OPEN ACCESS

Radiologist-initiated double reading of abdominal CT: retrospective analysis of the clinical importance of changes to radiology reports

Peter Mæhre Lauritzen,[1,2] Jack Gunnar Andersen,[3] Mali Victoria Stokke,[4] Anne Lise Tennstrand,[5] Rolf Aamodt,[6] Thomas Heggelund,[6] Fredrik A Dahl,[7] Gunnar Sandbæk,[3,8] Petter Hurlen,[1] Pål Gulbrandsen[2,7]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2015-004536).

For numbered affiliations see end of article.

Correspondence to
Peter Mæhre Lauritzen, Department of Diagnostic Imaging, Akershus University Hospital, P. O. Box 1000, Lørenskog 1478, Norway; peter.m.lauritzen@gmail.com

Received 26 June 2015
Revised 8 January 2016
Accepted 21 January 2016
Published Online First
24 March 2016

Linked

▸ http://dx.doi.org/10.1136/bmjqs-2016-005301

CrossMark

To cite: Lauritzen PM, Andersen JG, Stokke MV, *et al. BMJ Qual Saf* 2016;**25**:595–603.

ABSTRACT

Background Misinterpretation of radiological examinations is an important contributing factor to diagnostic errors. Consultant radiologists in Norwegian hospitals frequently request second reads by colleagues in real time. Our objective was to estimate the frequency of clinically important changes to radiology reports produced by these prospectively obtained double readings.

Methods We retrospectively compared the preliminary and final reports from 1071 consecutive double-read abdominal CT examinations of surgical patients at five public hospitals in Norway. Experienced gastrointestinal surgeons rated the clinical importance of changes from the preliminary to final report. The severity of the radiological findings in clinically important changes was classified as increased, unchanged or decreased.

Results Changes were classified as clinically important in 146 of 1071 reports (14%). Changes to 3 reports (0.3%) were critical (demanding immediate action), 35 (3%) were major (implying a change in treatment) and 108 (10%) were intermediate (requiring further investigations). The severity of the radiological findings was increased in 118 (81%) of the clinically important changes. Important changes were made less frequently when abdominal radiologists were first readers, more frequently when they were second readers, and more frequently to urgent examinations.

Conclusion A 14% rate of clinically important changes made during double reading may justify quality assurance of radiological interpretation. Using expert second readers and a targeted selection of urgent cases and radiologists reading outside their specialty may increase the yield of discrepant cases.

INTRODUCTION

Surgeons often rely on radiology as a source of diagnostic information in the work-up and follow-up of their patients. Because the radiologists who interpret the examinations are human beings, they are not exempt from discrepancies or even error. The reports: 'To err is human' and 'An Organization with a Memory' increased awareness of medical errors and the importance of learning from them.[1, 2] An autopsy study of patients dying in hospital showed that radiological misinterpretation caused 8% and contributed to another 33% of diagnostic errors in patients with relevant imaging.[3] In a recent report, the Institute of Medicine finds that the occurrence of diagnostic errors has been largely unappreciated in efforts to improve the quality and safety of healthcare.[4]

Double reading is a practice in which two readers interpret an imaging examination that reduces errors and increases sensitivity.[5] Although the concept is simple, double reading can be conducted in several ways. There are large variations in the reported effect of double reading in different settings, and the cost effectiveness is not well established.[6–8] Applied prospectively, it may be used for quality assurance of radiology reports, and it is routine in the education of residents.[9, 10] Some mammography screening programmes conduct independent double reading, in which the readers are blinded to the interpretation of their colleague.[11]

In the USA, it is a requirement for department credentialing by the Joint Commission on Accreditation of

Healthcare Organizations that all staff participate in continuous peer review of 5% of randomly selected cases.[12] In order to meet this standard, and to minimise its impact on workflow, peer review programmes such as RADPEER use retrospective double reading (review) of previous examinations when they are compared with the current ones being interpreted.[13] The reviewing radiologist selects the examinations, and the goals are quality improvement through shared learning from discrepancies and benchmarking of performance, rather than quality assurance of the individual report.

Similarly, in the UK, The Royal College of Radiology recommends that all radiology departments aim to implement 'peer feedback' with a systematic review of 5% of reports by December 2018, and that this effort should be coupled with regular 'Learning from Discrepancies meetings'.[14 15]

In Norway, the approach to double reading in clinical radiology is somewhat different. When reading an examination, a consultant radiologist may choose to finalise the report directly or to request a second reading.[9] The decision is based on the consultant's judgement of whether this quality assurance is warranted or not. The request may be explicit by directly contacting a specific colleague, or implicit by choosing not to sign the report, in which case the examination is routed to a queue for second reading. Fellow consultants at the same hospital carry out the second readings, and most consultants contribute as second readers, usually within their own field of expertise. Second readers have access to the preliminary report and updated information in the electronic patient record. The preliminary report, which is available in the electronic patient record, is substituted by the final report when the second reading is completed.

