the reference for the training and evaluation of the classifiers.

Performance of the classifiers was measured by the area under the receiver operating characteristic curve (AUC) on the training set using 10-fold cross-validation, and on the test set. Sensitivity or recall (proportion of positive cases that are correctly classified), specificity (proportion of negative cases that are correctly classified), positive predictive value (PPV) or precision (proportion of positive classified cases that are correct), and negative predictive value (NPV) (proportion of negative classified cases that are correct) were computed for specific probability cutoffs.

For statistical tests, a P value ≤ 0.05 was considered significant.

Results

Annotations

The confusion matrices of the radiologists' annotations of actionable findings in the training and test sets are shown in Table 1. The accuracy and kappa coefficient for the training set were 0.866 (1,131/1,306) and 0.715, respectively, and for the test set 0.868 (868/1,000) and 0.696. Of the discrepancies between the radiologists in the test set, the majority (94/132) were annotations where one radiologist indicated a category 3 actionable finding and the other no actionable finding. Most of these differences may be explained by the usage of qualitative modifiers in the lists of actionable findings provided in the Appendix of the ACR Work Group report, e.g. clinically significant, highly suggestive, probable, suspected, or mild. The interpretation of these terms can vary between the annotators, leading to reduced agreement scores. Of the differences between category 2 and no actionable finding annotations (38/132), the majority could be attributed to a different assessment of whether a finding was previously known (and thus whether it should be labelled as actionable). Differences between category 1 and no actionable finding annotations did not occur. Two examples of annotator disagreement are given in the Supplementary Material.

Based on the final annotations, the training set contained 555 reports with actionable findings, and the test set 325. Mentions of communication between interpreting radiologist and referring physician were annotated in 125 reports of the training set, and in 44 reports of the test set.

Actionable finding classifier

The classifier for actionable findings was trained on the training set using default hyperparameters (number of trees = 500, unlimited tree depth, minimum node size = 1, gini splitting rule), except for mtry (the number of variables to possibly split at in each node) which was optimized. We constructed various feature sets by taking all combinations of the following settings: removal of stopwords yes or no; word stemming yes or no; removal of negated, known or stable findings yes or no; and use of unigrams only, unigrams and bigrams, or unigrams, bigrams and trigrams. The highest 10-fold cross-validated performance, with an AUC of 0.894 (95% confidence interval (CI) 0.876-0.911), was obtained for a feature set of unigrams after stemming and removal of negated, known and stable findings, and an mtry of 20. Testing this classifier on the test set resulted in an AUC of 0.876 (95% CI 0.854-0.898). Figure 1 shows the corresponding ROC curve, and Figure 2 the precision-recall curve. The curves show the different combinations of sensitivity, specificity and PPV that can be obtained by varying the probability cutoff. For example, a classifier with 0.8 specificity will have a sensitivity of 0.791 and a PPV of 0.647, or a classifier with 0.9 specificity will have 0.610 sensitivity and 0.745 PPV. A few examples of actionable finding classifications are provided in the Supplementary Material. The model and the script to generate the feature sets are available at github.com/mi-erasmusmc.nl/ActionableFindingsClassifier.

We have done a subanalysis of the classifier performance per modality. The report distribution across modalities is given in Table 2 for the training and test set. This random sample reflects the distribution of reports among the different modalities in routine clinical practice at our institution. Conventional X-ray was the modality with the largest number of reports, while ultrasonography and MRI had the smallest number of reports. On the test set, the AUCs of the actionable finding classifier were 0.874 for X-ray, 0.897 for CT, 0.871 for ultrasonography, and 0.784 for MRI.

To determine whether the classifier performs differently for the three categories of actionable findings, we analyzed the distribution of the classifier probabilities for each category. Figure 3 shows that the classifier probabilities for category 1 are generally higher than those for categories 2 and 3, indicating that the classifier, given a probability cut-off, is less likely to miss actionable findings in category 1 than in categories 2 and 3. We also compared relative word frequencies in reports that were classified false-negatively and true-positively, and in reports classified false-positively and true-negatively. We made scatter plots of the relative word frequencies in each pair of report sets, showing words that are overrepresented or underrepresented in the false-positive and false-negative reports (see Figures S1 and S2 in the Supplementary Material).

Communication mention classifier

The classifier for the detection of communication mentions was trained on the training set after negation removal and using unigrams as features. Figure 4 gives the resulting decision rule (with English translations of the Dutch unigrams in square brackets). The 10-fold cross-validation estimate for sensitivity was 0.913 (114/125) (95% CI 0.849-0.950) and for specificity 0.993 (1,173/1,181) (95% CI 0.987-0.997). On the test set, sensitivity and specificity were 0.841 (37/44) (95% CI 0.706-0.921) and 0.990 (946/956) (95% CI 0.981-0.994), respectively, giving a PPV of 0.787 (37/47) (95% CI 0.651-0.880) and an NPV of 0.993 (946/953) (95% CI 0.985-0.996). The differences in sensitivity and specificity between training and test sets were not statistically significant (P > 0.10, Fisher’s exact test).

