European Journal of Radiology 129 (2020) 109109

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Assessment of actionable findings in radiology reports

Jacob J. Visser[a,*], Marianne de Vries[a], Jan A. Kors[b]

[a] Department of Radiology and Nuclear Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
[b] Department of Medical Informatics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Radiology information systems
Incidental findings
Prevalence
Interobserver variability

ABSTRACT

Purpose: The American College of Radiology (ACR) Actionable Reporting Work Group defined three categories of imaging findings that require additional, nonroutine communication with the referring physician because of their urgency or unexpectedness. The objective of this study was to determine the prevalence of actionable findings in radiology reports, and to assess how well radiologists agree on the categorisation of actionable findings.

Method: From 124,909 consecutive radiology reports stored in the electronic health record system of a large university hospital, 1000 reports were randomly selected. Two radiologists independently annotated all actionable findings according to the three categories of urgency defined by the ACR Work Group. Annotation differences were resolved in a consensus meeting and a final category was established for each report. Interannotator agreement was measured by accuracy and the kappa coefficient.

Results: The prevalence of the three categories of actionable findings together was 32.5 %. Of all reports, 10.9 % were from patients seen in the emergency department. Prevalence of actionable findings for these patients (45.9 %) was considerably higher than for patients in routine clinical care (30.9 %). Interannotator agreement scores on the categorisation of actionable findings were 0.812 for accuracy and 0.616 for kappa coefficient.

Conclusions: The prevalence of actionable findings in radiology reports is high. The interannotator agreement scores are moderate, indicating that categorisation of actionable findings is a difficult task. To avoid unneeded increase in the workload of radiologists, in particular in routine practice, clinical context may need to be considered in deciding whether a finding is actionable.

1. Introduction

The radiologist's interpretation of an imaging examination can only impact and improve patient care if the referring physician is notified of the results of the requested examination. Guidelines for the reporting of imaging findings have been established [1]. However, some of these findings may require additional, nonroutine communication with the referring physician because of their urgency or unexpectedness. Such findings that necessitate special communication are called actionable findings [2,3].

There has been considerable variation in nomenclature and classification of actionable findings [4–8]. Larson et al. [3], representing the American College of Radiology (ACR) Actionable Reporting Work Group (hereafter called the ACR Work Group), proposed a classification in three categories based on the timing of communication: findings that require communication within minutes (category 1), within hours (category 2), or within days (category 3). They also provided an extensive list of actionable findings for each category. The ACR Work Group indicated that these findings should be new or known to have significantly worsened since a prior study. A stable finding that was previously known and appropriately communicated may not require additional nonroutine communication despite the severity of the disease process.

To assess the impact of the ACR Work Group recommendations on the radiologist's workflow, information about the prevalence of actionable findings as defined by the ACR Work Group is needed. Many studies have reported on the prevalence of incidental imaging findings [8,9], often focussing on a single imaging modality or subspecialty. Only few studies have assessed the prevalence of the full scope of actionable findings in radiology reports [2,10,11]. The prevalence estimates in these studies vary widely, between 1.5 % and 13.1 %. Also, it is not clear to what extent the prevalences reported in these studies

Abbreviations: ACR, American College of Radiology

* Corresponding author at: Department of Radiology and Nuclear Medicine, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA, Rotterdam, the Netherlands.
E-mail addresses: j.j.visser@erasmusmc.nl (J.J. Visser), m.devries.2@erasmusmc.nl (M. de Vries), j.kors@erasmusmc.nl (J.A. Kors).

https://doi.org/10.1016/j.ejrad.2020.109109
Received 27 January 2020; Received in revised form 20 May 2020; Accepted 31 May 2020

reflect the classification of actionable findings provided by the ACR Work Group, and in how far there is agreement between radiologists in the categorisation of these findings.

Our aim was to determine the prevalence of the three levels of urgency of actionable findings as defined by the ACR Work Group in a random sample of all radiology reports generated in a large university hospital, and to assess how well radiologists agree on the categorisation of actionable findings.