Consultant radiologists in Norwegian hospitals submit 39% of CT examinations for a second reading in this manner.[16] For all examination techniques together, the practice consumes 20%–25% of consultant working hours.[16] The main goal is quality assurance of the report before it is finalised. Less than 10% of departments record discrepancy rates or engage in benchmarking of radiologist performance.[16]

The objective of this study was to estimate the proportion of radiology reports that were changed during prospective double reading of current abdominal CT examinations of surgical patients and to assess the potential clinical impact of these changes. We also aimed to explore whether characteristics of examinations or radiologists were associated with a higher proportion of clinically important changes.

METHODS

Study design

In this retrospective multicentre study, preliminary and final radiology reports from 1071 consecutive double-read abdominal CT examinations were collected and compared for changes (figure 1). Experienced gastrointestinal surgeons rated the clinical importance of the changes made to radiology reports following double reading. In order for the clinical raters to act within their area of expertise, all patients were inpatients or outpatients from the department of surgery and were aged 18 years or older. We only included examinations of the entire abdominal cavity (excluding isolated examinations of the liver). Repeated examinations on the same patient were not included.

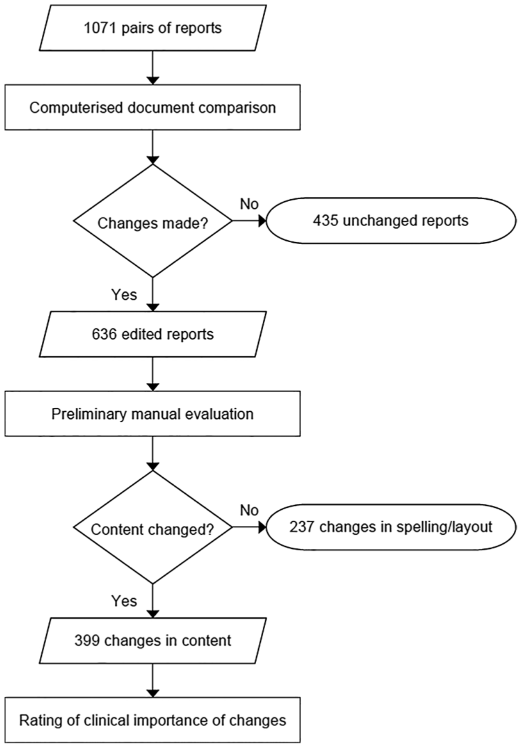
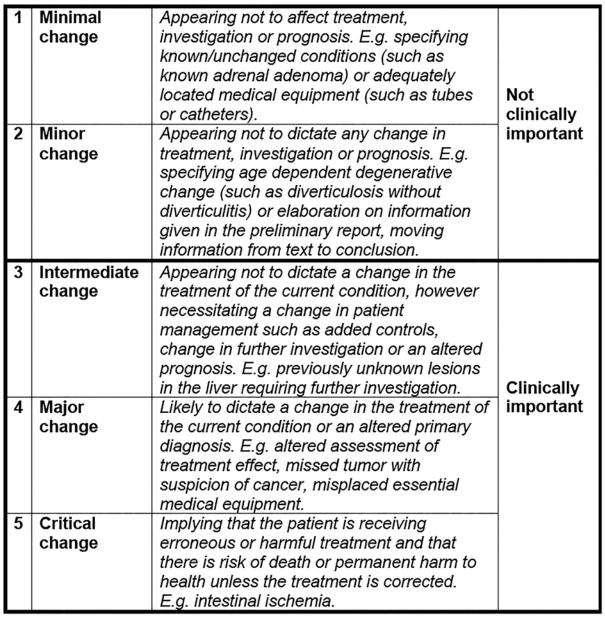

1071 pairs of reports
Computerised document comparison
Changes made?
No
435 unchanged reports
Yes
636 edited reports
Preliminary manual evaluation
Content changed?
No
237 changes in spelling/layout
Yes
399 changes in content
Rating of clinical importance of changes

Figure 1 Selection of radiology reports for clinical rating.

Data were collected from the Radiology Information System and Electronic Patient Records at five public hospitals with a combined catchment population of 1.2 million. The number of reports collected from each hospital was in relative proportion to the number of consultant full-time equivalents in the radiology department. All included examinations were conducted between 1 September 2011 and 27 March 2013, and had been double read by two consultant radiologists as routine quality assurance. The first reader selected which examinations to submit for this quality assurance according to their own judgement, as there are no established selection criteria. Accordingly, the reasons for submitting and the number of examinations submitted vary among radiologists. Approval for the study and waiver of informed consent was obtained from the Regional Ethics Committee and the Data Protection Officer.

