Inter-observer agreement for abdominal CT in unselected patients with acute abdominal pain

van Randen, A.; Laméris, W.; Nio, C.Y.; Spijkerboer, A.M.; Meier, M.A.; Tutein Nolthenius, C.; Smithuis, F.; Bossuyt, P.M.; Boermeester, M.A.; Stoker, J.

DOI
10.1007/s00330-009-1294-9

Publication date
2009

Document Version
Final published version

Published in
European Radiology

Eur Radiol (2009) 19: 1394–1407
DOI 10.1007/s00330-009-1294-9

Inter-observer agreement for abdominal CT in unselected patients with acute abdominal pain

Adrienne van Randen
Wytze Laméris
C. Yung Nio
Anje M. Spijkerboer
Mark A. Meier
Charlotte Tutein Nolthenius
Frank Smithuis
Patrick M. Bossuyt
Marja A. Boermeester
Jaap Stoker

Received: 17 July 2008
Accepted: 8 November 2008
Published online: 21 February 2009

A. van Randen · W. Laméris · C. Y. Nio · A. M. Spijkerboer · M. A. Meier · C. Tutein Nolthenius · F. Smithuis · J. Stoker
Department of Radiology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

A. van Randen · W. Laméris · M. A. Boermeester
Department of Surgery, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

P. M. Bossuyt
Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

A. van Randen (✉)
Academic Medical Center,
Meibergdreef 9,
1105 AZ Amsterdam, The Netherlands
e-mail: a.vanranden@amc.uva.nl
Tel.: +31-20-5662630
Fax: +31-20-5669119

Abstract The level of inter-observer agreement of abdominal computed tomography (CT) in unselected patients presenting with acute abdominal pain at the Emergency Department (ED) was evaluated. Two hundred consecutive patients with acute abdominal pain were prospectively included. Multi-slice CT was performed in all patients with intravenous contrast medium only. Three radiologists independently read all CT examinations. They recorded specific radiological features and a final diagnosis on a case record form. We calculated the proportion of agreement and kappa values, for overall, urgent and frequently occurring diagnoses. The mean age of the evaluated patients was 46 years (range 19–94), of which 54% were women. Overall agreement on diagnoses was good, with a median kappa of 0.66. Kappa values for specific urgent diagnoses were excellent, with median kappa values of 0.84, 0.90 and 0.81, for appendicitis, diverticulitis and bowel obstruction, respectively. Abdominal CT has good inter-observer agreement in unselected patients with acute abdominal pain at the ED, with excellent agreement for specific urgent diagnoses as diverticulitis and appendicitis.

Keywords Interobserver · CT-abdomen · Acute abdominal pain

Introduction

Acute abdominal pain is a common emergency that can be caused by a wide variety of conditions, ranging from self-limiting to life-threatening disease. For patients presenting to the Emergency Department (ED) with acute abdominal pain, fast and accurate work-up is needed to plan treatment. Imaging will be used to differentiate between urgent conditions (requiring immediate treatment) and non-urgent conditions, and for determining the diagnosis and extent of disease.

The diagnostic work-up for acute abdominal pain has changed over the last 10 years, with a fourfold increase in the use of computed tomography (CT) [1, 2]. High accuracy of CT has been reported for appendicitis and for acute diverticulitis (98%) [3, 4]. A study evaluating the diagnostic value of abdominal CT in patients with acute abdominal pain in general also showed good results, with an accuracy of 78% for CT, including clinical evaluation [5]. Although the accuracy reported in the literature is good, this does not automatically imply that reproducibility is good as well and that accuracy results reported in the literature can be generalised to the local clinical situation.

Only a few studies evaluated the reproducibility of CT results. These studies focused on selected patients, with a specific suspected diagnosis, such as appendicitis or diverticulitis [6–9]. There were considerable differences in the reported inter-observer agreement in these studies, ranging from fair to excellent [6–9]. Fair inter-observer agreement was also reported for so-called difficult cases [6]. The latter study included CT examinations that were

equivocal for the diagnosis of appendicitis after the first analysis.

These previous inter-observer studies all evaluated patients suspected of one specific disease and the evaluation of the test results focused merely on the presence or absence of that disease. An evaluation of inter-observer agreement in unselected patients with acute abdominal pain presenting at the ED is probably more relevant as it is closest to clinical practice, where consecutive patients usually have different diagnoses, and most likely clinical diagnosis may not be evident after clinical evaluation in a substantial number of cases.

The purpose of this study was to document the level of inter-observer agreement in abdominal CT in unselected patients with acute abdominal pain presenting at the ED. We evaluated agreement of all diagnoses, as well as agreement on urgent versus non-urgent diagnoses, on general radiological features, and on frequent diagnoses in this patient population.

Method and materials

Patients presenting at the ED with acute abdominal pain for more than 2 h and less than 5 days were eligible for this study. Patients who were discharged by the treating physician at the ED without any diagnostic imaging (including plain radiography and ultrasonography), patients under 18 years, pregnant women, patients with a blunt or penetrating trauma, and patients in haemorrhagic shock caused by a gastrointestinal bleeding or acute abdominal aneurysm, were excluded Furthermore, only patients with abdominal pain were eligible for this study; if a patient had just flank pain, that patient was not invited. The included patients were part of a larger trial to document the diagnostic accuracy of imaging modalities in the work-up of patients with acute abdominal pain [10]. In this trial 1,101 consecutive patients underwent plain abdominal and chest radiography, abdominal US and abdominal CT. The first 200 patients of the initiating hospital entered this retrospective inter-observer study.