2. Material and methods

This study was approved by the Medical Ethics Committee of the Erasmus MC. Since all data were anonymised and retrospectively collected, informed consent of the subjects was not required according to Dutch legislation.

2.1. Study population

A sample of 1000 radiology reports was randomly selected from 124,909 consecutive radiology reports stored in the electronic health record system of the Department of Radiology and Nuclear Medicine of the Erasmus MC from June 2017 until March 2018. The reports are from both inpatients and outpatients, and cover all radiological imaging modalities and specialties in the Erasmus MC, one of the largest academic centers in The Netherlands. All reports have a standard layout with four sections: indication, clinical history, findings, and impression. The sections contain free text provided by the interpreting radiologist.

2.2. Categorisation of actionable findings

Two general radiologists (one with subspecialty musculoskeletal and five years of experience, the other with subspecialty abdomen and eight years of experience) independently annotated all actionable findings in the 1000 reports and categorised them in three groups according to lists of actionable findings that were developed by the ACR Work Group [3]. Briefly, 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 not immediately threatening. As these findings are incidental, there is a risk of their being overlooked by the care provider who is responsible for follow-up. Category 3 findings are therefore required to be communicated within days. Note that the annotators did not annotate stable findings, in accordance with the ACR Work Group recommendations. However, they did not have access to prior reports and had to tell from the given report whether a finding was previously known.

Both radiologists used brat, a web-based, open-source annotation tool [12]. 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. If the report did not contain an actionable finding, an arbitrary phrase in the report (usually the conclusion header) was marked as category 4.

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. Reasons for differences of more than one category level (e.g. category 2 vs. category 4) were elucidated and grouped. If a report was annotated with more than one actionable finding, the most severe category was taken as the final category.

Table 1
Characteristics of the study population.

Variable	Value (n = 1000)
Age (y)	47.5 ± 23.7 (0−91)[a]
Sex	
Men	516
Women	484
Modality	
Conventional X-ray	519
Computed tomography	238
Ultrasonography	122
Magnetic resonance imaging	121
Subspecialty	
Musculoskeletal	274
Chest	252
Gastrointestinal	201
Neurologic/head and neck	170
Cardiac/vascular	44
Breast	39
Genitourinary/obstetric	11
Unknown	9
Emergency department	109

[a] Values are mean ± standard deviation, with range in parentheses.

2.3. Statistical analysis

The category annotations of the two annotators were collected in a 4×4 confusion matrix and, after merging category 1, 2, or 3 annotations in a single category indicating presence of an actionable finding, in a 2×2 confusion matrix. From these matrices, interannotator agreement scores were derived: accuracy (proportion of agreement) and the kappa coefficient (proportion agreement corrected for chance agreement). Prevalence estimates of the categories were based on the final annotations.

3. Results

Table 1 shows age and sex of the study population, the distributions according to imaging modality and subspecialty, and the number of individuals who were seen in the emergency department of our hospital. Of the 1000 reports that were annotated by the two radiologists, 297 reports contained actionable findings according to one radiologist, and 339 according to the other. The confusion matrix of the category annotations of the two radiologists is shown in Table 2. The accuracy and kappa coefficient across all categories was 0.812 and 0.616, respectively. For 148 of the 188 reports where the radiologists disagreed, their annotations differed one category level (e.g. category 1 by one radiologist and category 2 by the other). For 40 reports, a two-level difference in annotations (category 1 vs. category 3 or category 2 vs. category 4) was present. A three-level difference (category 1 vs. category 4) did not occur. For binary annotations (absence or presence of

Table 2
Confusion matrix of the category annotations of actionable findings by two radiologists in 1000 radiology reports.

	Annotator 2			
Annotator 1	Category 1	Category 2	Category 3	Category 4
Category 1	12	1	0	0
Category 2	11	74	5	10
Category 3	2	37	110	35
Category 4	0	28	59	616

Table 3
Distribution of actionable findings across categories for patients seen in the emergency department and patients seen in routine clinical practice.

