Effect of Directed Training on Reader Performance for CT Colonography: Multicenter Study[1]

European Society of Gastrointestinal and Abdominal Radiology CT Colonography Study Group Investigators

Purpose: To define the interpretative performance of radiologists experienced in computed tomographic (CT) colonography and to compare it with that of novice observers who had undergone directed training, with colonoscopy as the reference standard.

Materials and Methods: Physicians at each participating center received ethical committee approval and followed the committees' requests regarding informed consent. Nine experienced radiologists, nine trained radiologists, and 10 trained technologists from nine centers read 40 CT colonographic studies selected from a data set of 51 studies and modeled to simulate a population with positive fecal occult blood test results: Studies were obtained in eight patients with cancer, 12 patients with large polyp, four patients with medium polyp, and 27 patients without colonic lesions. Findings were verified with colonoscopy. An experienced radiologist used 50 endoscopically validated studies to train novice observers before they were allowed to participate. Observers used one software platform to read studies over 2 days. Responses were collated and compared with the known diagnostic category for each subject. The number of correctly classified subjects was determined for each observer, and differences between groups were examined with bootstrap analysis.

Results: Overall, 28 observers read 1084 studies and detected 121 cancers, 134 large polyps, and 33 medium polyps; 448 healthy subjects were categorized correctly. Experienced radiologists detected 116 lesions; trained radiologists and technologists detected 85 and 87 lesions, respectively. Overall accuracy of experienced observers (74.2%) was significantly better than that of trained radiologists (66.6%) and technologists (63.2%). There was no significant difference ($P = .33$) between overall accuracy of trained radiologists and that of technologists; however, some trainees reached the mean performance achieved by experienced observers.

Conclusion: Experienced observers interpreted CT colonographic images significantly better than did novices trained with 50 studies. On average, no difference between trained radiologists and trained technologists was found; however, individual performance was variable and some trainees outperformed some experienced observers.

[1] The complete list of investigators and affiliations is listed at the end of this article. Received June 15, 2005; revision requested August 15; revision received February 10, 2006; accepted March 6; final version accepted July 13. Supported by a grant from the European Association of Radiology administered by the European Society of Gastrointestinal and Abdominal Radiology and a Kodak Scholarship administered by the Royal College of Radiologists, United Kingdom. **Address correspondence to** Steve Halligan, MD, FRCP, FRCR, Department of Radiology, University College London, Level 2 Podium, 235 Euston Rd, London NW1 2BU, England (e-mail: *s.halligan@ucl.ac.uk*).

Computed tomographic (CT) colonography is available for colorectal cancer screening in the United States (1,2), and a survey revealed that 36% of radiology departments in the United Kingdom offered this service (3). To date, most research has focused on the technical capabilities of CT colonography; however, it is increasingly being realized that observer experience and training are equally important (4,5). A study in which 18 radiologists were asked to interpret CT colonographic images revealed that observer performance was related to prior experience (6).

At the time of this writing, there were no evidence-based guidelines for training; however, a working group suggested that supervised interpretation of at least 40 validated studies might be adequate for this purpose (7). However, when this suggestion was tested on a small scale, it was shown that observer response to such training is highly unpredictable and that performance may even deteriorate (8). Also, a recent study revealed that some novice observers exposed to a training module could outperform their more-experienced colleagues (9).

The European Society of Gastrointestinal and Abdominal Radiology is interested in developing evidence-based guidelines for training and accrediting radiologists in the interpretation of CT colonographic studies. With this goal in mind, the purpose of our study was to define the interpretative performance of radiologists with experience in CT colonography and to compare their performance with that of novice observers who had undergone directed training, with colonoscopy serving as the reference standard.

Advances in Knowledge

- Experienced observers interpreted CT colonography studies significantly better than did novice readers trained with 50 studies.
- On average, we found no difference between trained radiologists and trained technologists; however, individual performance was variable and some trainees outperformed some experienced observers.

Materials and Methods

Physicians at each participating center received ethical committee approval and followed the committee's requests regarding informed patient consent. Voxar (Edinburgh, Scotland) provided the software that was used with laptop computers, and E-Z-Em (Westbury, NY) provided the two workstations and software that were used at the trial office. The authors had full control of all data and information submitted for publication.