Eligible patients were informed about the study and asked for consent. Consenting patients underwent CT within a few hours after ED presentation. A multidetector-row four-slice helical CT (SOMATOM Sensation 4; Siemens Medical Systems, Forchheim, Germany) was used in all patients. The CT protocol was as follows: effective mAs level of 165, 120 kV, (4×) 2.5-mm collimation, (4×) 3-mm slice width and 0.5-s rotation time. A total of 125 ml contrast medium (Visipaque 320, General Electric Healthcare, Chicago, Ill.) was injected intravenously at 3 ml/s and the CT was performed after a 60-s delay; no oral or rectal contrast agents were used. The effective dose of this CT protocol was 11 mSv, with a DLP of 640 mGy·cm.

Only patients with known renal failure received an unenhanced CT.

The CT examinations were reviewed 3 months or more after the initial presentation at the ED, to diminish recall bias of the observers, as some of them could have been involved in the initial diagnostic work-up of these patients. All CT examinations were interpreted using a picture archiving and communication system (PACS; Agfa-Gevaert, Brussels, Belgium), on which observers evaluated the axial-CT images; however, they had access to three-dimensional reformats. The CT examinations were independently read by three radiologists, blinded for the results of the co-observers. Two observers both had 12 years of experience in abdominal radiology, in which they had evaluated approximately 5,200 abdominal CT studies for acute abdomen and 13,000 abdominal CT examinations in general. The third observer had 2 years of experience as a general radiologist (fellow interventional radiology), and had evaluated approximately 175 abdominal CT studies indicated for acute abdomen, and 1,000 abdominal CT examinations in general. The observers were blinded for other imaging examinations performed in the diagnostic work-up on the day of presentation of these patients, imaging performed during follow-up and other findings during follow-up.

Observers had access to a summary of the patients' clinical history, physical examination (both performed by an attending surgical resident) and to the laboratory findings of the day of presentation, as in normal clinical practice observers also would have access to clinical information of the patient [11]. An example of summary patient information is provided in Appendix I.

Image and CT characteristics

For a standardized evaluation of the CT examinations, image characteristics were assessed and recorded on a digital case record form. The following general image findings and specific radiological features were assessed: image quality, fat infiltration, free fluid, fluid collections, free intra-peritoneal air, and whether fistulas could be visualized. Image characteristics were also assessed for abnormalities per organ: gallbladder, bile duct, liver, pancreas, appendix, gastrointestinal tract (without appendiceal abnormalities), lymph nodes, vascular system, kidneys, and if appropriate, female genitalia.

In case of abnormalities further specification on the observed abnormality was warranted. All observers also recorded their final CT diagnosis, and, if applicable, two differential diagnoses. These diagnoses were selected from a list of diagnoses provided with the online case record form. All possible diagnoses in the list of diagnoses on the online case record form had been classified a priori as urgent or non-urgent. Diagnoses were classified as urgent when immediate treatment, within 24 h, was needed, whereas in patients with a non-urgent diagnosis a general consensus exists that treatment, if any, can safely be delayed beyond 24 h.

Final diagnosis

An independent expert panel, consisting of two experienced gastrointestinal surgeons and an experienced abdominal radiologist, assigned a final diagnosis. The members of this panel individually evaluated all available data for each patient. Final consensus on the final diagnosis was reached in a consensus meeting. Information was provided to the expert panel in a standardized way and consisted of clinical findings, laboratory findings, image findings, surgery (if any), histopathology (if any) and the results from 6 months of outpatient follow-up. Panel members selected the final diagnosis from the same list of diagnoses as provided to the initial observers. None of the panel members had been involved in the work-up of the included patients or in reading the CT images in this study.

Analysis

In the analysis, our focus was on overall inter-observer agreement, on agreement on urgent diagnoses, and on frequently occurring diagnoses. Inter-observer agreement was also evaluated for specific radiological features, such as fat infiltration, free fluid, fluid collections, and free intraperitoneal air.

Frequent diagnosis within the population under study was defined as diagnosis with a prevalence >5%. Frequencies of diagnoses or specific features were calculated for each observer. The number of cases in which an observer recorded a specific diagnosis or feature (e.g. fat infiltration) was recorded per observer. It is thought that if different observers record a specific diagnosis or feature in a similar number of the patients, agreement will be good as well. This evaluation of frequencies is used as a rough measurement to indicate inter-observer agreement.

Agreement was calculated and expressed as percentage observed agreement (i.e. the number of CT examinations at which both observers scored a feature as present or absent divided by the total of 200 CT examinations evaluated by both observers) and with kappa statistics (i.e. the observed agreement adjusted for chance). If prevalence is very high or very low, chance on agreement increases, thereby lowering kappa values. Kappa values were calculated for each observer. Median kappa values were calculated for all three observers. Kappa values can be calculated for a 2×2 table as well as for more extensive tables [12].