Discussion

We developed classifiers to detect actionable findings in routinely delivered radiology reports, and to detect mentions of communication between the interpreting radiologist and the referring physician about these findings. Using a simple, straightforward approach, we obtained good to excellent performance results on an independent test set.

We are not the first to automatically detect actionable findings in radiology reports. Previous studies have reported different approaches, often with excellent results [2-11]. However, almost all of these studies focused on the detection of a limited set of findings, often related to a specific modality or organ system. We aimed at detecting all actionable findings in routinely delivered radiology reports, which makes our classification task more difficult and performance results difficult to compare.

Our natural language processing approach is straightforward and practical. We used simple bag-of-words features (n-grams), and only performed basic preprocessing steps. More sophisticated features have been proposed, including those based on topic modelling and word embeddings, but bag-of-words based models achieved competitive results in a comparative study of different feature sets for the classification of critical findings in head CT reports [8]. In the past, very good results have been reported for rule-based systems, in which the rules were manually crafted [2; 6; 17]. However, this approach is time-consuming and requires active involvement of radiologists, which may be problematic in a demanding clinical environment. We only asked radiologists to label the reports in the training and test sets, which takes considerably less time and effort.

Very few studies investigated the automatic detection of communication mentions of actionable findings in radiology reports. Some researchers addressed related tasks, such as the automatic identification of clinically important recommendations [18], recommendations for additional imaging [19] or subsequent action [2]. To our knowledge, only two studies tried to detect the communication of critical results [11; 20]. Lakhani et al. [20] developed and iteratively refined a set of hand-crafted pattern-matching rules that combined words and phrases used to indicate nonroutine communication. The system performed extremely well, with a recall of 98.2% and a precision of 97.0%, but was not made publicly available. Meng et al. [11] used a keyword-based approach to identify documentation of communication in reports with findings that require “prompt communication”. They selected keywords based on similarity of distributional semantic representations with the word “communicated”, but did not report the performance of their approach.

Regarding our machine-learning algorithms, we chose to use random forests for classifying the actionable findings. Random forests have been shown to be among the best classifiers on a wide variety of problems [21]. For the communication detection, we used a decision-rule learner. Random forests also performed well on this task (data not shown), but the decision rules have the advantage that they are comprehensible and can manually be optimized. For example, based on an analysis of the errors on the test set, we noticed four additional keywords that appeared to be related to communication mentions and were missing from our original decision rule. When these were added, sensitivity improved to 97.7% while specificity remained the same at 99.0%. Of note, these results should still be corroborated on a new, independent test set.

The interannotator agreement scores on our training and test sets are moderate, indicating that the classification of actionable findings is a difficult task. The performance of our classifier should be judged in view of this interannotator agreement. While performance of the classifier is not yet on par with human performance, the difference is not large.

The performance subanalysis of the specific categories of actionable findings indicated that our classifier is less likely to miss actionable findings in category 1 than in categories 2 and 3. While we consider a high sensitivity for detecting actionable findings in the most severe category beneficial, actionable findings in any category can have clinical consequences and adequate detection and communication for all categories remains important. The subanalysis of the performance per modality showed that our classifier performs best for CT reports, slightly less for X-ray and ultrasonography, and least well for MRI (although the AUC of 0.784 may still be considered acceptable). The lower performance for MRI might be due to the relatively small number of MRI reports in our training set, possibly in combination with a high level of detail in reports on MRI examinations that may obfuscate actionable findings.

As to the practical use of the classifiers, we plan to incorporate them in a dashboard for quality control of radiological reporting that is currently being developed in our hospital. The classifier for detection of actionable findings will provide an estimate of the prevalence of actionable findings, while the classifier for detection of communication mentions will allow to assess trends and guideline adherence for reporting of actionable findings.

A possible quality control use scenario is the detection and validation of reports that contain an actionable finding without mention of nonroutine communication. The sensitivity of the actionable finding classifier determines how many reports with an actionable finding will be missed (falsenegatives), and has to be balanced against the specificity, which determines the number of reports without an actionable finding that are classified as containing an actionable finding (false-positives). The sensitivity and specificity of the classifier can be adjusted by changing the probability threshold, dependent on user preferences. The communication mention classifier would then filter the reports that do not contain nonroutine communication. For this, a high NPV is wanted, that is, a high probability that the report does not contain nonroutine communication if the classifier does not find it. This is indeed the case for the mention classifier, which has an NPV of 0.993.

Another application for daily clinical practice would be the incorporation of these algorithms into reporting systems. Although further performance improvement is required, these types of algorithms could alert the reporting radiologist to actionable findings so that appropriate action is assured. Alternative approaches to natural language processing, based on standard operating procedures and structured reporting [22], may also help to improve the reporting of actionable findings, e.g., by providing a button in the radiology information system to report an actionable finding, or by using a keyboard shortcut to generate a standard statement on nonroutine communication. In practice, however, these procedures may not always be adhered to by radiologists and a quality control tool to check whether actionable findings are reported according to the guidelines will still be relevant.