	Actionable finding				
Setting	Category 1	Category 2	Category 3	Category 4	Total
Emergency	6 (5.5)[a]	38 (34.9)	6 (5.5)	59 (54.1)	109
Routine	17 (1.9)	93 (10.4)	165 (18.5)	616 (69.1)	891
Total	23 (2.3)	131 (13.1)	171 (17.1)	675 (67.5)	1000

[a] Data are numbers of findings, with percentages per category in parentheses.

any actionable finding), interannotator agreement scores increased to 0.868 for accuracy and 0.696 for the kappa coefficient.

Final annotations were established in a consensus meeting as described above. Of the 1000 radiology reports, 23 were annotated as category 1, 131 as category 2, 171 as category 3, and 675 as category 4. Thus, the total number of reports with actionable findings in our study population was 325, giving a prevalence of 32.5 %. Table 3 shows the actionable findings for patients seen in the emergency department and in routine clinical care. The prevalence of actionable findings for patients with an emergency examination was 45.9 % (50/109), whereas the prevalence for patients with a routine examination was 30.9 % (275/891). The more urgent findings (category 1 and 2) were relatively much more frequent in the emergency setting than in the routine setting (40.3 % vs. 12.3 %), whereas the category-3 findings were relatively infrequent (5.5 % vs. 18.5 %). Of all actionable findings in the emergency setting, 88 % (44/50) were in category 1 or 2, whereas in the routine setting, 40 % (110/275) of the findings were in category 1 or 2.

Analysis of the 40 reports for which the annotation categories differed by two levels revealed that for 26 discrepancies the radiologists differed in their assessment of whether a finding was previously known (and thus whether it should be labelled as actionable). Six differences resulted from different interpretations of the definition of the actionable finding, four were due to non-conclusive imaging results, and three occurred because an annotation mistake had been made by either one of the annotators. In one report, one of the radiologists annotated “pulmonary edema” as a category 2 finding, although pulmonary edema is not part of the list of actionable findings proposed by the ACR Work Group [3]. The other radiologist initially annotated this report as category 4 but then agreed on category 2 as the final annotation.

In Table 4, the distribution of actionable findings across categories is shown for the different imaging modalities. CT and MRI show higher prevalences of actionable findings (50.4 % and 38.0 %) compared to conventional X-ray and ultrasound (28.7 % and 23.9 %). All six category-1 findings from the emergency department were based on CT, accounting for 46 % (6/13) of the CT-related findings in category 1. Of the category-2 findings, 35 % (13/37) came from the emergency department for CT, 35 % (6/17) for ultrasonography, 27 % (19/70) for X-ray, and 0% (0/7) for MRI.

Table 5 shows the distribution of actionable findings across categories for different subspecialties. Gastrointestinal has the highest prevalence of actionable findings (44.8 %), whereas musculoskeletal and breast have the lowest prevalences (19.3 % and 20.5 %). Five of the category-1 findings from the emergency department were neurological, accounting for 35 % (5/13) of the total neurological-related findings in category 1. The highest proportions of category-2 findings from the emergency department were seen for musculoskeletal (52 %, 15/29), neurological (42 %, 5/12), and gastrointestinal (29 %, 10/35).

Table 4
Distribution of actionable findings across categories for different imaging modalities.

	Actionable finding			
Modality	Category 1	Category 2	Category 3	Category 4
Conventional X-ray	5 (1.0)[a]	70 (13.5)	49 (9.4)	395 (76.1)
Computed tomography	13 (5.5)	37 (15.5)	70 (29.4)	118 (49.6)
Ultrasonography	0 (0.0)	17 (13.9)	18 (14.8)	87 (71.3)
Magnetic resonance imaging	5 (4.1)	7 (5.8)	34 (28.1)	75 (62.0)

[a] Data are numbers of findings, with percentages per modality in parentheses.