Kappa (κ) values can be classified according to the level of agreement as $\kappa<0.20$ poor agreement, $\kappa=0.21$–0.40 fair agreement, $\kappa=0.41$–0.60 moderate agreement, $\kappa=0.61$–0.80 good agreement, $\kappa=0.81$–1.00 excellent agreement, according to Landis and Koch [13].

For all analyses the statistical software package SPSS 12.0.2. was used (SPSS, Chicago, Ill.)

Results

The mean age of the 200 included patients was 46 years (range 19–94) and 54% ($n=107$) of the patients were female. In 17 patients it was not possible to obtain a final diagnosis because of incomplete patient data from initial clinical history and physical examination at the ED (discharge diagnoses were non-specific abdominal pain (NSAP) in six, pneumonia in two, miscellaneous in nine).

The most frequent final diagnoses, assigned by the expert panel, were acute appendicitis in 41 (22%) patients, NSAP in 32 (17%), and acute diverticulitis in 20 (11%) patients (Table 1). Of the 200 patients evaluated in the inter-observer analysis, 193 patients had received intravenous contrast agent, whilst seven (3.5%) had unenhanced CT (five with renal failure; two with inappropriate position of intravenous cannula).

Table 1 Final diagnoses after 6 months follow-up, reference standard

Diagnoses	Frequency (*n*)
Appendicitis	41
NSAP[a]	32
Diverticulitis	20
Bowel obstruction	14
inflammatory bowel obstruction	1
bowel obstruction adhesion-related	9
bowel obstruction non specified	3
Hepatic pancreatic biliary disorder[a]	14
hepatitis	1
cholecystolithiasis	10
chronic pancreatitis	1
choledocholithiasis	2
Gastrointestinal disorder; non-urgent[a]	12
gastritis	1
gastroenteritis	4
epiploic appendagitis	1
bowel inflammation; non-specified	6
Cholecystitis	8
Pancreatitis	7
Urological disorder; non-urgent[a]	7
urinary tract infection	6
renal stones without obstruction	1
Gynecological disorder; urgent	6
bleeding - rupture ovarian cyst	2
adnexal torsion	1
pelvic inflammatory disease	3
Urological disorder, urgent	5
urinary tract stones with obstruction	4
pyelonephritis	1

Table 1 (continued)

Diagnoses	Frequency (*n*)
Abscess	3
intra-abdominal abscess	1
tubo-ovarian abscess	1
abdominal wall abscess	1
Bowel ischemia	3
Gynecological disorder; non-urgent[a]	3
endometriosis	2
uterus myoma	1
Pneumonia	2
IBD	1
ulcerative colitis	1
Malignancy[a]	1
pancreas carcinoma	1
Perforated viscus	1
Peritonitis	1
Miscellaneous[a]	3
other[b]	2
mesenteric vein thrombosis	1
Total	183

[a]Diseases classified as non-urgent, meaning no treatment was needed within 24 h
[b]Other disease consisted of renal infarction and non-specified post-operative abdominal pain in the other patient

Overall agreement

Overall inter-observer agreement on diagnoses was good, with a kappa value of 0.66 (95% CI: 0.60–0.75) for observers 1 and 2, a kappa value of 0.63 (95% CI: 0.58–0.69) for observers 1 and 3 and a kappa value of 0.67 (CI: 0.63–0.75) for observers 2 and 3 (median kappa value of 0.66). The observed proportion of agreement was 0.71, 0.67 and 0.71 for observer 1 and 2, 1 and 3 and, 2 and 3, respectively (Table 2).

An overall cross-classification of all diagnoses is provided for all three observer couples in Appendix II. For urgent versus non-urgent diagnoses agreement was moderate, with a median kappa for all three observers of 0.59 (see also Table 2). The observed agreement for urgent diagnoses was excellent, with a median agreement of 0.83 (Table 2). Furthermore, observers assigned an urgent diagnosis approximately to the same number of patients (Table 3).

Radiological features

Inter-observer agreement for specific radiological features, such as fat infiltration, free fluid, free intra-peritoneal air and fluid collections, is reported in Table 4. Detection of fat infiltration had a good level of agreement, with a median kappa value of 0.70. Agreement on free fluid was moderate, with a median kappa value of 0.58. The frequency in which observers recorded presence of free fluid, differed considerably, ranging from 26% to 46% (Tables 3, 4). Fluid collection and free intra-peritoneal air both had an extremely high observed agreement. Because the prevalence of fluid collections and free intra-peritoneal air within this study population were very low, the corresponding kappa values were low as well.

