Time Requirement and Feasibility of a Systematic Quality Peer Review of Reporting in Radiology

Zeitaufwand und Machbarkeit einer systematischen Qualitätsbeurteilung radiologischer Befunde

Authors

Martin H. Maurer[1], Michael Brönnimann[1], Christophe Schroeder[1], Ehssan Ghadamgahi[2], Florian Streitparth[3], Johannes T. Heverhagen[1], Alexander Leichtle[4], Maximilian de Bucourt[5], Tobias Philipp Meyl[6]

Affiliations

1 Department of Diagnostic, Interventional and Paediatric Radiology, Inselspital, Bern Universtity Hospital, University of Bern, Switzerland
2 Radiology, CT-MRT Institute, Berlin, Germany
3 Department of Radiology, LMU University Hospital, Munich, Germany
4 Institute of Clinical Chemistry, Inselspital, Bern Universtiy Hospital, University of Bern, Switzerland
5 Institute for Diagnostic and Interventional Radiology, Charité University Medicine Berlin, Germany
6 Medical Department, Medical Strategy, Inselspital, Bern University Hospital, University of Bern, Switzerland

Key words

socioeconomic issues, CT, mammography, radiography, health policy and practice, MR imaging

received 12.01.2020
accepted 04.05.2020

Bibliography

DOI https://doi.org/10.1055/a-1178-1113
Published online: 2020
Fortschr Röntgenstr

ISSN 1438-9029

Correspondence

Prof. Dr. Dr. Martin H. Maurer
Department für Diagnostische, Interventionelle und Pädiatrische Radiologie, Universitätsspital Bern, Inselspital, Freiburgstr. 10, 3010 Bern, Switzerland
Tel.: ++ 41/31/6 32 71 80
Fax: ++ 41/31 63 2 48 74
martin.maurer@insel.ch

ZUSAMMENFASSUNG

Zielsetzung Ziel dieser Studie war es, den Aufwand für eine retrospektive Qualitätsprüfung mittels des RADPEER-Systems für verschiedene Prozentsätze der Gesamtmenge an radiologischen Befunden in der Klinik für Radiologie an der Universitätsklinik Bern (Schweiz) abzuschätzen.

Material und Methoden Drei Fachärzte für Radiologie (Bewerter 1 bis 3) bewerteten retrospektiv die Qualität der Befunde von insgesamt 150 radiologischen Untersuchungen (5 verschiedene Untersuchungsarten: Computertomografie (CT) des Abdomens, CT des Thorax, Mammografien, konventionelle Röntgenbilder und abdominale Magnetresonanztomografien (MRT)). Jedem Befund wurde eine RADPEER-Bewertung der Kategorien 1 bis 3 zugewiesen (Kategorie 1: stimmt mit der vorherigen Interpretation überein; Kategorie 2: Diskrepanz in der Interpretation/Beurteilung wäre nicht in jedem Fall zu erwarten gewesen; Kategorie 3: Diskrepanz in der Interpretation/Beurteilung wäre in den meisten Fällen zu erwarten gewesen) und die für jede Überprüfung erforderliche Zeit in Sekunden (s) dokumentiert. Die durchschnittliche Zeit für jede der 5 verschiedenen radiologischen Untersuchungsarten und die entsprechenden Bewertungen von 1 bis 3 wurden verglichen. Eine Sensitivitätsanalyse wurde durchgeführt, um die Gesamtarbeitsbelastung für die Überprüfung verschiedener Prozentsätze des gesamten jährlichen Befundvolumens der Klinik zu berechnen.

Ergebnisse Von den insgesamt 450 analysierten Befunden erhielten 91,1 % (410/450) eine Bewertung von 1 und 8,9 % (40/450) eine Bewertung von 2 oder 3. Die durchschnittliche Zeit (in Sekunden) für eine Bewertung betrug 60,4 s (min. 5 s, max. 245 s). Der Bewerter mit der längsten klinischen Erfahrung benötigte signifikant weniger Zeit für die Prüfung der Befunde als die beiden Gutachter mit kürzerer klinischer Erfahrung ($p < 0,05$). Die durchschnittlichen Bewertungszeiten waren länger für nicht übereinstimmende Bewertungen mit den Kategorien 2 oder 3 ($p < 0,05$). Der Gesamtzeitaufwand für die Überprüfung sämtlicher Befunde der 5 verschiedenen Untersuchungsarten eines Jahres würde mehr als 1200 Arbeitsstunden in Anspruch nehmen.

Schlussfolgerung Eine retrospektive Begutachtung von radiologischen Befundberichten mit dem RADPEER-System erfordert erhebliche personelle Ressourcen. Um die Befundqualität zu verbessern, scheint es jedoch möglich zu sein, zumindest einen Teil der Gesamtbefundungsleistung auch großer radiologischer Abteilungen routinemäßig einer Zweitbegutachtung zu unterziehen.

Thieme

Kernaussagen:

- Eine systematische retrospektive inhaltliche Begutachtung von radiologischen Befunden mit dem RADPEER-System ist mit einem hohen Personalaufwand verbunden.
- Die retrospektive Begutachtung sämtlicher Befunde einer Klinik bzw. Praxis erscheint aufgrund des Mangels an hochspezialisiertem Personal unrealistisch.
- Mit dem Ziel der inhaltlichen Qualitätsverbesserung sollte jedoch zumindest ein Teil aller Befunde einer zweiten retrospektiven Begutachtung unterzogen werden.