Data Set Composition and Accrual

We collated a data set of normal and abnormal CT colonographic studies submitted by the seven participating centers. In this data set, prevalence and morphology of neoplasia were modeled to simulate those expected in patients with positive fecal occult blood test results (prevalence of cancer, 10%; prevalence of large polyps, 30%; prevalence of medium polyps, 10%; prevalence of normal colorectum, 50%) (10–12). The aim of this procedure was to create a mix of studies with normal and abnormal findings to investigate sensitivity for different classes of neoplasia and specificity between observer groups. This mix also ensured that the data set was clinically relevant.

Physicians at each center were asked to submit 10 studies that were obtained in subjects aged 50–69 years and that matched the expected prevalence of neoplasia on the basis of fecal occult blood test results (ie, one patient with cancer, three patients with large polyps, one patient with medium polyps, and five subjects with no colonic lesion) (10–12). The four diagnostic categories were as follows: cancer, large polyps, medium polyps, and normal. In line with the results of fecal occult blood test trials, a large polyp was defined as a polyp measuring 10 mm or more in diameter and a medium polyp was defined as a polyp measuring less than 10 mm in diameter (6–9 mm in diameter for the purposes of this study).

To reflect normal variation in data quality, subjects from each center were recruited in a strictly chronologically consecutive fashion. That is to say, consecutive subjects were assigned to an appropriate diagnostic category until all four categories were full. Thus, the first patient with cancer completed recruitment to this category, whereas three consecutive patients with large polyps were necessary to complete recruitment to this category. Patients with multiple lesions were assigned to a category according to the largest lesion detected, which was referred to as the index lesion. To ensure that studies accurately reflected the natural and inevitable technical variation found in day-to-day practice, centers were obliged to submit all eligible studies, with the exception of those that were deemed nondiagnostic (ie, any study in which the local principal investigator would normally recommend repeat colonography or another examination because of insurmountable technical problems, such as segmental collapse or retained fluid). In all subjects, CT findings were defined by subsequent same-day colonoscopic findings obtained by experienced practitioners. Polyp size was based on the colonoscopic measurement, which was estimated with adjacent biopsy forceps.

Participating centers were chosen because they had active CT colonography research programs at the time the study protocol was developed and because they could contribute studies. We also stated that as long as studies were chronologically consecutive, centers could

Published online
10.1148/radiol.2421051000

Radiology 2007; 242:152–161

Author contributions:
Guarantors of integrity of entire study, S.H., D.B.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, S.H., D.B., W.A.; clinical studies, S.H., D.B., H.F., R.F., S.T., D.N., J.W.B., J.F., M.P., V.V.d.H., P.L., J.M., G.D., A.O., S.F., E.N., P.V., R.I., F.M., D.R.; statistical analysis, S.H., D.G.A., P.B.; and manuscript editing, S.H., D.B., W.A., C.B., H.F., A.L., J.S.

See Materials and Methods for pertinent disclosures.

submit retrospective studies, provided the technical stipulations for CT colonography were satisfied. Five centers submitted only retrospective data, and two centers submitted only prospective data. Ethical permission for data sharing was covered by the local stipulations at each center. Four centers that submitted retrospective data did not require additional specific ethical committee approval for this study because data were collected as part of a local study, and ethical committee approval and patient informed consent were applicable for additional analyses and data sharing. The fifth center that submitted retrospective data obtained ethical committee approval and verbal consent via telephone from the subjects selected. Physicians at the two centers that submitted prospective data obtained patient informed consent, ethical committee approval, and permission for additional analyses and data sharing with ongoing studies, provided patient identifying information was removed before data sharing. We applied this stipulation to all data collected for this study.

Images were acquired with the patient in the prone and supine positions with full bowel purgation, a collimation that was no greater than 2.5 mm, and use of a multi–detector row CT scanner. Gas insufflation and spasmolytic use were left to the discretion of local physicians. Low-radiation-dose protocols were permissible, but administration of intravenous contrast material was impermissible on the grounds that contrast agents are unlikely to be used in a screening program. Fecal tagging was not permitted since this procedure was not common practice at the time of data accrual (May to November 2003). Studies were archived on a compact disk and transferred to the trial office; technical and diagnostic category data were included for each study.