Table 5
Distribution of actionable findings across categories for different subspecialties.

	Actionable finding			
Subspecialty	Category 1	Category 2	Category 3	Category 4
Musculoskeletal	1 (0.4)[a]	29 (10.6)	23 (8.4)	221 (80.7)
Chest	6 (2.4)	50 (19.8)	35 (13.9)	161 (63.9)
Gastrointestinal	2 (1.0)	35 (17.4)	53 (26.4)	111 (55.2)
Neurologic/head and neck	13 (7.6)	12 (7.1)	39 (22.9)	106 (62.4)
Cardiac/vascular	1 (2.3)	4 (9.1)	10 (22.7)	29 (65.9)
Breast	0 (0.0)	0 (0.0)	8 (20.5)	31 (79.5)
Genitourinary/obstetric	0 (0.0)	0 (0.0)	3 (27.3)	8 (72.7)
Unknown	0 (0.0)	1 (11.1)	0 (0.0)	8 (88.9)

[a] Data are numbers of findings, with percentages per subspecialty in parentheses.

4. Discussion

We found a prevalence of actionable findings in our population of 32.5 %. Prevalence for patients seen in the emergency department (45.9 %) was considerably higher than for patients in routine clinical care (30.9 %). Furthermore, the radiologists initially disagreed in nearly 19 % of their categorisations of actionable findings, yielding a moderate interannotator agreement.

This is the first study to report on the prevalence of actionable findings according to the ACR Work Group guidelines in radiology reports. Our prevalence estimates are much higher than previously reported. Anthony et al. [2] searched a random sample of 16,983 reports from a tertiary academic medical center for critical results, defined as new or unexpected findings that could result in mortality or significant morbidity without appropriate follow-up, or interpretations differing from a previously communicated interpretation. They found a prevalence of 9.6 %. In a follow-up study by the same institution to evaluate the impact of an alert notification system, a prevalence of 13.1 % was reported [10]. Both studies distinguished three levels of urgency (red, orange, and yellow alerts), roughly corresponding with the categories defined by the ACR Work Group, but did not provide prevalence figures for separate categories. In another study to assess the effect of a critical results communication system, only 1.5 % of the 467,134 reports that were generated after system implementation, contained an actionable finding [11]. The difference between these prevalence estimates and ours may possibly be explained, at least partly, by differences in the definitions of actionable findings. We adhered to the lists of actionable findings that are given in the Appendix of the ACR Work Group report [2], while for the earlier reports such lists were not specified or could not be retrieved. Also, our study population was taken from a tertiary care center, which may have led to a higher prevalence.

The interannotator agreement scores indicate that categorisation of actionable findings is a difficult task. On average, the two radiologists judged differently on an actionable finding in almost one out of five radiology reports. Part 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 [2], e.g. clinically significant, highly suggestive, probable, suspected, mild, or moderate. The interpretation of these terms can vary between the annotators, leading to reduced agreement scores. However, most of the discrepancies were relatively small, with annotations that differed by one

category level. Of the larger differences, the majority could be attributed to a difference in opinion about the presence of the finding in the previous examination. Only one difference was due to a finding (pulmonary edema) that was not on the list of actionable findings of the ACR Work Group. We considered this finding actionable, keeping in mind that the ACR Work Group does not claim their list to be definitive, stating that "radiologists should always use their judgment and treat similarly important findings that are not on the list in the same manner when required for optimal patient care" [2].

The prevalence of actionable findings among patients seen in the emergency department proved to be much higher than among patients seen in routine clinical practice. Also, relatively many actionable findings in the emergency department were urgent (category 1 and 2), whereas non-urgent category-3 findings were few. These urgent findings in the emergency setting were mostly found for CT and contributed to the high prevalence of actionable findings for this modality. With regard to subspecialty, urgent findings in the emergency setting were mostly related to neurological and musculoskeletal specialties.