ABSTRACT

Objective To estimate the human resources required for a retrospective quality review of different percentages of all routine diagnostic procedures in the Department of Radiology at Bern University Hospital, Switzerland.

Materials and Methods Three board-certified radiologists retrospectively evaluated the quality of the radiological reports of a total of 150 examinations (5 different examination types: abdominal CT, chest CT, mammography, conventional X-ray images and abdominal MRI). Each report was assigned a RADPEER score of 1 to 3 (score 1: concur with previous interpretation; score 2: discrepancy in interpretation/not ordinarily expected to be made; score 3: discrepancy in interpretation/should be made most of the time). The time (in seconds, s) required for each review was documented and compared. A sensitivity analysis was conducted to calculate the total workload for reviewing different percentages of the total annual reporting volume of the clinic.

Results Among the total of 450 reviews analyzed, 91.1 % (410/450) were assigned a score of 1 and 8.9 % (40/450) were assigned scores of 2 or 3. The average time (in seconds) required for a peer review was 60.4 s (min. 5 s, max. 245 s). The reviewer with the greatest clinical experience needed significantly less time for reviewing the reports than the two reviewers with less clinical expertise ($p < 0.05$). Average review times were longer for discrepant ratings with a score of 2 or 3 ($p < 0.05$). The total time requirement calculated for reviewing all 5 types of examination for one year would be more than 1200 working hours.

Conclusion A retrospective peer review of reports of radiological examinations using the RADPEER system requires considerable human resources. However, to improve quality, it seems feasible to peer review at least a portion of the total yearly reporting volume.

Key Points:

- A systematic retrospective assessment of the content of radiological reports using the RADPEER system involves high personnel costs.
- The retrospective assessment of all reports of a clinic or practice seems unrealistic due to the lack of highly specialized personnel.
- At least part of all reports should be reviewed with the aim of improving the quality of reports.

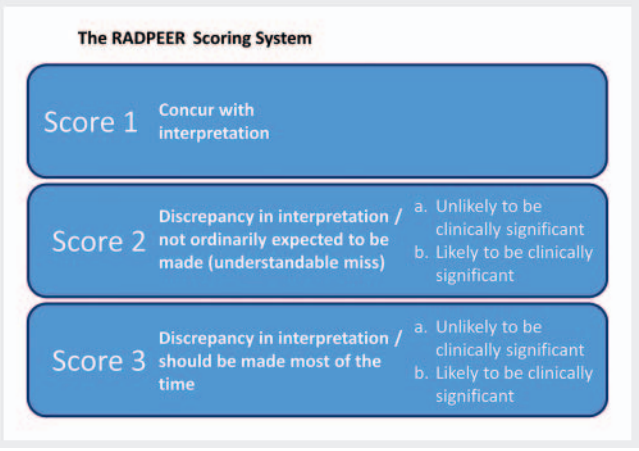
Citation Format

- Maurer MH, Brönnimann M, Schroeder C et al. Time Requirement and Feasibility of a Systematic Quality Peer Review of Reporting in Radiology. Fortschr Röntgenstr 2020; DOI 10.1055/a-1178-1113

Introduction

According to reports by the Institute of Medicine (IOM) in 1999 and 2015, up to 100 000 patients die every year in the USA alone due to avoidable treatment errors and up to 10 % of these cases are caused by diagnostic errors [1, 2]. The regional office for Europe of the World Health Organization (WHO) reported that up to 95 000 deaths per year could be avoided within the European Union by strictly applying strategies to avoid adverse events in patient treatment [3]. In Germany, the scientific institute of the AOK, one of the largest German health insurers, reported up to 18 800 yearly preventable deaths due to treatment errors [4]. In light of this, different attempts have been made to continuously improve quality in radiology, mostly focusing on improving examination procedures and reporting workflows [5–9]. As various specialist disciplines have developed in medicine over the last decades and the degree of subspecialization has massively increased among referring physicians, radiologists are expected not only to have ever-increasing specialized diagnostic knowledge but also to take into account very complex treatment pathways [10–12]. In this context, new approaches aim at improving not only the processes and workflows in radiology, but also the content of radiological reports either by using double reading like in different Scandinavian countries or using systematic retrospective peer reviews [13–15]. In an extensive review, Geijer et al. [14] found discrepancy rates of up to 22 % when comparing reports of double readings.

One of the most widely used systems with the aim of continuously monitoring and improving quality in radiology is RADPEER, which was introduced in the United States on the initiative of the American College of Radiology (ACR) in 2002. This system has been available in a web-based format (known as eRADPEER) since 2005 and was revised in 2009 [16–18]. At present, more than 18 000 participating radiologists in more than 1100 participating clinics and practices use RADPEER in the US. So far, more than 30 million radiological reports have been reviewed using this system [19]. In the recently adapted current version of RADPEER, the retrospective evaluation of radiological reports is carried out on the basis of three categories according to agreement with the previous report [19]: (1.) Score 1: concur with previous interpretation; (2.) Score 2: discrepancy in interpretation/not ordinarily to be made (understandable miss); and (3.) Score 3: discrepancy in interpretation/should be made most of the time. The reviewer assigning Scores 2 and 3 can optionally distin-



▶ **Fig. 1** The RADPEER scoring system.