Of the seven centers that submitted data, three provided data only from symptomatic patients, one provided data only from asymptomatic patients, and three provided data from both symptomatic and asymptomatic patients. Four centers submitted studies obtained in a full data set of 10 subjects; however, the file containing one study could not be opened at the trial office. The three remaining centers submitted five, four, and three studies because of difficulties satisfying protocol requirements, notably, those related to age. Thus, our study included 51 subjects, of whom 27 (53%) had no colonic lesion and 24 (47%) had an index lesion. Of the 24 patients with an index lesion, eight had cancer, 12 had large polyps, and four had medium polyps. Seven (29%) of the patients with an index lesion had a second lesion: Two patients with cancer each had an additional large polyp, and one patient with cancer and four patients with large polyps each had an additional medium polyp.

Observers

The data set was interpreted by the following three groups of observers: experienced radiologists and trained radiologists and radiologic technologists. Nine centers (including all seven that submitted imaging studies) provided an observer for each group; one center provided two technologists.

An experienced radiologist was defined as a radiologist who had considerable practical and/or research experience with CT colonography prior to this study. Individual experience ranged from evaluation of 325 to evaluation of 1200 studies (median, 750 studies), with between 120 and 600 studies (median, 200 studies) validated with colonoscopy.

Each experienced radiologist identified a local radiologist and radiologic technologist who had interpreted 10 or fewer studies prior to this study. We stipulated that radiologists be familiar with the interpretation of standard abdominopelvic CT studies and that technologists be familiar with the acquisition of abdominopelvic CT studies. The experienced radiologists used normal and abnormal studies that had been acquired locally and verified with subsequent colonoscopy to train inexperienced radiologists and technologists to interpret CT colonographic images. There was no attempt to use the same training data set at all participating centers because we wanted to emulate existing training programs for conventional CT (in which trainees generally learn by using studies acquired locally). However, we did stipulate that 50 individual studies should be interpreted; interpretation was to be unaided initially and then followed by face-to-face discussion with the local trainer on a patient-by-patient basis, so as to closely mimic standard day-to-day training practice. Trainers and trainees used the preferred local reading platform, in line with everyday practice. We stipulated that training should occur over several separate sessions and several weeks to reflect standard teaching practice.

Reading Conditions and Outcome Measures

After training, an individualized test data set of 40 studies was prepared by the trial coordinator for each participating center. These 40 studies were sampled from the data set of 51 studies and balanced in terms of the prevalence of abnormalities; studies submitted by a center were excluded from the data set sent to that center. The order of studies was randomized to mix abnormal and normal cases, and all readers read the studies in the same order. All patient identifiers were removed. The experienced radiologist ($n = 9$), trained radiologist ($n = 9$), and trained technologist(s) ($n = 10$) at each center then interpreted this data set over 2 days. The trial coordinator visited each center to supervise reading, which was conducted with individual laptop computers equipped with 17-inch (43.18-cm) screens and software that allowed a primary two-dimensional analysis, with three-dimensional analysis available for problem solving (Voxar ColonScreen, version 2.2; Barco, Edinburgh, Scotland). Observers were familiarized with the software, when necessary, and the supervisor was available at all times. Reading was performed in a quiet environment with ambient light. Observers were asked to read at their own pace, with no requirement to finish within a prespecified time. Observers had read the study protocol and knew that studies obtained at their own institutions (if any) had been excluded, but they had no

specific information about the composition of their individualized data set.

Observers used a data sheet to categorize each subject as either healthy or unhealthy. Subjects designated as unhealthy were further categorized as having cancer, a large polyp, or a medium polyp. Large polyps had a maximal two-dimensional transverse diameter of 10 mm or larger, whereas medium polyps had a diameter of 6–9 mm; software calipers were used to obtain these measurements. Observers noted any polyp that measured 5 mm or less but categorized subjects with such polyps as healthy; this practice allowed false-negative findings due to measurement error to be distinguished from false-negative findings due to perceptual error. Observers were unaware of each other's responses. Prone and supine image coordinates and segmental location were recorded for each perceived abnormality so that false-positive responses could be distinguished from true-positive responses in the same patient. Multiple responses were possible. There were six bowel segments (rectum, sigmoid colon, descending colon, transverse colon, ascending colon, and cecum), and observers were provided with an annotated diagram of segmental definitions. Observers were free to classify a study as technically inadequate, although steps had been taken when designing the study protocol to avoid including nondiagnostic studies.