Agreement on specific diagnoses

Kappa values and observed agreement are listed in Table 5 for diagnoses with prevalence higher than 5% within this study population. Median kappa values for specific urgent diagnoses, such as appendicitis, diverticulitis and bowel obstruction were 0.84, 0.90, and 0.81, respectively, which implies excellent agreement (Figs. 1, 2). Median kappa values for non-urgent diagnoses were moderate to fair (Table 5). This difference between urgent and non-urgent diagnoses could not be derived from the frequencies of these specific diagnoses assigned by radiologists only. Frequencies between urgent and non-urgent diagnoses did not differ at large. The table in Appendix II shows cross tables of

Table 2 Observer agreement of all diagnoses and of urgent diagnoses

	Observer 1 and observer 2		Observer 1 and observer 3		Observer 2 and observer 3		Median
	Observed agreement[a]	Kappa (95% CI)	Observed agreement[a]	Kappa (95% CI)	Observed agreement[a]	Kappa (95% CI)	Kappa
Overall diagnoses	0.71	0.66 (0.59–0.73)	0.67	0.63 (0.56–0.70)	0.71	0.67 (0.60–0.74)	0.66
Urgent diagnosis	0.86	0.69 (0.58–0.79)	0.81	0.59 (0.47–0.71)	0.83	0.59 (0.50–0.73)	0.59

[a]Observed agreement was calculated as the number of patients identified with the diagnosis by both observers or if both scored an urgent diagnosis as present or absent divided by a total of 200 diagnoses

Table 3 Frequencies of urgent diagnoses, CT examinations with a high level of confidence, frequent occurring diagnoses and radiological features

General	Observer 1		Observer 2		Observer 3	
	%[a]	*n*[b]	%[a]	*n*[b]	%[a]	*n*[b]
Urgent diagnoses	63%	125	65%	130	66%	131
Frequent diagnosis at CT						
Appendicitis	28%	55	25%	49	28%	56
Diverticulitis	11%	22	12%	24	10%	20
Bowel obstruction	6%	11	7%	14	8%	15
Hepatic pancreatic biliary tract disorder	5%	10	4%	8	6%	12
Gastrointestinal tract disorder	6%	11	6%	12	7%	13
NSAP	19%	37	17%	33	15%	30
Feature at CT						
Fat infiltration	58%	116	54%	108	49%	98
Mild	24%	47	19%	37	18%	36
Moderate	23%	46	25%	49	17%	34
Severe	11%	22	9%	17	14%	27
Free intra-peritoneal air	5%	10	6%	11	4%	7
Free fluid	35%	70	46%	109	26%	52
Mild	22%	43	37%	73	16%	32
Moderate	9%	18	11%	22	8%	15
Severe	2%	3	1%	2	2%	4
Fluid collection	5%	10	2%	4	4%	7

[a]Percentage is *n*/200
[b]Number of times seen or diagnosed at CT by the observer in 200 CT scans

diagnoses assigned per observer and, thereby, the difference in agreement between urgent and non-urgent diagnoses.

Discussion

In this study, in unselected patients with acute abdominal pain presenting to the ED, abdominal CT was found to have good inter-observer agreement, with excellent inter-observer agreement for urgent diagnoses, such as appendicitis, diverticulitis and bowel obstruction. For non-urgent diagnoses, such as hepatic pancreatic biliary disorders, gastrointestinal tract disorders and NSAP, CT had good but not excellent agreement of abdominal CT in patients with acute abdominal pain. One should be aware that most of these non-urgent diagnoses, such as gastro-enteritis, can not be readily made by CT. Therefore, efforts must be made to adequately select patients with acute abdominal pain at the ED that will benefit from CT. Inter-observer agreement was generally moderate for individual radiological features, but that did not negatively affect agreement on urgent diagnoses. It is most important that an urgent diagnosis is recognised by all observers. Patients with an urgent disease need immediate treatment, whereas patients with a non-urgent cause of acute abdominal pain, treatment can be safely delayed beyond 24 h. In these patients, more time is available to establish the correct diagnosis.

This study has some potential limitations that have to be taken into account. First, we did not evaluate intra-observer

Table 4 Observer agreement on specific radiological features

Agreement on specific radiological features	Observer 1 and observer 2		Observer 1 and observer 3		Observer 2 and observer 3		Median kappa
	Observed agreement[a]	Kappa (95% CI)	Observed agreement[a]	Kappa (95% CI)	Observed agreement[a]	Kappa (95% CI)	
Fat infiltration	0.85	0.70 (0.60–0.80	0.82	0.64 (0.53–0.75)	0.87	0.74 (0.65–0.83)	0.70
Free intra-peritoneal air	0.97	0.65 (0.39–0.90)	0.96	0.45 (0.10–0.80)	0.98	0.77 (0.54–0.99)	0.65
Free fluid	0.79	0.58 (0.47–0.69)	0.84	0.63 (0.51–0.74)	0.68	0.38 (0.25–0.50)	0.58
Fluid collection	0.96	0.41 (0.01–0.81)	0.94	0.20 (−0.22–0.62)	0.97	0.35 (−0.13–0.82)	0.35

[a]Observed agreement is calculated as follows: both observers scored a feature as present or absent divided by the total of 200 diagnoses

Table 5 Observed agreement and agreement according to kappa statistic for frequent occurring disease