▶ **Abb. 1** Das RADPEER-Bewertungssystem.

guish between findings that seem to be of unlikely clinical significance for the patient (Scores 2a and 3a) and findings that seem be of likely clinical significance (2b and 3b) (▶ **Fig. 1**).

However, implementation of such review programs is seriously hampered by the current shortage of board-certified radiologists and other qualified staff [20, 21]. In the UK, for example, the Royal College of Radiologists even warned that patient care is at risk due to a severe shortage of trained radiologists [22, 23]. Systematic double reading or peer reviews of radiological examinations like with RADPEER would further aggravate the existing shortage of skilled personnel and make the situation even more threatening in terms of adequate patient care. Furthermore, it is not even clear at present how many additional board-certified radiologists would be needed to establish systematic peer review of even a very small subset of all examinations.

Therefore, the aim of this analysis is to investigate whether a systematic secondary peer review reading in radiology is feasible, taking into account different types of examinations, costs, and the availability of qualified staff.

Methods

For retrospective peer review using the RADPEER system, a total of 150 reports – 30 each of 5 different types of radiological examination (abdominal computed tomography (CT), chest CT, mammograms, conventional X-rays and magnetic resonance imaging (MRI) of the abdomen) performed from January to June 2019 – were randomly selected from the Radiological Information System (RIS) of the Department of Radiology at Bern University Hospital (Switzerland). The corresponding image datasets were retrieved from the department's Picture Archiving and Communication System (PACS). All reports selected for this retrospective analysis were initially written by a radiologist not involved in this study.

All 150 examinations selected for this review were reviewed by three board-certified radiologists (one with 14 years and two with 5 years of clinical experience) to assess reporting quality by reviewing the corresponding imaging dataset and comparing it with the original report. Conventional X-ray examinations consisted of up to 2 radiographies and mammograms of 4 images (craniocaudal (cc) and mediolateral oblique (mlo) views of both sides) each. A selection of suitable image sequences was made in advance for the cross-sectional imaging procedures (CT and MRI), including 4 image series each for CT examinations of the lungs and abdomen and up to 8 image series for MRI examinations. All examinations were hung up in advance in the PACS for a smooth workflow during the assessments. For each of the 150 examinations, each of the three radiologists independently evaluated the report and assigned a score in accordance with the RADPEER system (▶ **Fig. 1**) and also documented the time (in seconds (s)) required for each review.

For each of the three reviewers, the number of examinations to which the Scores 1, 2 and 3 were assigned was documented. The number of Scores 2 and 3 was also documented on a per-patient basis considering single and double mentions. Wherever possible, the clinical outcome of patients was analyzed also considering further follow-up imaging and other clinical examinations. In addition, the average times required for peer review were calculated and compared for all examinations, for each of the five types of imaging examinations and for the three different scores (1 to 3).

To calculate the total effort of the review process, the total time required for all 450 peer reviews was summed up (in seconds and hours). For the subsequent sensitivity analysis, at first, the total number of examinations of the 5 different types of examination that were reviewed was determined for the Department of Radiology in Bern for the entire 2018 annual period. The time required for peer review using RADPEER was then calculated for six different percentages of this total volume of examinations (0.5 %, 1 %, 2 %, 3 %, 5 %, and 10 %).

Data analysis and statistical analysis: For all evaluation times for the three different reviewers, the five different examination types and the three different RADPEER scores (1–3), the mean, minimum (min), maximum (max), median and the quartiles 1 (Q1) and 3 (Q3) were calculated. As normal distribution of data was tested and it was shown that there was no normal distribution of data, an ANOVA test with Tukey's post-hoc test was performed for comparison of the different groups. Statistical significance was assumed for a level of $p < .05$. Results in seconds and minutes were rounded to one decimal place.

Results

For all 450 reviews taken together (150 examinations, each evaluated by three reviewers), the scores were distributed as follows: RADPEER Score 1 was assigned in 410 instances (410/450, 91.1 %), Score 2 in 33 instances (33/450, 7.3 %), and Score 3 in 7 instances (7/450, 1.6 %) (▶ **Table 1**). Reviewer 1 assigned a score of 2 to 2 abdominal CT scans, 2 conventional radiographs, and 1 abdominal MRI. Reviewer 2 assigned a score of 2 to a total of 6 abdominal CT scans, 2 chest CT scans, 1 mammogram, 1 conventional radiograph, and 3 abdominal MRI scans and a score of 3 to 3 abdominal CT scans. Reviewer 3 assigned a score of 2 to a total of 6 abdominal CT scans, 3 chest CT scans, 1 mammogram, 2 radiographs and 3 abdominal MRI scans and a score of 3 to 3 ab-

▶ **Table 1** Distribution of RADPEER scores 1 to 3 for all peer reviews (total of 450 including percentages) and for each of the three reviewers. Analysis also on a per-patient basis (n = 150), examples and outcome given for scores 2 and 3.