Patient and examination data

We collected data on patient gender and age, inpatient/outpatient status, urgency of examination



1	**Minimal change**	*Appearing not to affect treatment, investigation or prognosis. E.g. specifying known/unchanged conditions (such as known adrenal adenoma) or adequately located medical equipment (such as tubes or catheters).*	**Not clinically important**
2	**Minor change**	*Appearing not to dictate any change in treatment, investigation or prognosis. E.g. specifying age dependent degenerative change (such as diverticulosis without diverticulitis) or elaboration on information given in the preliminary report, moving information from text to conclusion.*	
3	**Intermediate change**	*Appearing not to dictate a change in the treatment of the current condition, however necessitating a change in patient management such as added controls, change in further investigation or an altered prognosis. E.g. previously unknown lesions in the liver requiring further investigation.*	**Clinically important**
4	**Major change**	*Likely to dictate a change in the treatment of the current condition or an altered primary diagnosis. E.g. altered assessment of treatment effect, missed tumor with suspicion of cancer, misplaced essential medical equipment.*	
5	**Critical change**	*Implying that the patient is receiving erroneous or harmful treatment and that there is risk of death or permanent harm to health unless the treatment is corrected. E.g. intestinal ischemia.*	

Figure 2 Clinical importance of rating scale.

(routine or urgent, defined as requested within 24 h), referral information, the identities of the first reader and second reader and the time of examination, time of preliminary and final reports (during working hours: 7:00 to 16:00, or out of working hours).

Text comparison

The pairs of preliminary and final reports were compared using 'Diff Doc Professional' (Softinterface, Los Angeles, California, USA), document comparison software, which labelled deletions, additions and changes in the reports by colour coding.

Clinical rating

All radiology reports with changes in content beyond simple corrections of misspelling and layout were submitted for clinical rating (figure 1). Two gastrointestinal surgeons independently rated the clinical importance of changes in content to the reports on an ordinal five-point scale. We designed the scale with the intention to be dichotomised in the statistical analysis (figure 2).

Report changes given discrepant ratings of two or lower by both raters were classified as 'clinically not important' and not resolved further. All discrepancies rated three or higher by at least one rater were resolved by obtaining a clinical rating from a third surgeon, and clinical importance was classified according to the median of the three ratings.

The three raters were specialists in gastrointestinal surgery, all with >10 years of surgical experience. They made their rating based on the radiology report with colour-coded changes, the referral and the patients' age and gender. To reduce bias, the source hospital of the reports were not disclosed to the raters and reports from the five hospitals were presented in a mixed sequence.

Clinical context

In addition to the rating, the surgeons made written comments about the assumed consequences of the changes they rated clinically important. With the aid of these comments we classified clinically important changes according to the clinical issues concerned. We also distinguished between increased, unchanged and decreased severity of the radiological findings resulting from clinically important changes. Changes considered an increase in severity were additional pathological findings or diagnostic suggestions leading to more comprehensive investigations or treatment. Changes considered a decrease in severity were removal or downgrading of initially reported pathological findings. Some changes could not be classified as either and were labelled unchanged severity.

We wished to explore the impact of reasons for referral on the frequency of clinically important changes. The first author reviewed referrals, and classified reasons for referral into four groups: acute presentations, non-acute presentations, follow-up and investigations after surgery or invasive procedures.

Radiologists

We classified the involved consultant radiologists based on experience as a consultant and subspecialty into four groups: inexperienced (<3 years as a consultant), general radiologist (≥3 years, not working within a limited field of expertise), abdominal radiologist (≥3 years, working predominantly with abdominal imaging) and other subspecialist (≥3 years, working within any other limited field of expertise).

Statistical analysis

The inter-rater agreement for the five-point scale was assessed using raw agreement and weighted κ.[17] We used a weight of $1-[(i-j)/(k-1)]^2$, where 'i' and 'j' index the rows and columns of the ratings by the two raters, and 'k' is the maximum number of possible ratings. Differences in ratings between the two initial raters were tested with a related samples Wilcoxon signed-rank test. Agreement and Cohen's κ were calculated for the dichotomised ratings.

Exploratory analysis of associations between clinical importance of changes and characteristics of patients, examinations and readers was performed with univariate logistic regression. Variables whose univariate test had a p value of <0.25 were entered as candidate variables in a multivariate logistic regression model. Subsequently, a stepwise removal of the candidate variable with highest p value was performed until only statistically significant variables remained.

Associations between reasons for referral and clinically important changes were explored by univariate logistic regression. The classification of reasons for referral is not a readily available parameter in a quality assurance setting, and we expected considerable overlap with more robust patient parameters such as urgency, admission status and examination time. Therefore we decided not to enter reasons for referral into the multivariate model.

We constructed two random effects logistic regression models to assess a possible association between readings of separate examinations by the same radiologist. The models tested whether there was clustering of clinically important changes in reports that were made or reviewed by individual radiologists. The significant variables from the multivariate analysis were included as fixed effects coefficients, and the random effects coefficients in the two models were the identity of the first reader and second reader, respectively.