Diagnosis at CT	Observer 1 and observer 2		Observer 1 and observer 3		Observer 2 and observer 3	
Urgent diagnoses	Observed agreement[a]	Kappa (CI)	Observed agreement[a]	Kappa (CI)	Observed agreement[a]	Kappa (CI)
Appendicitis	0.73 (44/60)	0.79 (0.70–0.89)	0.77 (49/64)	0.84 (0.75–0.92)	0.88 (49/62)	0.91 (0.84–0.98)
Diverticulitis	0.80 (24/30)	0.85 (0.74–0.97)	0.91 (20/22)	0.95 (0.87–1.00)	0.83 (20/24)	0.90 (0.80–0.99)
Bowel obstruction	0.71 (12/17)	0.87 (0.73–1.00)	0.53 (9/17)	0.67 (0.44–0.90)	0.71 (12/17)	0.81 (0.65–0.98)
Non-urgent diagnoses						
Hepatic pancreatic biliary tract disorder	0.29 (4/14)	0.42 (0.07–0.78)	0.29 (5/17)	0.42 (0.11–0.74)	0.54 (7/13)	0.42 (0.11–0.74)
Gastrointestinal tract disorder	0.28 (5/18)	0.54 (0.22–0.85)	0.14 (3/21)	0.23 (−0.13–0.59)	0.19 (13/21)	0.23 (−0.13–0.59)
NSAP	0.40 (18/45)	0.52 (0.35–0.68)	0.40 (19/49)	0.48 (0.31–0.66)	0.38 (18/47)	0.49 (0.31–0.97

[a]Observed agreement was calculated as the number of patients identified with the diagnosis by both observers divided by the total number of patients identified with the diagnosis by both and/ or one of the observers

agreement. As inter-observer agreement was good overall, information on intra-observer agreement may not be crucial. Another potential limitation of this study was the spectrum of experience of observers. All observers were radiologists, no radiological resident read the CT images for study purposes. Radiological residents are usually supervised by a radiologist in the evaluation of abdominal CT. Thirdly, the CT protocol within this study included intravenous contrast medium only. Oral or rectal contrast medium is not a prerequisite, although helpful in some conditions. In this study, CT examinations were read after 3 months, which may have caused some bias. Although, work level and time of day differed between initial review and cold review, methods of review were identical for all there observers. Furthermore, kappa values did not differ significantly between observer 3 and the observer at the ED (data not shown). Finally, because not all patients had a

Fig. 1 Case of agreement. A 52-year-old male with abdominal pain for 2 days in the right lower quadrant. He had complaints of nausea and vomiting and a temperature of 38°C. At physical examination he had right lower quadrant tenderness with guarding. The C-reactive protein was 206. All observers correctly diagnosed this patient with acute appendicitis (*arrow*). *C* cecum

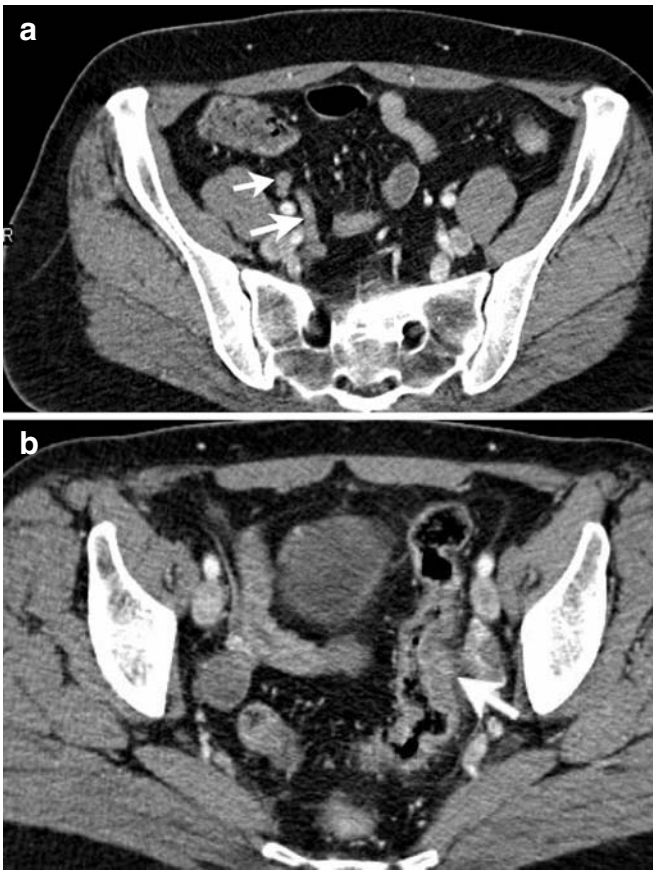

Fig. 2 Case of disagreement. A 57-year-old female, with right lower quadrant pain for 3 days and a temperature of 37.6°C. At physical examination both lower abdominal quadrants were tender, with rebound tenderness, but without guarding. C-reactive protein was 112 mg/l and WBC 8×10^9 g/l. **a** Observer 1 diagnosed this patient with appendicitis (*arrows* point at the appendix). **b** Observer 2 diagnosed her with diverticulitis (*arrow* points to a diverticula with adjacent inflammatory changes). Observer 3 diagnosed this patient with inflammation of the sigmoid, but without diverticulitis. The final diagnosis of the expert panel in this patient was acute diverticulitis. The patient was treated conservatively with rest and a liquid diet and recovered uneventfully

surgical and histopathological proven final diagnosis, an expert panel was used to assign a final diagnosis. For this reason, a follow-up period of 6 months was chosen to collect additional data.

Our results closely reflect daily clinical practice, as we invited all consecutive patients presenting with acute abdominal pain and made no a priori selection. In the literature [6–8] kappa values have been reported for selected patients with a suspicion of one specific condition, e.g. patients suspected of appendicitis or diverticulitis, or in studies in which abdominal CT were reviewed to identify the appendix [14].