▶ **Tab. 1** Verteilung der verschieden RADPEER-Bewertungen 1 bis 3 (insgesamt n = 450 inkl. prozentuale Verteilung) durch die 3 verschiedenen Bewerter. Auswertung auch auf Basis der Patientenzahl (n = 150), Beispiele und weiterer klinischer Verlauf bei Bewertungen mit den Scores 2 und 3.

n = 450	score 1 (n/percentages)	score 2 (n/percentages)	score 3 (n/percentages)
all reviewers	410 (91.1 %)	33 (7.3 %)	7 (1.6 %)
reviewer 1	145 (32.2 %)	5 (1.1 %)	–
reviewer 2	134 (29.8 %)	13 (2.9 %)	3 (0.6 %)
reviewer 3	131 (29.1 %)	15 (3.3 %)	4 (0.8 %)
per-patient basis (n = 150) considering single and double mentions	129 (86 %)	17 (11.3 %)	4 (2.7 %)
Examples (and outcomes)		Aneurysm of the abdominal aorta (stable for at least 5 years) Possible pleural empyema (not confirmed/ resolved in follow-up imaging and on clinical examination) Possible diverticulitis (not confirmed/ resolved in follow-up) Intrahepatic bilioma (correct, without therapeutic consequences) Small fracture of the clavicula (correct, without therapeutic consequences) Kidney cyst should be rated as Bosniak IIF (case was presented in X-ray demonstration, MRI was recommended which did not confirm the diagnosis) Cardiomegaly in chest X-ray not mentioned (correct, elderly patient with known cardiomegaly, under therapy) Residual abdominal abscess not mentioned (patient was known to have a history of abdominal abscess which had already nearly completely resolved before, not mentioning the residuum was without consequence for the patient)	Double duct sign in pancreas (retrospective analysis showed a stable situation for 3 years, endo-ultrasound negative) Possible small carcinoma of urinary bladder wall (patient with known recurrent small bladder wall carcinoma, tumor was confirmed histologically after a biopsy during cystoscopy) Diverticulitis of the sigma (complete restitution in the follow-up) Increasing intrahepatic cholestasis (stable situation for years, no clear cause found)

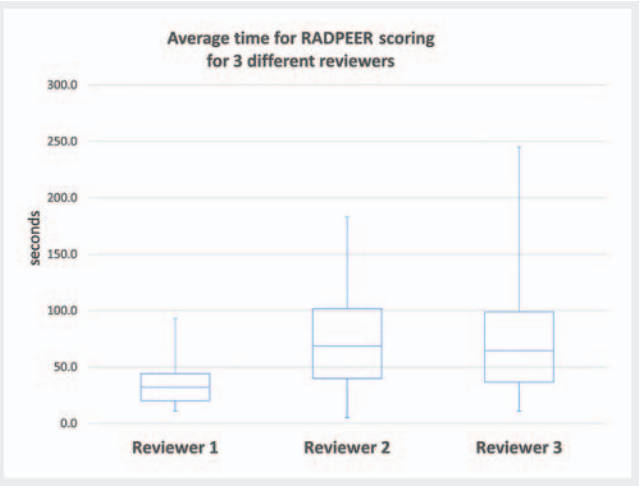
dominal CT scans and 1 abdominal MRI scan. On a per-patient basis, there were 17 examinations rated with a score of 2, and 4 examinations with a score of 3 (▶ **Table 1**).

In our peer review study, Score 2, for example, was assigned in a patient with multiple cystic liver lesions. The radiologist writing the initial report failed to assess whether this might be an echinococcosis infestation of the liver, although this was explicitly asked in the indication. Score 3, for example, was assigned in a patient with pelvic pain in whom diverticulitis was not recognized as a possible cause. Further examples and the clinical outcomes are given in ▶ **Table 1**.

The average time to review a radiological examination regardless of the score assigned was 60.4 seconds (min.: 5 s; max.: 245 s) for all types of examination and all three reviewers taken together. The mean evaluation time was 34.3 seconds for reviewer 1 (min.: 11 s; max.: 93 s; median: 32 s; Q1: 20 s; Q3: 44 s), 74.4 seconds for reviewer 2 (min.: 5 s; max.: 183 s; median: 68.5 s; Q1: 39.5 s; Q3: 101.8 s) and 72.5 seconds for reviewer 3 (min.: 11 s; max.: 245 s; median: 64.5 s; Q1: 36.5 s; Q3: 98.8 s) (▶ **Fig. 2**). Compared with reviewer 1, the average review time was significantly longer for both reviewer 2 and reviewer 3 ($p < 0.05$). On the other hand, reviewers 2 and 3 do not differ significantly with respect to their required assessment times ($p = 0.79$).