Statistical analysis was done using IBM SPSS Statistics (V.22; IBM Corp, Somers, New York, USA) and Stata (V.12.1; StataCorp, College Station, Texas, USA). All p values are two-sided. A p value of <0.05 indicates statistical significance.

RESULTS

A total of 7838 abdominal CT examinations were conducted at the five hospitals in the time span from which we collected the reports. About 4102 of these were referred from the departments of surgery, from which 1970 (48%) were read by residents. We included pairs of reports from the 1071 examinations (26%), which were read by two consultant radiologists consecutively. Descriptive statistics regarding examinations, patients, hospitals and radiologists are shown in tables 1 and 2. The median delay between the preliminary and final reports was 19 h and 56 min. Details of report turnaround times are shown in online supplementary appendix 1.

Changes to reports

There were no changes made to 435 reports (41%). There were simple orthographical corrections or changes in layout for 237 reports (22%). In 399 reports (37%), the content had been changed, and these were submitted for clinical rating. A flow chart depicting this is shown in figure 1.

Clinical rating

On the five-point scale, the two raters were in agreement on 245 ratings (61%), and the weighted κ score for the inter-rater agreement was 0.60 (95% CI 0.53 to 0.66). Rater 2 gave lower ratings than rater 1 for 91 reports and gave higher ratings for 63 reports (p=0.049). On the dichotomised scale, there was agreement on 297 ratings (74%), and the κ score was 0.50 (95% CI 0.42 to 0.58).

The 154 discrepant ratings were resolved as follows: 10 reports with a mean rating of 1.5 were considered unequivocally 'not clinically important' and were not resolved further. A total of 144 reports

Table 1 Descriptive statistics of examinations, patients and readers, number (%) unless stated otherwise

	Double read*	All consultant read examinations†	All examinations‡
Examinations			
Conducted during ordinary working hours§	636 (59)	1172 (61)	1837 (45)
First reading during ordinary working hours§¶	519 (49)	–	–
Second reading during ordinary working hours§,**	839 (79)	–	–
Urgent referral††	722 (67)	1053 (55)	2642 (64)
Time from preliminary to final report‡‡, median	19 h 56 min	N/A	N/A
Patients			
Age, mean (SD)	60.6 (17.4) years	61.6 (17.1) years	60.5 (17.7) years
Female gender	526 (49)	953 (50)	2060 (50)
Inpatients	849 (79)	1343 (70)	3182 (78)
Specialist experience of readers, mean (SD)			
First readers	5.5 (6.9) years§§	N/A	N/A
Second readers	9.2 (9.4) years¶¶	10.5 (10.3) years***	9.3 (9.6) years†††

*Abdominal CT, referred from surgical department, double read by consultants (n=1071).
†Abdominal CT, referred from surgical department, read by consultants (n=1920).
‡Abdominal CT, referred from surgical department (n=4102).
§Monday to Friday 7:00–16:00.
¶n=1055.
**n=1061.
††Urgent: Requested within 24 h.
‡‡n=1055 (see online supplementary appendix 1 shows details of delay and turnaround times).
§§n=1042.
¶¶n=1060.
***n=1877.
†††n=3725.

Table 2 Descriptive statistics of hospitals and radiologists

Hospitals (n=5), median (range)	
No of beds, per hospital, surgical department	75 (17–144)
Annual output, surgical department*	5930 (1913–18 152)
No of annual CT exams, per hospital†	13 006 (5862–43 584)
Catchment population, per hospital	209 072 (77 836 – 471 661)
No of involved radiologists, per hospital	18 (6–31)
No of reports collected, per hospital	194 (43–414)
Proportion of double reading‡	0.33 (0.12–0.47)
Subspecialty of radiologists (n=87), number (%)	
Inexperienced consultant	26 (30)
General radiologist	23 (26)
Abdominal radiologist	23 (26)
Other subspecialty	15 (17)
Role of radiologists (n=90), number (%)	
First readings only	15 (17)
Second readings only	7 (8)
Both first and second readings	68 (76)
Gender of radiologists (n=90), number (%)	
Female	38 (42)

*Diagnosis-related group (DRG)—weighted (no of admissions×DRG-index).
†Norwegian Classification of Radiological Procedures (NCRP) 2012.
‡Abdominal CT, referred from surgical department, double read by consultants.

with discrepant ratings were submitted for a third rating. In the final classification, changes to 146 reports (14%, 95% CI 11.6% to 15.8%) from 1071 double-read examinations were clinically important. Changes to 108 reports (10%, 95% CI 8.3% to 12.0%) were intermediate, 35 (3%, 95% CI 2.3% to 4.5%) were major and 3 (0.3%, 95% CI 0.06% to 0.8%) were critical.

Clinical context

The clinical issues concerned in changes classified as clinically important are presented in table 3. Among the 146 clinically important changes, the severity of the radiological findings was increased in 118 (81%), decreased in 11 (8%), and unchanged in 17 (12%). All three critical changes implied an increase in severity. In one of the critical changes, the preliminary reported normal postoperative findings were changed to suspected anastomotic leakage.