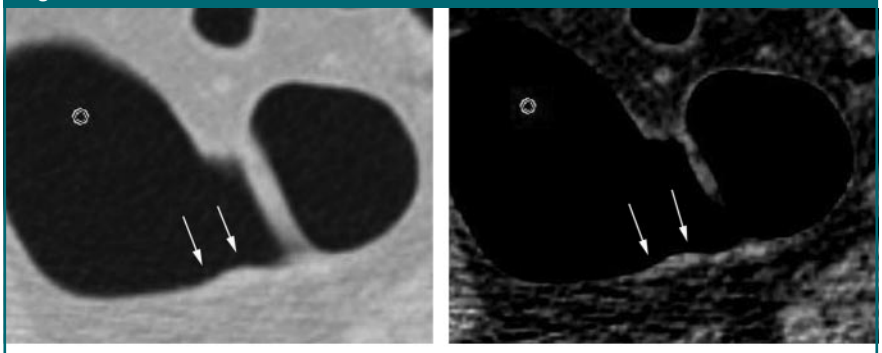
Data sheets were collated, and observers' responses were compared with the known diagnostic category. The trial coordinator (who had experience with more than 300 endoscopically verified studies) independently evaluated each study to confirm both the CT findings reported by the submitting center and the CT coordinates of the abnormality, which were then used to determine whether observers' responses were true-positive or false-positive. All but one of the endoscopically validated lesions could be identified. However, observers encountered difficulty locating four flat adenomas (which measured 40, 30, 15, and 12 mm in diameter), one of which was only visible when standard abdominal CT window settings were used (window level, 40 HU; window width, 400 HU) (Fig 1). One flat adenoma (40 mm) could not be identified despite good bowel preparation and distention and a thorough review of endoscopic data.

Statistical Analysis

Observer responses were compared with the known diagnostic category and lesion coordinates; the numbers of true-positive, true-negative, false-positive, and false-negative classifications were determined. Individual and group performance was determined by calculating the number and percentage of studies in which the index lesion (and second lesion in seven patients) was correctly identified and the number and percentage of normal studies that were correctly categorized. The number of false-positive classifications in healthy subjects and in patients known to have an index lesion was determined.

Two measures were derived for each reader: sensitivity for lesions (number of lesions correctly seen divided by number of lesions present) and accuracy (the overall percentage of correct categorizations). For both measures, studies classified as technically inadequate by readers were included. For the most part, observers read the same studies and observations were correlated to some extent; therefore, a bootstrap analysis was used to investigate differences between observer groups. A total of 1999 samples were redrawn randomly from the original sample, with replacement and analysis of each resultant data set. The results of interest were calculated for each bootstrap sample, and the distribution of values was used to obtain a bootstrap confidence interval. A probability value was also calculated by considering how many of the values were farther from zero than the actual value observed with the data. Results were considered statistically significant at a probability level of 5%. Statistical analysis was performed with Stata, version 8.0, software (Stata, College Station, Tex).

Figure 1



a. b.

Figure 1: Transverse CT images acquired with a four–detector row scanner (100 mA, 120 kV) in a 62-year old woman with a 12-mm-diameter sigmoid flat adenoma. All observers missed this polyp. **(a)** The adenoma (arrows) is barely visible with standard CT colonography window settings (window level −150 HU; window width, 1500 HU). **(b)** The adenoma (arrows) is more clearly visible with standard abdominal CT window settings (window level, 40 HU; window width, 400 HU).

Results

The 28 observers read a total of 1084 individual studies; 22 (79%) readers (including all nine experienced observers) read all 40 studies assigned to them, two read 39, one read 37, one read 35, and two read 27 because of time constraints.

Overall, 736 (68%) patients were correctly classified (Table 1). The number of lesions correctly classified declined in conjunction with decreased size of the index lesion: Cancer was detected in 121 (79%) patients, large pol-

yps were detected in 134 (47%), and medium polyps were detected in 33 (36%) (Table 1). In the remaining 348 patients, the index lesion was missed in 239, findings were false-positive in 73, and studies were deemed technically inadequate in 36. Of the 36 technically inadequate studies, 23 (64%) related to one subject who had no colonic lesion. Of the 13 other technically inadequate studies, 11 related to subjects who had no colonic lesion. Overall, the false-positive rate was 13% (22 of 176 studies) for experienced radiologists, 12% (21 of 169 studies) for trained radiologists, and 16% (30 of 188 studies) for technologists. One experienced radiologist and two technologists did not assign any false-positive diagnoses. Six readers (one experienced observer, two radiologists, and three technologists) had false-positive rates of 20% or more.