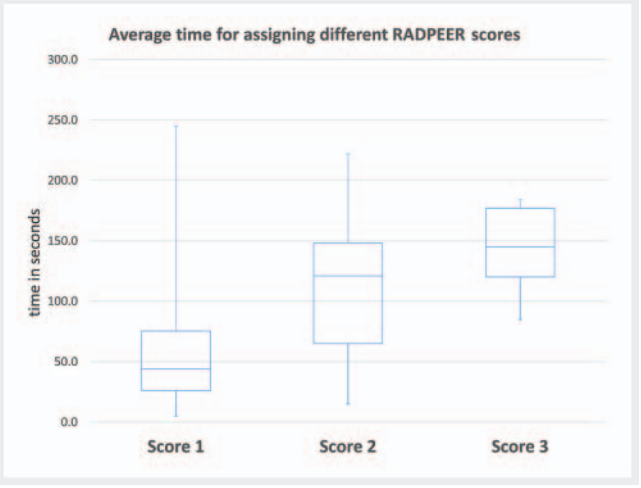
The mean review times varied significantly with the scores assigned. Averaging 54.9 seconds for Score 1 (range: 5 s–245 s; median: 44 s; Q1: 26 s; Q3: 75.3 s), mean review times were significantly longer for Score 2 (mean: 105.3 s; range: 15 s–212 s; median: 111 s; Q1: 65 s; Q3: 138 s) and Score 3 (mean: 144.2 s; range: 85 s–184 s; median: 145 s; Q1: 120 s; Q3: 177 s) for all three reviewers ($p < .05$ each) (▶ **Fig. 3**). In contrast, on average, the evaluation with a Score of 3 did not take significantly longer than a Score of 2 ($p = 0.11$).

The distribution of review times according to the different types of imaging examination is given in ▶ **Fig. 4**. The average



▶ **Fig. 2** Average time for RADPEER scoring for three different reviewers.

▶ **Abb. 2** Benötigte Durchschnittszeit für die RADPEER-Bewertungen durch die 3 verschiedenen Bewerter.

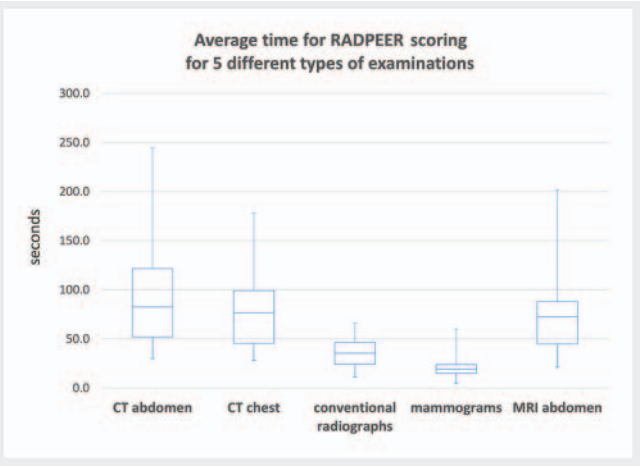


▶ **Fig. 3** Average time for assigning different RADPEER scores.

▶ **Abb. 3** Durchschnittszeiten für die RADPEER-Bewertungen für eine der 3 verschiedenen Bewertungsklassen.

time required for reviewing mammograms (mean: 20.3 s; range: 5–60 s; median: 19s; Q1: 15 s; Q3: 23.8 s) and conventional radiographs (mean: 36.3 s; range: 11–66 s; median: 35.5 s; Q1: 24 s; Q3: 46.5 s) was significantly shorter compared with the average time for each of the three cross-sectional imaging methods (abdominal CT (mean: 89.6 s; range: 30–245 s; median: 82.5 s; Q1: 52.5 s; Q3: 121.5 s), abdominal MRI (mean: 77 s; range: 21–212 s; median: 72.5 s; Q1: 44.8 s; Q3: 98 s) and chest CT (mean: 78.9 s; range: 28–178 s; median: 76.5 s; Q1: 45.5 s; Q3: 99 s)).

A sensitivity analysis was performed to determine the relationship between the overall time needed for peer review of the 5 types of examination that were analyzed and their percentage share of all radiological examinations of the full year in 2018. At the Department of Radiology, University Hospital of Bern, a total of 6193 abdominal CT scans, 5274 chest CT scans, 2902 mammograms, 85 378 conventional radiographies, and 3480 abdominal MRI examinations were performed in 2018 (equivalent to 100 %) (▶ **Table 2**). For the sample of 450 reviews (90 of each type), a total review time of 27 198 seconds (equivalent to 7.55 hours) was calculated. These data and results were the basis for calculating the total time expenditure that would be required for a peer review covering various percentages of all performed examinations. For example, a peer review of 1 % of the 5 types of examination selected for the current analysis would require approx. 12.2 hours; for peer review of 5 % of the examinations, the time needed would be approx. 61.1 hours. In the extreme case of double reading all examinations, the time required would be approx. 1221.5 hours.



▶ **Fig. 4** Average time for RADPEER scoring for five different types of examination.

▶ **Abb. 4** Durchschnittliche Zeitdauer für die Bewertung von 5 verschiedenen Untersuchungsarten mit dem RADPEER-Bewertungssystem.

Discussion

Our systematic peer review of radiological reports reveals that, if the RADPEER system were used to review all reports of 5 selected types of imaging examination performed during a one-year period (2018, n = 102 227) at Bern University Hospital, more than 1200 hours of work by specialized radiologists would be needed. Performing peer reviews of smaller percentages of all examinations would be less time-consuming. The average time per peer review was about 60 seconds.