Another study that has evaluated inter-observer variability of abdominal CT in general found a good inter-observer agreement for presence or absence of abdominal pathology, as measured on a five-point scale [15]. However, in that study abdominal pathology was not specified, and no diagnosis was assigned by the observers.

We chose to report level of agreement in kappa values, because they are widely used in the literature and well known to clinicians and radiologists. Cohen's kappa statistics express agreement adjusted for chance. We also reported observed agreement (percentage agreement) alongside kappa values. Kappa values are influenced by disease prevalence. If the disease prevalence is high in the study population, expected agreement will be high as well, and this will lower the corresponding kappa value. On the other end of the spectrum the same holds true: if disease prevalence is very low in the population under study, expected agreement will be high, thereby lowering the corresponding kappa value. Therefore, it is assumed that kappa values of urgent diagnoses in our study are lowered because of high prevalence of urgent diagnoses instead of actual moderate agreement.

Agreement between radiologists on the CT diagnoses was good, but excellent inter-observer agreement could have been expected, because of the excellent accuracy reported in literature, which presupposes excellent agreement. In the present study, accuracy was not a primary study aim, and the 200 abdominal CTs that had been assessed by three observers were nevertheless not enough to evaluate accuracy.

Accuracy studies can be prone to observer bias, when only highly experienced observers are used to evaluate the test. In this study no difference in level of agreement was found between all three observer couples, which suggests that agreement does not depend highly on additional years of experience.

CT images were evaluated with information on clinical history, physical and laboratory examination provided to the observers. Reading with clinical information can inflate test accuracy due to clinical review bias [16]. Test reading is influenced by clinical information in the perception of abnormalities and in the interpretation of abnormalities. Our results may have been influenced by clinical review bias, as CT scans were evaluated with knowledge of clinical information, but this situation reflects normal practice.

In conclusion, we can say that overall inter-observer agreement of radiologists for CT is good in patients with acute abdominal pain presenting at the emergency department and, most importantly, excellent agreement was found for urgent diagnoses. Therefore, if CT images suggest an urgent diagnosis in a patient with acute abdominal pain, it can safely be assumed that different radiologists would assign the same diagnosis, but opinions are more likely to differ for non-urgent diagnoses.

Acknowledgments The Dutch Organization for Health Research and Development, Health Care Efficiency Research programme funded this study (ZonMw, grant number 945-04-308).

Appendix I

Example of patient information provided to the reviewers of CT examinations

Name	Name, female		
Patient id number	00000000		
Date of birth	00-00-0000		
Previous clinical history	None		
Clinical history			
Duration of pain	4 h		
Localization of pain	RLQ, umbilical pain		
Pain characteristics	Progressive; stinging; pain on movement		
Gastrointestinal tract	Nausea; vomiting		
Other tracts	No abnormalities		
Physical examination			
General	Temp: 37.6 °C	blood pressure: 136/76	heart rate: 68/min
Surgical scars	Yes, peri-umbilical		
Abdominal examination	Painful palpation RLQ; rebound tenderness		
Further examination	No abnormalities		
Abnormal laboratory findings			
	ASAT	58 U/l 37C (elevated)	
	Creatinine	58 μmol/l (lowered)	
	Urine sediment		
	Leucocytes (stick)	+/−	
	Albumin (stick)	+/−	

Information as described on original application:
Pain for 4 days, which started in the umbilical region and migrated to the right lower quadrant. Painful RLQ on palpation.
Surgical history of umbilical hernia correction.
Question: Appendicitis, salpingitis, cystitis?

Appendix II

Diagnoses		Observer 2								
		Peritonitis	Perforated viscus	Bowel ischemia	Appendicitis	Diverticulitis	Bowel obstruction	Cholecystitis	Pancreatitis	Gynecological disease urgent
Observer 1	Peritonitis	1	–	–	–	–	–	–	–	–
	Perforated viscus	–	1	–	–	–	–	–	–	–
	Bowel ischemia	–	–	1	–	–	–	–	–	–
	Appendicitis	–	–	–	44	2	1	–	–	–
	Diverticulitis	–	1	–	–	20	–	1	–	–
	Bowel obstruction	–	–	–	–	–	11	–	–	–
	Cholecystitis	–	–	–	–	–	–	8	–	–
	Pancreatitis	–	–	–	–	–	–	–	6	–
	Gynecological disease urgent	–	–	–	1	1	–	–	–	1
	Urinary tract disease urgent	–	–	–	–	–	–	–	–	–
	Abscess	–	–	–	–	–	–	–	–	–
	Extra abdominal	–	–	–	–	–	–	–	–	–
	Gastro intestinal non-urgent	–	–	–	2	–	1	–	–	–
	IBD	–	–	–	–	–	1	–	–	–
	HPB	–	–	–	–	–	–	2	2	–
	Malignancy	–	–	–	–	–	–	–	–	–
	Gynecological disease non-urgent	–	–	–	–	–	–	–	–	–
	Urinary tract disease non-urgent	–	–	–	–	1	–	–	–	–
	NSAP	–	–	–	2	–	–	1	–	1
	Other	–	–	–	–	–	–	1	–	–
	Total	1	2	1	49	24	14	13	8	2