Among changes classified as major, 30 (86%) implied an increase in severity, and in 5 (14%) the severity was unchanged. In one of the major changes, the preliminary reported possible (but unlikely) large bowel obstruction was changed to large bowel obstruction caused by a constricting tumour of the sigmoid colon with suspected metastases.

Among the changes classified as intermediate, 85 (79%) implied an increase in severity, 12 (11%) implied unchanged severity and 11 (10%) implied a decrease in severity. In one of the intermediate changes, the preliminary reported normal imaging findings were changed to a suspected cystadenoma in the head of the pancreas. More examples of report changes with description of clinical presentation and corresponding classification of clinical importance and change in severity are shown in online supplementary appendix 2.

The distribution of reasons for referral (n=1069) was acute presentations 349 (33%), non-acute presentations 211 (20%), follow-up 204 (19%) and investigations after surgery or invasive procedures 305 (29%). There was an association (p <0.01) between reasons for referral and clinically important change, with changes made less frequently to reports in a follow-up setting (OR: 0.4, p<0.001) than in the setting of acute presentations.

Factors associated with clinical importance

Associations between clinical importance of changes and characteristics of patients, examinations and readers are shown in table 4. The multivariate analysis showed that more clinically important changes were made to urgent referrals. Subspecialties of both first and second readers were associated with the rate of clinically important changes. Important changes were made less frequently when abdominal radiologists were first readers and more frequently when they were second readers.

Examination and first reading out of working hours and inpatient status were associated with higher rates of clinically important changes in the univariate model, but not in the multivariate model. The random effects logistic regression model did not show a significant clustering effect neither with regards to the identity of the first reader (p=0.3) nor with the second reader (p=0.1).

DISCUSSION

We found that prospective double reading of radiologist-selected examinations produced clinically important changes to 14% of radiology reports. Although our data stem from a different approach both to double reading and rating of discrepancies, the results are not significantly different from a previously reported 11.8% pooled total discrepancy rate for CT of the abdomen and pelvis, suggesting that some quality assurance of radiological interpretation may be justified.[18]

Changes to 10% of reports were rated intermediate, necessitating added controls or a change in investigations or prognosis. Although the results of these investigations are not known, they are not inconsequential neither with regards to the patients nor to resource consumption. Changes to 4% of reports were rated major or critical, implying changes in conservative or invasive treatment.

We rated discrepancies based on the potential clinical consequences of discrepancies, and used

Table 3 Clinical characteristics of report changes classified as clinically important

Category of pathology	Subcategory of pathology	Severity		
		Increased	Unchanged	Decreased
Cancer*(28)	Presence of tumour/suspicion of cancer (17)	15	2	
	Progression/extent/recurrence/metastases (9)	9		
	Altered suspected origin/location (2)		2	
Possible premalignancies†‡§¶ (21)	Pancreatic tumour/cyst (5)	5		
	Adrenal incidentaloma (4)	4		
	Liver lesion (4)	4		
	Gynaecology—Ovarian cyst (4)	4		
	Urology—Kidney cysts (2)	2		
	Other (2) (Spleen lesion, Lung nodulus)	2		
Infection**,††(17)	Abscess (6) (incl. 1 tubo-ovarian abscess)	3		3
	Appendicitis (6)	2	2	2
	Diverticulitis (2)		2	
	Cholecystitis/cholangitis with liver abscess (2)	1	1	
	Pneumonia (1)	1		
Vascular (16)	Pulmonary embolism (4)	4		
	Venous thrombosis (Hepatic 2, Mesenterial 1, Portal 1)	2	1	1
	Mesenterial claudication (1)	1		
	Ischaemia of small or large intestine (4)	4		
	Aneurysms (abdominal aorta, iliac and cystic artery) (3)	2		1
Pancreaticobiliary system‡,†† (12)	Biliary obstruction (4)	3		1
	Pancreatic duct obstruction/dilatation (2)	2		
	Contrast accumulation in gall bladder (1)			1
	Pancreatitis and sequela (5)	5		
Leak/perforation (11)	Anastomotic leakage (5)	3	1	1
	Perforated diverticulitis (4)	4		
	Perforation, cause unknown (1)		1	
	Perforated duodenal ulcer (1)	1		
Intestinal‡‡ obstruction* (9)	Small intestinal obstruction (6)	5	1	
	Large intestinal obstruction (2)	2		
	Subileus (1)	1		
Hernia§§ (9)	Hernia without obstruction/incarceration (9)	9		
Intestinal‡‡ inflammation (5)	Colitis (3)	2	1	
	Inflammation of jejunum (1)	1		
	Pouchitis (1)	1		
Other (18)	Enlarged lymph nodes (3), Mesenterial adenitis (2)	3	1	1
	Urology¶ (3): Hydronephrosis, Uroplania, Undescended testis	2	1	
	Intussusception without obstruction (3)	3		
	Skeletal abnormalities (3)	3		
	Oesophagus, contrast enhancement (1)	1		
	Pyloric ulcer (1)	1		
	Accessory spleen (1)	1		
	Gynaecology'§,**—Endometrioma (1)		1	
Total		118	17	11