As expected, the distribution of the 3 RADPEER scores shows that the vast majority of reports were found to be acceptable (Score 1, (410/450, 91.1 %) (▶ **Table 1**)). However, there was a higher proportion of reviews with assignment of scores of 2 (33/450, 7.3 %) or even 3 (7/450, 1.5 %) compared to the published data. Jackson et al. [17] found a total disagreement rate of 2.91 % with Score 2 being assigned in 2.51 % and Score 3 in 0.39 % of all cases reviewed using the RADPEER system. Soffa et al. [24] found an overall disagreement rate of 3.48 % in a total of 6703 ex-

▶ **Table 2** Sensitivity analysis for calculation of the total time requirement for peer reviewing various percentages of five types of radiological examination using RADPEER. CT = computed tomography; MRI = magnetic resonance imaging.

▶ **Tab. 2** Sensitivitätsanalyse zur Berechnung des Gesamtzeitbedarfs für die Begutachtung verschiedener Prozentsätze von 5 Arten radiologischer Untersuchungen mit dem RADPEER-System. CT = Computertomografie; MRT = Magnetresonanztomografie.

number of examinations	percentage of overall examinations	overall review time in seconds (and in hours)
6193 abdominal CT scans 5274 chest CT scans 2902 mammograms 85 378 conventional radiographies 3480 abdominal MRI examinations	100 %	4397 600 (1221.5)
90 abdominal CT scans 90 chest CT scans 90 mammograms 90 conventional radiographies 90 abdominal MRI examinations	0.145 % 0.171 % 0.311 % 0.011 % 0.259 %	27 198 (7.55)
30.9 abdominal CT scans 26.4 chest CT scans 14.5 mammograms 427 conventional radiographies 17.4 abdominal MRI examinations	0.5 %	21 988 (6.1)
61.9 abdominal CT scans 52.7 chest CT scans 29.0 mammograms 854 conventional radiographies 34.8 abdominal MRI examinations	1 %	43 976 (12.2)
123.8 abdominal CT scans 52.8 chest CT scans 58 mammograms 854 conventional radiographies 34.8 abdominal MRI examinations	2 %	87 952 (24.4)
etc.	3 %	131 928 (36.6)
etc.	5 %	219 880 (61.1)
etc.	10 %	439 760 (122.2)

aminations evaluated by 26 different radiologists. Swanson et al. [31] identified a discrepancy rate of 3.8 % between the original interpretation and the peer review (Score 2: 3.6 %, Score 3: 0.2 %). In our study, the overall disagreement rate (8.9 %) was almost twice as high as in the quoted studies (▶ **Table 1**). However, it is worth noting that the most experienced reviewer in our analysis (reviewer 1 with 14 years of clinical experience in diagnostic radiology) assigned Score 2 in 5 of 150 reviews (3.3 %) while never assigning Score 3. The results for this reviewer are roughly comparable to the findings of other large studies. In contrast, reviewers 2 and 3 had a Score 2 rate of 9.3 % (28 of 300 reports) and a Score 3 rate of 2.3 % (7 of 300 reports). A possible explanation for this difference between more experienced and less experienced radiologists might be that the latter are more cautious and therefore more readily assign a disagreement score to be on the safe side.

In addition, the two less experienced reviewers (2 and 3 in our study) required significantly more time for the reviews (74.4 s and 72.5 s vs. 34.3 s for reviewer 1) (▶ **Fig. 2**). Again, this observation suggests that there is a correlation with clinical experience. Of note, all 3 reviewers took significantly longer for peer review when they assigned a disagreement score (2 or 3) compared with reviews assigned a score of 1 ($p < 0.05$ each) (▶ **Fig. 3**). This was true for all three reviewers even when comparing the average review time for a score of 2 compared to that for a score of 1 and for reviewers 2 and 3 when comparing the times for a score of 3 compared to a score of 1. This suggests that, overall, it takes significantly longer to allocate a disagreement score of 2 or 3. Before such scores are assigned, the reviewers have to thoroughly compare images and findings, identify, and weigh possible deficiencies, and then assign the rating.

It is also plausible that it takes longer to review reports of cross-sectional imaging examinations, such as CT and MRI, compared with mammograms and conventional X-ray images. This was consistently observed for all three reviewers in our study (an average of 20.3 s for the evaluation of a mammogram or 36.3 s for a conventional radiography vs. 60.4 s on average for all other types of examination) as well as for each reviewer considered separately (▶ **Fig. 4**). This can be explained by the fact that a CT examination generates a vast number of thin-layer images, which allow multi-dimensional reformation but also take longer to interpret. Modern MRI examinations with specific protocols, e. g., for evaluation of the liver, including more than 20 different pulse sequences are even more complex.

To retrospectively peer review all five examination types analyzed here for a whole year (2018) at Bern University Hospital would involve a workload of more than 1200 hours or the equivalent work of half of a full-time radiologist. This is not realistic under routine conditions. It should also be borne in mind that the peer reviews in the framework of this study were carried out under optimal conditions. If one were to integrate the peer review process as proposed by the ACR into the daily reporting process, this would probably require even more time. It must also be taken into consideration that the workflow of radiologists is already highly fragmented. If double reading using RADPEER is added to the already heavy workload of a radiologist at a university hospital, this can lead to faster fatigue and decreasing quality of subsequent reports. Under these conditions, a routine peer review system would not improve quality but might even have the opposite effect.