IBD inflammatory bowel disorders, *HPB* hepatic pancreatic biliary disease

Urinary tract disease urgent	Abscess	Extra abdominal	Gastro intestinal non-urgent	IBD	HPB	Malignancy	Gynecological disease non-urgent	Urinary tract disease non-urgent	NSAP	Other	Total
–	–	–	–	–	–	–	–	–	–	–	1
–	–	–	–	–	–	–	–	–	–	–	1
–	–	–	1	–	–	–	–	–	–	–	2
–	1	–	1	–	–	–	–	1	5	–	55
–	–	–	–	–	–	–	–	–	–	–	22
–	–	–	–	–	–	–	–	–	–	–	11
–	–	–	–	–	–	–	–	–	–	–	8
–	–	–	–	–	1	–	–	–	1	–	8
–	1	–	–	–	–	–	1	–	1	–	6
6	–	–	–	–	–	–	–	–	–	–	6
–	1	–	–	–	–	–	–	–	–	–	1
–	–	1	–	–	–	–	–	–	–	–	1
–	–	–	5	1	–	–	–	–	–	2	11
–	–	–	–	2	–	–	–	–	–	–	3
–	–	–	–	–	4	–	–	–	1	1	10
–	–	–	–	–	–	1	–	–	–	–	1
–	–	–	–	–	–	–	3	–	–	–	3
1	–	–	1	–	–	–	–	1	2	–	6
–	1	1	4	–	2	–	1	3	21	–	37
–	–	–	–	–	1	–	–	–	2	3	7
7	4	2	12	3	8	1	5	5	33	6	200

		Observer 3								
		Perforated viscus	Appendicitis	Diverticulitis	Bowel obstruction	Cholecystitis	Pancreatitis	Gynecological disease urgent	Urinary tract disease urgent	Abscess
Observer 1	Peritonitis	1	–	–	–	–	–	–	–	–
	Perforated viscus	–	–	–	–	–	–	–	–	–
	Bowel ischemia	–	–	–	1	–	–	–	–	–
	Appendicitis	–	49	–	1	–	–	1	–	–
	Diverticulitis	–	–	20	–	1	–	–	–	–
	Bowel obstruction	–	–	–	9	–	–	–	–	–
	Cholecystitis	–	–	–	–	8	–	–	–	–
	Pancreatitis	–	–	–	–	–	6	–	–	–
	Gynecological disease urgent	–	1	–	–	–	–	1	–	2
	Urinary tract disease urgent	–	–	–	–	–	–	–	4	–
	Abscess	–	–	–	–	–	–	–	–	1
	Extra abdominal	–	–	–	–	–	–	–	–	–
	Gastro intestinal non-urgent	–	3	–	1	–	–	1	–	–
	IBD	–	–	–	1	–	–	–	–	–
	HPB	–	–	–	–	–	2	–	–	–
	Malignancy	–	–	–	–	–	–	–	–	–
	Gynecological disease non-urgent	–	–	–	–	–	–	–	–	–
	Urinary tract disease non-urgent	–	–	–	–	–	–	–	1	–
	NSAP	–	3	–	2	–	–	2	1	1
	Other	–	–	–	–	–	1	1	–	–
	Total	1	56	20	15	9	9	6	6	4

Extra abdominal	Gastrointestinal non-urgent	IBD	HPB	Malignancy	Gynecological disease non-urgent	Urinary tract disease non-urgent	NSAP	Other	Total
–	–	–	–	–	–	–	–	–	1
–	1	–	–	–	–	–	–	–	1
–	1	–	–	–	–	–	–	–	2
–	1	–	–	–	–	–	3	–	55
–	1	–	–	–	–	–	–	–	22
–	–	–	1	–	–	–	1	–	11
–	–	–	–	–	–	–	–	–	8
–	–	–	1	–	–	–	–	1	8
–	–	–	–	–	1	–	1	–	6
–	–	–	–	–	–	–	2	–	6
–	–	–	–	–	–	–	–	–	1
1	–	–	–	–	–	–	–	–	1
–	3	2	–	–	–	–	1	–	11
–	1	1	–	–	–	–	–	–	3
1	1	–	5	1	–	–	–	–	10
–	–	–	–	1	–	–	–	–	1
–	–	–	–	–	3	–	–	–	3
–	–	–	–	–	–	3	2	–	6
–	4	–	2	–	2	–	19	1	37
–	–	–	3	–	–	–	1	1	7
2	13	3	12	2	6	3	30	3	200