*Large intestinal obstruction due to colon cancer (1) classified as obstruction, not cancer.
†Possible premalignancy: defined as lesions requiring further investigations or controls in order to evaluate risk of malignancy.
‡Pancreatic cysts/tumours (5) classified as premalignancy, not pancreaticobiliary system.
§Ovarian cysts (4) classified as premalignancy, not gynaecology.
¶Complex kidney cysts (2) classified as premalignancy, not urology.
**Tubo-ovarian abscess (1) classified as infection, not gynaecology.
††Cholecystitis/cholangitis with liver abscess (2) classified as infection, not pancreaticobiliary system.
§§Incarcerated hernia with intestinal ischaemia (1) classified as vascular: intestinal ischaemia, not hernia.
‡‡Intestinal=small or large intestine.

experienced gastrointestinal surgeons as raters. This is logical as surgeons have superior clinical knowledge, are the typical recipients of these reports and are accustomed to making clinical decisions partly founded on their content. Traditionally, radiologists have rated discrepancies of interpretation according to the magnitude of the error in question.[13] Such rating is subjective and may be perceived as punitive.[19] Previously reported inter-rater agreement is slight to fair with a κ of 0.17–0.2.[17 20 21] The clinical rating system in the present study was more reliable, achieving a moderate to substantial inter-rater agreement, with a κ of 0.5–0.6.[17] In a quality assurance perspective there might be mutual benefits from bringing clinicians into the feedback loop. It may increase awareness among clinicians of the limitations of

Table 4 Associations between clinically important report changes and characteristics of examinations, patients and readers

		Logistic regression analysis					
		Univariate			Multivariate (n=1055)		
Variable	n	OR	95% CI	p Value	OR	95% CI	p Value
Examination							
Urgency (urgent referral* vs not†)	1071	2.0	1.3 to 3.1	0.001	1.6	1.0 to 2.5	0.05
Examination time (out of working hours‡ vs during†)	1071	1.8	1.3 to 2.6	0.001			
Time of first reading (out of working hours‡ vs during†)	1055	1.6	1.1 to 2.3	0.01			
Time of second reading (out of working hours‡ vs during†)	1061	0.8	0.5 to 1.3	0.4			
Patient							
Age (increase of 10 years)	1071	1.1	1.0 to 1.2	0.2			
Gender (female vs male†)	1071	0.8	0.6 to 1.1	0.2			
Admission status (inpatients vs outpatients†)	1071	1.8	1.1 to 2.9	0.03			
First reader, gender (female vs male†)	1071	1.1	0.8 to 1.6	0.6			
First reader subspecialty	1064			<0.001			0.001
Inexperienced consultant†	408	1.0			1.0		
General radiologist	202	0.4	0.2 to 0.7	0.001	0.6	0.4 to 1.7	0.7
Abdominal radiologist	383	0.5	0.3 to 0.8	0.001	0.4	0.3 to 0.6	<0.001
Other subspecialty	71	1.0	0.5 to 1.9	0.9	1.1	0.6 to 2.2	0.7
Second reader, gender (female vs male†)	1071	1.0	0.7 to 1.4	0.9			
Second reader subspecialty	1062			<0.001			0.002
Inexperienced consultant†	222	1.0			1.0		
General radiologist	235	0.3	0.2 to 0.6	<0.001	0.3	0.2 to 0.7	0.002
Abdominal radiologist	535	0.8	0.6 to 1.3	0.4	1.1	0.7 to 1.7	0.7
Other subspecialty	70	0.2	0.1 to 0.6	0.01	0.2	0.1 to 0.8	0.02

*Urgent: Requested within 24 h.
†Reference in the logistic regression model.
‡Working hours: 7:00–16:00.

radiology, and among radiologists of the discrepancies that matter most to clinicians and patients.[19]

Our data result from routine quality assurance as it is practiced, and the results should be representative of everyday clinical practice in these departments. The first reader selected the cases for double reading, but we do not know their reasons or thresholds for doing so. One might expect that complex cases be selected more frequently, which might increase the rate of interpretation discrepancies. However, this is not necessarily the case. Autopsy studies have shown that in almost half of autopsies requested by clinicians they were ‘fairly certain’ of the main diagnosis, and that the degree of clinical confidence was an inadequate predictor of diagnostic errors.[22–25]

Less-experienced consultants submitted more cases for double reading, and more experienced radiologists tended to conduct the second reading, indicating that the task was not randomly assigned. The higher rate of clinically important changes made by abdominal radiologists as second readers may therefore partly be due to intentional routing of complex cases to these readers as well as their competence in detection, interpretation and reporting. Similarly the lower rate of clinically important changes made to abdominal radiologists as first readers may result from higher performance or a tendency by the second readers to put more trust in their judgement and less scrutiny in their work.