Alternatively, one should consider systematic peer review of only part of all examinations. If only 10 % of the 5 types of examinations analyzed here were evaluated, the time required would be around 120 hours, which seems to be quite reasonable (▶ **Table 2**). To select cases with the highest probability of errors, Sheu et al. [25] created a mathematical model for radiologists based on past frequencies of interpretive errors. In addition, the

time required for peer review could be reduced further by working exclusively with very experienced reviewers. In our study, the average review time for the most experienced radiologist was almost 50 % lower than for the two less experienced reviewers. However, even if the workload can be reduced by the measures just outlined, double reading involves significant staff costs. In addition, there are always license fees for the use of RADPEER.

Ultimately, in addition to calculating the additional costs and effort, as carried out in our study, the potential benefit of a double reading has to be weighed against its potential benefits for patients. RADPEER was first introduced to identify patterns in frequently observed errors and to learn from past mistakes to prevent these in the future [9, 10]. What is important for radiologists is to be aware that, while reporting errors can occur, there are efforts to minimize them [17]. Quality measures can be adapted to the local requirements that can be very specific in relation to the referrers [26]. From Scandinavia, where in different countries (e. g., Sweden and Norway) double reading for different radiological examination procedures is widespread, extensive data is available which show high discrepancy rates of up to over 20 % between the compared radiological reports [14, 15]. As an added value for patients, an increased detection rate of breast cancer was demonstrated when using a systematic double reading of mammograms in breast cancer screening [27]. Moreover, Lauritzen et al. [28] found that double reading of chest CT examinations led to important changes in the clinical treatment pathway in 9 % of all patients, whereas in CT examinations of the abdomen even in 14 % of all patients a significant change in the clinical treatment path was due to the result of the double reading [29]. In trauma CT examinations, missing findings were found in up to 47 % of all patients comparing the initial short reports and the final reports, even though many of these discrepancies were findings of the musculoskeletal system with minor importance [30]. However, in the context of trauma, Yoon et al. [31] found significant changes in patient management based on double reading of trauma CT examinations in 7.8 % of all patients.

However, it has to be considered that besides RADPEER, there are other well-established measures to improve the quality of radiological reports like daily clinical demonstrations and tumor boards that include as in the case of RADPEER a “second look” on radiological examinations, often by a different radiologist [32]. Morbidity & mortality conferences provide a systematic approach to analyze errors in radiology with the aim of reducing these in the future. All these measures to improve quality can even be used for advertising purposes. For patients, it may be important to know that, in addition to various quality certificates of the processes, the actual content of radiological reports is rigorously checked.

Our study has various limitations. First of all, it has to be considered that the RADPEER scoring system was meant to be carried out during everyday work in clinics or practices in parallel with the ongoing reporting process. However, the effort required for the RADPEER scores is not provided but should not be neglected. In this regard, the sample analyzed in this study is small compared with the data that have so far been collected with RADPEER from a large number of sites over a period of many years. Therefore, our results for the small subsets of reviews assigned scores of 2 or 3 should be interpreted with caution. However, our primary aim was to obtain a basis for determining whether a systematic peer review of a certain percentage of the total radiological examination volume in a large clinic would be feasible, especially when taking into account the current lack of radiologically qualified staff for such a task. It was intended to provide guidance to quickly estimate the additional expenses that can be expected in radiological institutes of variable size when using RADPEER as a quality measure. Our approach to use RADPEER scoring slightly differs from the “traditional” use during the ongoing reporting procedure. However, both are retrospective assessment methods. Our method might even underestimate the personnel effort, as imaging material in the PACS was probably more available than during ongoing reporting processes. The usual approach of using RADPEER also leads to faster radiologist fatigue as the reporting workflow is repeatedly interrupted. Finally, only five different types of examinations were included, accounting for about 60 % of the total volume at Bern University Hospital. We intended to use a sample of particularly common types of examinations. Certainly, other types of examination could easily be taken into account, which would of course take even more personnel effort.

In summary, performing a retrospective peer review of radiology reports using the RADPEER system is associated with significant personnel effort. However, if one optimizes the workflow of the review process and only reviews a certain proportion of the total volume of imaging examinations performed by an institution, the effort can be quite feasible, and its benefits may justify the expenditure incurred by routine peer review.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Institute of Medicine Committee on Quality of Health Care in A. In: Kohn LT, Corrigan JM, Donaldson MS, Hrsg To Err is Human: Building a Safer Health System. Washington (DC): National Academies Press (US); 2000

[2] Committee on Diagnostic Error in Health C, Board on Health Care S, Institute of M et al.In: Balogh EP, Miller BT, Ball JR, Hrsg Improving Diagnosis in Health Care. Washington (DC): National Academies Press (US); 2015

[3] World Health Organisation (WHO). http://www.euro.who.int/en/health-topics/Health-systems/patient-safety/data-and-statistics

[4] Klauber J, Geraedts M, Friedrich J et al. Krankenhausreport 2014 – Schwerpunkt Patientensicherheit. Stuttgart: Schattauer Verlag; 2014

[5] Maurer MH, Hamm B, Teichgraber U. ServiceBlueprinting as a service management tool in radiology. European journal of radiology 2011; 79: 333–336