Our prevalence results suggest that adherence to the ACR Work Group guideline may increase the workload of radiologists and significantly interfere with their daily clinical activities. It should be noted, however, that most actionable findings at the emergency department are urgent and will be communicated with the referring physician in person quite often. For actionable findings in the routine setting, the referring physicians will be disturbed more frequently if the actionable findings were communicated according to the guideline. To avoid unneeded increase of the radiologist's workload and unnecessary interference with the activities of referring physicians, it could be helpful to take the clinical context into account. For example, if a patient with fever and cough who is suspected to have a pneumonia, indeed has a pneumonia, no additional, nonroutine communication is required. Whereas, if somebody is presented as suspected of a pneumothorax and the chest X-ray is suggestive of a pneumonia, nonroutine communication is required to assure that appropriate treatment will be started. To determine the potential effect of this recommendation, one radiologist reassessed the 325 final annotations of actionable findings and considered 230 (3 category 1; 81 category 2; 146 category 3) not to be actionable in view of the clinical context. Thus, taking clinical context into account can greatly reduce the number of findings that are actionable, but even then actionable findings will remain prevalent in routine radiology reports and require substantial effort of radiologists to report and communicate. Information and communication technology may offer solutions to better organise and facilitate the reporting process [3,11,13]. Automatic detection of actionable findings using natural language processing may further support the radiologist in consistently detecting and reporting actionable findings [14–17]. Structured reporting also has potential value in reporting actionable findings [18]. Usage of standardised terms and structured reporting would improve the consistency of reporting of actionable findings and therefore improve interobserver variability. In addition, as structured reporting standardizes the report format, referring physicians would be able to more easily find the actionable findings in the report.

Our study has a number of strengths. The annotations of actionable findings were based on well-established guidelines of the ACR Work Group, and were made in a random sample of all radiology reports generated in our hospital. Our results thus reflect the prevalence of actionable findings in clinical care. Also, we separately analysed reports from emergency examinations and from routine examinations, indicating a higher prevalence of actionable findings, most of them urgent, in the emergency setting. Finally, we determined interannotator agreeement scores, showing the difficulty of the classification task and providing a yardstick for the performance of automatic classification algorithms in the future.

Our study also has several limitations. The results were obtained in one single tertiary care center. The patient mix in our hospital may have led to higher prevalence estimates than would be obtained in urban or rural general hospitals. Furthermore, actionable findings were annotated by only two radiologists. Involvement of more radiologists could result in more robust estimates of interannotator agreement. Finally, the number of reports in this study is relatively small, in particular for assessing the prevalence of actionable findings in subgroups of modality or subspecialty. Increasing the sample size would improve the precision of the prevalence estimates.

In conclusion, we found high prevalences of actionable findings as defined by the ACR Work Group in radiology reports. The agreement among radiologists on the classification of actionable findings was moderate. To reduce the workload for radiologists, clinical context may need to be taken into account in deciding whether a finding is actionable or not.