The non-random selection of cases and readers renders our data unsuitable for benchmarking of performance, and the outcomes may not pertain to all abdominal CTs performed. However, retrospective peer review systems, which are frequently used for this purpose, are also vulnerable to selection bias due to radiologists’ intentional avoidance of cases taking more time to review and conscious selection of less-time-intensive cases.[26] A similar reluctance has been reported in physicians failing to participate in adverse events reporting due to risk of liability exposure or professional embarrassment, burdensome reporting methods, time required for reporting, perceptions of the clinical import of adverse events and lack of sense of ownership in the process.[27]

The median delay between the preliminary and final reports was approximately 20 h. Meantime it is possible that the discrepancy be discovered based on clinical factors, or that the opportunity to intervene be missed. However, for most findings the information will still be relevant, and patient treatment may still be corrected. This opportunity to prevent patient harm directly may facilitate a more wholehearted participation by radiologists, and may also reduce concerns over medico-legal issues.

Clinically important changes were made more often to the reports from urgent investigations. This may be attributed to a higher frequency of new findings in these examinations or to a less favourable working environment of the on-call radiologist. Regardless it is worth considering urgent examinations especially for quality assurance.

This study was limited to the preliminary and final radiology reports, and did not consider any supplementary communication between radiologists and clinicians. Since there is a delay between the first and second reading, second readers may have gained information on patient development through clinical conferences or subsequent investigations, and some report changes may not result from the second reading only.

Another limitation of our study is that the actual impact of the report changes is unknown. It is questionable whether patient records can be relied on to establish this retrospectively. Records may be incomplete regarding decisions and their justifications, and courses of action may change before they are recorded. In the absence of a gold standard we cannot confirm that the second reading was the correct one. There are studies in which discrepancies between preliminary interpretations of residents and final interpretations of staff radiologists have been compared with those of consensus reference panels. The panels confirmed the second reading in 64%–85%, and were more likely to confirm a second reading pointing out false-positive than false-negative and false indeterminate preliminary reports.[28–30] Accordingly, in some cases report changes may have resulted in increased costs or even harm without benefit to the patient. This underlines the importance of establishing a feedback system involving the first and second readers and of course the clinicians.

We conclude that a 14% rate of clinically important changes made during double reading suggest that some quality assurance of radiological interpretation is justified. Using expert second readers, and targeting urgent cases and radiologists reading outside their specialty may increase the yield of discrepant cases. Establishing additional objective selection criteria would require further studies.

Author affiliations
[1]Department of Diagnostic Imaging, Akershus University Hospital, Lørenskog, Norway
[2]Institute of Clinical Medicine, University of Oslo, Campus Ahus, Lørenskog, Norway
[3]Department of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway
[4]Department of Radiology, Vestre Viken, Drammen Hospital, Drammen, Norway
[5]Department of Radiology, Vestre Viken, Bærum Hospital, Sandvika, Norway
[6]Department of Gastrointestinal Surgery, Akershus University Hospital, Lørenskog, Norway
[7]Health Services Research Unit, Akershus University Hospital, Lørenskog, Norway
[8]Institute of Clinical Medicine, University of Oslo, Oslo, Norway

Acknowledgements This study was made possible by research grants from the Norwegian Medical Association's Fund for Patient Safety and Quality Improvement and from the Norwegian Society of Radiology. We would like to thank Øystein Andreas Ødegaard for his contribution as third clinical rater.

Contributors PML: Substantial contributions towards concept, design, data retrieval, data processing/analysis, statistical analysis, interpretation, drafting manuscript, revising and approval. JGA, MVS and ALT: Substantial contributions towards design, data retrieval, interpretation, revising manuscript and approval. RA and TH: Substantial contributions towards design (rating scale), data processing, data analysis, revising manuscript and approval. PH and GS: Substantial contributions towards concept, design, interpretation, revising manuscript and approval. FAD: Substantial contributions towards design, data processing/analysis, statistical analysis, interpretation, revising manuscript and approval. PG: Substantial contributions towards concept, design, data processing/analysis, statistical analysis, interpretation, drafting manuscript, revising and approval. Øystein Andreas Ødegaard: contribution as the third clinical rater (see Acknowledgement).

Funding The Norwegian Society of Radiology and The Norwegian Medical Association (Fund for Patient Safety and Quality Improvement) (KSF1: 12/1318).

Competing interests PML reports research grants from the Norwegian Medical Association and the Norwegian Society of Radiology for the conduct of this study.

Ethics approval South-East Regional Ethics Committee, Norway (ref: 2012/1986) and the Data Protecion Officer at each hospital.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Provided relevant permissions are obtained; an anonymised version of the dataset will be made available on request should anyone wish to inspect our work or conduct further studies.