		Observer 3								
		Perforated viscus	Appendicitis	Diverticulitis	Bowel obstruction	Cholecystitis	Pancreatitis	Gynecological disease urgent	Urinary tract disease urgent	Abscess
Observer 2	Peritonitis	1	–	–	–	–	–	–	–	–
	Perforated viscus	–	–	–	–	–	–	–	–	–
	Bowel ischemia	–	–	–	1	–	–	–	–	–
	Appendicitis	–	49	–	–	–	–	–	–	–
	Diverticulitis	–	1	20	–	–	–	–	1	–
	Bowel obstruction	–	–	–	12	–	–	–	–	–
	Cholecystitis	–	–	–	–	9	–	–	–	–
	Pancreatitis	–	–	–	–	–	8	–	–	–
	Gynecological disease urgent	–	–	–	–	–	–	1	–	–
	Urinary tract disease urgent	–	–	–	–	–	–	–	4	–
	Abscess	–	–	–	1	–	–	1	–	2
	Extra abdominal	–	–	–	–	–	–	–	–	–
	Gastro intestinal non-urgent	–	1	–	1	–	–	2	–	–
	IBD	–	–	–	–	–	–	–	–	–
	HPB	–	–	–	–	–	–	–	–	–
	Malignancy	–	–	–	–	–	–	–	–	–
	Gynecological disease non-urgent	–	–	–	–	–	–	–	–	–
	Urinary tract disease non-urgent	–	1	–	–	–	–	–	1	–
	NSAP	–	3	–	–	–	–	2	–	2
	Other	–	1	–	–	–	1	–	–	–
	Total	1	56	20	15	9	9	6	6	4

Extra abdominal	Gastro intestinal non-urgent	IBD	HPB	Malignancy	Gynecological disease non-urgent	Urinary tract disease non-urgent	NSAP	Other	Total
–	–	–	–	–	–	–	–	–	1
–	2	–	–	–	–	–	–	–	2
–	–	–	–	–	–	–	–	–	1
–	–	–	–	–	–	–	–	–	49
–	1	–	–	–	–	–	1	–	24
–	–	–	1	–	–	–	1	–	14
–	1	–	3	–	–	–	–	–	13
–	–	–	–	–	–	–	–	–	8
–	–	–	–	–	1	–	–	–	2
–	–	–	–	–	–	1	2	–	7
–	–	–	–	–	–	–	–	–	4
1	–	–	–	–	–	–	1	–	2
–	4	2	–	–	–	–	2	–	12
–	1	1	–	–	–	–	1	–	3
–	1	–	7	–	–	–	–	–	8
–	–	–	–	1	–	–	–	–	1
–	–	–	–	–	4	–	1	–	5
–	–	–	–	–	–	1	2	–	5
1	2	–	1	–	1	1	18	2	33
–	1	–	–	1	–	–	1	1	6
2	13	3	12	2	6	3	30	3	200

References

1. McDonald GP, Pendarvis DP, Wilmoth R, Daley BJ (2001) Influence of preoperative computed tomography on patients undergoing appendectomy. Am Surg 67:1017–1021
2. Weyant MJ, Eachempati SR, Maluccio MA et al (2000) Interpretation of computed tomography does not correlate with laboratory or pathologic findings in surgically confirmed acute appendicitis. Surgery 128:145–152
3. Rao PM, Rhea JT, Novelline RA, Mostafavi AA, Lawrason JN, McCabe CJ (1997) Helical CT combined with contrast material administered only through the colon for imaging of suspected appendicitis. AJR Am J Roentgenol 169:1275–1280
4. Rao PM, Rhea JT, Novelline RA, Mostafavi AA, McCabe CJ (1998) Effect of computed tomography of the appendix on treatment of patients and use of hospital resources. N Engl J Med 338:141–146
5. Ng CS, Watson CJ, Palmer CR et al (2002) Evaluation of early abdominopelvic computed tomography in patients with acute abdominal pain of unknown cause: prospective randomised study. BMJ 325:1387
6. Keyzer C, Tack D, De Maertelaer V, Bohy P, Gevenois PA, Van GD (2004) Acute appendicitis: comparison of low-dose and standard-dose unenhanced multi-detector row CT. Radiology 232:164–172
7. Tack D, Bohy P, Perlot I et al (2005) Suspected acute colon diverticulitis: imaging with low-dose unenhanced multi-detector row CT. Radiology 237:189–196
8. Weltman DI, Yu J, Krumenacker J Jr, Huang S, Moh P (2000) Diagnosis of acute appendicitis: comparison of 5- and 10-mm CT sections in the same patient. Radiology 216:172–177
9. Wise SW, Labuski MR, Kasales CJ et al (2001) Comparative assessment of CT and sonographic techniques for appendiceal imaging. AJR Am J Roentgenol 176:933–941
10. Laméris W, van Randen A, Dijkgraaf MG, Bossuyt PM, Stoker J, Boermeester MA. Optimization of diagnostic imaging use in patients with acute abdominal pain (OPTIMA): Design and rationale (2007). BMC Emerg Med 7:9 (Aug 6)
11. Irwig L, Macaskill P, Walter SD, Houssami N (2006) New methods give better estimates of changes in diagnostic accuracy when prior information is provided. J Clin Epidemiol 59:299–307
12. Altman D (1997) Practical statistics for medical research, 1st edn. Chapman and Hall, London, pp 403–407
13. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
14. Benjaminov O, Atri M, Hamilton P, Rappaport D (2002) Frequency of visualization and thickness of normal appendix at nonenhanced helical CT. Radiology 225:400–406
15. Zangos S, Steenburg SD, Phillips KD et al (2007) Acute abdomen: Added diagnostic value of coronal reformations with 64-slice multidetector row computed tomography. Acad Radiol 14:19–27
16. Loy CT, Irwig L (2004) Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA 292:1602–1609