Funding

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

CRediT authorship contribution statement

Jacob J. Visser: Conceptualization, Methodology, Data curation, Writing - original draft, Funding acquisition. **Marianne de Vries:** Data curation, Writing - review & editing. **Jan A. Kors:** Conceptualization, Methodology, Formal analysis, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] American College of Radiology, ACR Practice Parameter for Communication of Diagnostic Imaging Findings, (2014) (Accessed 12 April 2020), https://www.acr.org/-/media/ACR/Files/Practice-Parameters/CommunicationDiag.pdf.
[2] S.G. Anthony, L.M. Prevedello, M.M. Damiano, et al., Impact of a 4-year quality improvement initiative to improve communication of critical imaging test results, Radiology 259 (2011) 802–807, https://doi.org/10.1148/radiol.11101396.
[3] P.A. Larson, L.L. Berland, B. Griffith, et al., Actionable findings and the role of IT support: report of the ACR Actionable Reporting Work Group, J. Am. Coll. Radiol. 11 (2014) 552–558, https://doi.org/10.1016/j.jacr.2013.12.016.
[4] L.S. Babiarz, S. Trotter, V.G. Viertel, et al., Neuroradiology critical findings lists: survey of neuroradiology training programs, Am. J. Neuroradiol. 34 (2013) 735–739, https://doi.org/10.3174/ajnr.A3300.
[5] V.G. Viertel, S.A. Trotter, L.S. Babiarz, et al., Reporting of critical findings in neuroradiology, Am. J. Roentgenol. 200 (2013) 1132–1137, https://doi.org/10.2214/AJR.12.9041.
[6] D.R. Murphy, H. Singh, L. Berlin, Communication breakdowns and diagnostic errors: a radiology perspective, Diagnosis (Berl.) 1 (2014) 253–261, https://doi.org/10.1515/dx-2014-0035.
[7] S. Waite, J.M. Scott, I. Drexler, et al., Communication errors in radiology – pitfalls and how to avoid them, Clin. Imaging 51 (2018) 266–272, https://doi.org/10.1016/j.clinimag.2018.05.025.
[8] J.W. O'Sullivan, T. Muntinga, S. Grigg, et al., Prevalence and outcomes of incidental imaging findings: umbrella review, BMJ 361 (2018) k2387, https://doi.org/10.1136/bmj.k2387.
[9] B. Lumbreras, L. Donat, I. Hernandez-Aguado, Incidental findings in imaging diagnostic tests: a systematic review, Br. J. Radiol. 83 (2010) 276–289, https://doi.org/10.1259/bjr/98067945.
[10] R. Lacson, L.M. Prevedello, K.P. Andriole, et al., Four-year impact of an alert notification system on closed-loop communication of critical test results, Am. J. Roentgenol. 203 (2014) 933–938, https://doi.org/10.2214/AJR.14.13064.
[11] S.J. Baccei, C. DiRoberto, J. Greene, et al., Improving communication of actionable findings in radiology imaging studies and procedures using an EMR-independent system, J. Med. Syst. 43 (2019) 30, https://doi.org/10.1007/s10916-018-1150-z.
[12] P. Stenetorp, S. Pyysalo, G. Topić, et al., Brat: a web-based tool for NLP-assisted text annotation, Proceedings of the Demonstrations Session at EACL, Association for Computational Linguistics (2012) 103–107.
[13] R. Lacson, S.D. O'Connor, V.A. Sahni, et al., Impact of an electronic alert notification system embedded in radiologists' workflow on closed-loop communication of critical results: a time series analysis, BMJ Qual. Saf. 25 (2016) 518–524, https://doi.org/10.1136/bmjqs-2015-004276.

[14] A.D. Pham, A. Neveol, T. Lavergne, et al., Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings, BMC Bioinform. 15 (2014) 266, https://doi.org/10.1186/1471-2105-15-266.
[15] J. Zech, M. Pain, J. Titano, et al., Natural language-based machine learning models for the annotation of clinical radiology reports, Radiology 287 (2018) 570–580, https://doi.org/10.1148/radiol.2018171093.
[16] M.E. Heilbrun, B.E. Chapman, E. Narasimhan, et al., Feasibility of natural language processing-assisted auditing of critical findings in chest radiology, J. Am. Coll. Radiol. 16 (2019) 1299–1304, https://doi.org/10.1016/j.jacr.2019.05.038.
[17] X. Meng, C.H. Ganoe, R.T. Sieberg, et al., Assisting radiologists with reporting urgent findings to referring physicians: a machine learning approach to identify cases for prompt communication, J. Biomed. Inform. 93 (2019) 103169, , https://doi.org/10.1016/j.jbi.2019.103169.
[18] European Society of Radiology, ESR paper on structured reporting in radiology, Insights Imaging 9 (2018) 1–7, https://doi.org/10.1007/s13244-017-0588-8.