REFERENCES

1 Kohn LT, Corrigan JM, Donaldson MS. *To err is human: building a safer health system*. Washington DC: Institute of Medicine (US), 2000.
2 Donaldson L, Appleby L, Boyce J. *An organization with a memory: report of an expert group on learing from adverse events in the NHS*. London, UK: Department of Health, 2000.
3 Heriot GS, McKelvie P, Pitman AG. Diagnostic errors in patients dying in hospital: radiology's contribution. *J Med Imaging Radiat Oncol* 2009;53:188–93.
4 Balogh E, Miller BT, Ball J. *Improving diagnosis in health care*. Washington DC: Institute of Medicine (U.S.). Committee on Diagnostic Error in Health Care, 2015.
5 Goddard P, Leslie A, Jones A, *et al*. Error in radiology. *Br J Radiol* 2001;74:949–51.
6 Federle MP, Gur D. Double reading of certain examinations such as barium enemas and mammograms can increase sensitivity at the expense of specificity. *AJR Am J Roentgenol* 1995;164:1291–2.

7 Agostini C, Durieux M, Milot L, *et al*. Value of double reading of whole body CT in polytrauma patients. *J Radiol* 2008;89(Pt 1):325–30.
8 Babiarz LS, Yousem DM. Quality control in neuroradiology: discrepancies in image interpretation among academic neuroradiologists. *AJNR Am J Neuroradiol* 2012;33:37–42.
9 Hurlen P, Østbye T, Borthne A, *et al*. Introducing PACS to the late majority. a longitudinal study. *J Digit Imaging* 2010;23:87–94.
10 Jakanani GC, Botchu R, Gupta S, *et al*. Out of hours multidetector computed tomography pulmonary angiography: are specialist resident reports reliable? *Acad Radiol* 2012;19:191–5.
11 Ballard-Barbash R, Klabunde C, Paci E, *et al*. Breast cancer screening in 21 countries: delivery of services, notification of results and outcomes ascertainment. *Eur J Cancer Prev* 1999;8:417–26.
12 Mahgerefteh S, Kruskal JB, Yam CS, *et al*. Peer review in diagnostic radiology: current state and a vision for the future. *Radiographics* 2009;29:1221–31.
13 Jackson VP, Cushing T, Abujudeh HH, *et al*. RADPEER scoring white paper. *J Am Coll Radiol* 2009;6:21–5.
14 The Royal College of Radiologists. *Quality assurance in radiology reporting: peer feedback*. London, UK: The Royal College of Radiologists, 2014:1–18.
15 The Royal College of Radiologists. *Standards for learning from discrepancies meetings*. London, UK: The Royal College of Radiologists, 2014:1–16.
16 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 Radiol* 2015;56:78–86.
17 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
18 Wu MZ, McInnes MD, Macdonald DB, *et al*. CT in adults: systematic review and meta-analysis of interpretation discrepancy rates. *Radiology* 2014;270:717–35.
19 Larson DB, Nance JJ. Rethinking peer review: what aviation can teach radiology about performance improvement. *Radiology* 2011;259:626–32.
20 Mucci B, Murray H, Downie A, *et al*. Interrater variation in scoring radiological discrepancies. *Br J Radiol* 2013;86:20130245.
21 Bender LC, Linnau KF, Meier EN, *et al*. Interrater agreement in the evaluation of discrepant imaging findings with the Radpeer system. *AJR Am J Roentgenol* 2012;199:1320–7.
22 Cameron HM, McGoogan E. A prospective study of 1152 hospital autopsies: I. Inaccuracies in death certification. *J Pathol* 1981;133:273–83.
23 Landefeld CS, Chren MM, Myers A, *et al*. Diagnostic yield of the autopsy in a university hospital and a community hospital. *N Engl J Med* 1988;318:1249–54.
24 Britton M. Diagnostic errors discovered at autopsy. *Acta Med Scand* 1974;196:203–10.
25 Shojania KG, Burton EC, McDonald KM, *et al*. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *JAMA* 2003;289:2849–56.
26 Eisenberg RL, Cunningham ML, Siewert B, *et al*. Survey of faculty perceptions regarding a peer review system. *J Am Coll Radiol* 2014;11:397–401.
27 Farley DO, Haviland A, Champagne S, *et al*. Adverse-event-reporting practices by US hospitals: results of a national survey. *Qual Saf Health Care* 2008;17:416–23.
28 Ojutiku O, Haramati LB, Rakoff S, *et al*. Radiology residents' on-call interpretation of chest Radiographs for pneumonia. *Acad Radiol* 2005;12:658–64.
29 Rufener SL, Patel S, Kazerooni EA, *et al*. Comparison of on-call radiology resident and faculty interpretation of 4- and 16-row multidetector CT pulmonary angiography with indirect CT venography. *Acad Radiol* 2008;15:71–6.
30 Chung JH, Strigel RM, Chew AR, *et al*. Overnight resident interpretation of torso CT at a level 1 trauma center. *Acad Radiol* 2009;16:1155–60.