Our study has several limitations. First, we developed and tested our classifiers in only one university hospital. Further testing in other hospitals is necessary to establish whether the classifiers can be applied in different settings. Second, we have only tested our algorithms on Dutch radiology reports. While the bag-of-words approach is relatively simple to apply in other languages, it remains to be evaluated whether the algorithms would produce equally good results. Third, our simple approach of phrase removal worked well for our set of reports, but may be too simplistic for other settings or languages. Finally, the size of our data set is still relatively small. More data may allow to further increase classifier performance. More data would also allow to train classifiers that discriminate between the three categories of actionable findings (or even between specific actionable findings), for which the current data set is not large enough.

In conclusion, automatic detection of actionable findings and subsequent communication in routinely delivered radiology reports is possible. This can serve quality control purposes and may alert radiologists to the presence of actionable findings during reporting.

Acknowledgements

This study was supported by a grant from Stichting Kwaliteitsgelden Medisch Specialisten (Foundation Quality Funds Medical Specialists) [grant number 45368564].

References

1 Larson PA, Berland LL, Griffith B, Kahn CE, Liebscher LA (2014) Actionable findings and the role of IT support: report of the ACR Actionable Reporting Work Group. J Am Coll Radiol 11:552-558

2 Dreyer KJ, Kalra MK, Maher MM et al (2005) Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 234:323-329

3 Lacson R, Sugarbaker N, Prevedello LM et al (2012) Retrieval of radiology reports citing critical findings with disease-specific customization. Open Med Inform J 6:28-35

4 Lakhani P, Kim W, Langlotz CP (2012) Automated detection of critical results in radiology reports. J Digit Imaging 25:30-36

5 Pham AD, Neveol A, Lavergne T et al (2014) Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 15:266

6 Chapman BE, Mowery DL, Narasimhan E, Patel N, Chapman WW, Heilbrun ME (2016) Assessing the feasibility of an automated suggestion system for communicating critical findings from chest radiology reports to referring physiciansProceedings of the 15th Workshop on Biomedical Natural Language Processing. ACL, Berlin, pp 181-185

7 Chen MC, Ball RL, Yang L et al (2018) Deep learning to classify radiology free-text reports. Radiology 286:845-852

8 Zech J, Pain M, Titano J et al (2018) Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology 287:570-580

9 Banerjee I, Ling Y, Chen MC et al (2019) Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med 97:79-88

10 Heilbrun ME, Chapman BE, Narasimhan E, Patel N, Mowery D (2019) Feasibility of natural language processing-assisted auditing of critical findings in chest radiology. J Am Coll Radiol 16:1299-1304

11 Meng X, Ganoe CH, Sieberg RT, Cheung YY, Hassanpour S (2019) Assisting radiologists with reporting urgent findings to referring physicians: a machine learning approach to identify cases for prompt communication. J Biomed Inform 93:103169

12 Visser JJ, de Vries M, Kors JA (2020) Assessment of actionable findings in radiology reports. Eur J Radiol 129:109109

13 Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsjujii J (2012) brat: a web-based tool for NLP-assisted text annotationProceedings of the Demonstrations Session at EACL. Association for Computational Linguistics, pp 103-107

14 Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1-26

15 Hall M, Frank E, Holmes G (2009) The WEKA data mining software: an update. SIGKDD Explorations 11:10-18

16 Cohen WW (1995) Fast effective rule induction. Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, pp 115-123

17 Lakhani P, Kim W, Langlotz CP (2012) Automated extraction of critical test values and communications from unstructured radiology reports: an analysis of 9.3 million reports from 1990 to 2011. Radiology 265:809-818

18 Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH (2013) A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform 46:354-362

19 Dutta S, Long WJ, Brown DF, Reisner AT (2013) Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. Ann Emerg Med 62:162-169

20 Lakhani P, Langlotz CP (2010) Automated detection of radiology reports that document non-routine communication of critical or significant results. J Digit Imaging 23:647-657

21 Fernandez-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15:3133-3181

22 European Society of Radiology (2018) ESR paper on structured reporting in radiology. Insights Imaging 9:1-7

Figure and table legends

Figure 1. ROC curve for the classifier of actionable findings on the test set.

Figure 2. Precision-recall curve for the classifier of actionable findings on the test set.

Figure 3. Distribution of the classifier probabilities in the test set for different categories of actionable findings according to the reference.

Figure 4. Automatically constructed decision rule for detecting documentation of communication between interpreting radiologist and referring physician in Dutch radiology reports. English translations of the Dutch terms are shown in brackets.

Table 1. Confusion matrices of annotations of actionable findings in radiology reports of the training and test sets by two radiologists.

Table 2. Number of reports per modality for the training and test sets.