[6] Pianykh OS, Jaworsky C, Shore MT et al. Improving Radiology Workflow with Automated Examination Tracking and Alerts. Journal of the American College of Radiology: JACR 2017; 14: 937–943

[7] Kansagra AP, Liu K, Yu JP. Disruption of Radiologist Workflow. Curr Probl Diagn Radiol 2016; 45: 101–106

[8] Swensen SJ, Johnson CD. Radiologic quality and safety: mapping value into radiology. Journal of the American College of Radiology: JACR 2005; 2: 992–1000

[9] Tamm EP, Szklaruk J, Puthooran L et al. Quality initiatives: planning, setting up, and carrying out radiology process improvement projects. Radiographics: a review publication of the Radiological Society of North America, Inc 2012; 32: 1529–1542

[10] Mabotuwana T, Hall CS, Flacke S et al. Inpatient Complexity in Radiology-a Practical Application of the Case Mix Index Metric. J Digit Imaging 2017; 30: 301–308

[11] Margulis AR. Subspecialization and certification in radiology. Am J Roentgenol 1992; 159: 1113–1114

[12] Meyl TP, de Bucourt M, Berghofer A et al. Subspecialization in radiology: effects on the diagnostic spectrum of radiologists and report turnaround time in a Swiss university hospital. Radiol Med 2019; 124: 860–869

[13] Gollub MJ, Panicek DM, Bach AM et al. Clinical importance of reinterpretation of body CT scans obtained elsewhere in patients referred for care at a tertiary cancer center. Radiology 1999; 210: 109–112

[14] Geijer H, Geijer M. Added value of double reading in diagnostic radiology,a systematic review. Insights Imaging 2018; 9: 287–301

[15] Lauritzen PM, Hurlen P, Sandbaek G et al. Double reading rates and quality assurance practices in Norwegian hospital radiology departments: two parallel national surveys. Acta radiologica 2015; 56: 78–86

[16] Borgstede JP, Lewis RS, Bhargavan M et al. RADPEER quality assurance program: a multifacility study of interpretive disagreement rates. Journal of the American College of Radiology: JACR 2004; 1: 59–65

[17] Jackson VP, Cushing T, Abujudeh HH et al. RADPEER scoring white paper. Journal of the American College of Radiology: JACR 2009; 6: 21–25

[18] Abujudeh H, Pyatt RS Jr, Bruno MA et al. RADPEER peer review: relevance, use, concerns, challenges, and direction forward. Journal of the American College of Radiology: JACR 2014; 11: 899–904

[19] Goldberg-Stein S, Frigini LA, Long S et al. ACR RADPEER Committee White Paper with 2016 Updates: Revised Scoring System, New Classifications, Self-Review, and Subspecialized Reports. Journal of the American College of Radiology: JACR 2017; 14: 1080–1086

[20] Moriarity AK, Brown ML, Schultz LR. Work and retirement preferences of practicing radiologists as a predictor of workforce needs. Acad Radiol 2014; 21: 1067–1071

[21] Saket DD. The provision of emergency radiology services and potential radiologist workforce crisis: is there a role for the emergency-dedicated radiologist? Semin Ultrasound CT MR 2007; 28: 81–84

[22] Gourd E. UK radiologist staffing crisis reaches critical levels. Lancet Oncol 2017; 18: e651

[23] Mooney H. More radiologists needed for improved cancer diagnosis, says royal college. Bmj 2016; 353: i2718

[24] Soffa DJ, Lewis RS, Sunshine JH et al. Disagreement in interpretation: a method for the development of benchmarks for quality assurance in imaging. Journal of the American College of Radiology: JACR 2004; 1: 212–217

[25] Sheu YR, Feder E, Balsim I et al. Optimizing radiology peer review: a mathematical model for selecting future cases based on prior errors. Journal of the American College of Radiology: JACR 2010; 7: 431–438

[26] Larson PA, Pyatt RS Jr, Grimes CK et al. Getting the most out of RADPEER. Journal of the American College of Radiology: JACR 2011; 8: 543–548

[27] Hofvind S, Geller BM, Rosenberg RD et al. Screening-detected breast cancers: discordant independent double reading in a population-based screening program. Radiology 2009; 253: 652–660

[28] Lauritzen PM, Stavem K, Andersen JG et al. Double reading of current chest CT examinations: Clinical importance of changes to radiology reports. European journal of radiology 2016; 85: 199–204

[29] Lauritzen PM, Andersen JG, Stokke MV et al. Radiologist-initiated double reading of abdominal CT: retrospective analysis of the clinical importance of changes to radiology reports. BMJ quality & safety 2016; 25: 595–603

[30] Eurin M, Haddad N, Zappa M et al. Incidence and predictors of missed injuries in trauma patients in the initial hot report of whole-body CT scan. Injury 2012; 43: 73–77

[31] Yoon LS, Haims AH, Brink JA et al. Evaluation of an emergency radiology quality assurance program at a level I trauma center: abdominal and pelvic CT studies. Radiology 2002; 224: 42–46

[32] Kelly AM, Cronin P. Practical Approaches to Quality Improvement for Radiologists. Radiographics: a review publication of the Radiological Society of North America, Inc 2015; 35: 1630–1642