Table 1

Relationship between Patient Category and Observer Assessment for All Observer Groups Combined

Patient Category	Correct Classification	Incorrect Classification	False-Positive Finding	Technically Inadequate Study	Total
Cancer	121 (79)	33 (21)	· · ·	0	154
Large polyp	134 (47)	147 (52)	· · ·	2	283
Medium polyp	33 (36)	59 (64)	· · ·	0	92
No colonic lesion	448 (81)	· · ·	73	34	555
Total	736	239	73	36	1084

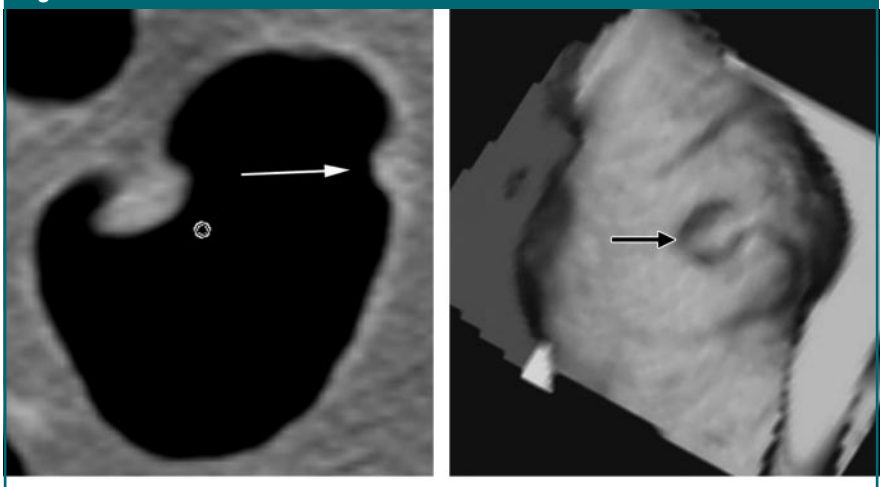
Note.—Data are numbers of studies. Data in parentheses are percentages.

Observer Performance

Overall, more lesions were detected by experienced radiologists (66%) than by trained radiologists (51%) or technologists (47%) (Table 2). This was also the case when all subgroups of lesions were considered individually.

Subset analysis revealed that some polyps were clearly more difficult to detect than others; this phenomenon applied across all observer groups. For example, in the 12 patients whose index lesion was a large polyp, two polyps were missed by all 24 observers who read these two studies. The large polyp in two of these 12 patients was identified and categorized correctly by only one observer (an experienced reader). All four of these difficult-to-detect polyps were morphologically flat. In the four patients whose index lesion was a medium polyp, only one (4%) of 23 observers identified the index lesion in one study (Fig 2), and only two (10%) of 21 observers identified the index lesion in another study. Thus, there were two difficult-to-detect medium polyps. Overall, the six difficult-to-detect polyps (four large and two medium polyps) had considerable influence on our results and decreased accuracy for observer groups and individuals.

Figure 2



a. **b.**

Figure 2: Transverse CT images acquired with a four–detector row scanner (100 mA, 120 kV) in a 66-year old woman with an 8-mm adenoma in the descending colon. This polyp was visible on only those images obtained with the patient in the prone position, and it was missed by all readers except one experienced observer. **(a)** Polyp (arrow) is visible with CT colonography window settings (window level, −150 HU; window width, 1500 HU). **(b)** Three-dimensional volume-rendered endoluminal view shows this same polyp (arrow).

Accuracy and Sensitivity

In regard to the bootstrap analysis, overall accuracy and sensitivity values were significantly higher for experienced radiologists than for trained radiologists or technologists (Table 3). This was the case for all analyses, regardless of whether the six difficult-to-detect polyps were included. However, there was no significant difference in measures of accuracy or sensitivity when the trained radiologists were compared with the trained technologists. Although these results show that experienced readers performed best on average, there was considerable overlap between the observer groups when individual performance was considered (Figs 3, 4); for example, the level of accuracy achieved by one technologist and two trained radiologists was higher than the mean accuracy achieved by experienced observers.

Secondary Lesions

The ability of observers to detect the seven secondary lesions is shown in Table 4. The secondary lesion was a large

polyp in 14 of the studies read by experienced radiologists, 14 of the studies read by trained radiologists, and 15 of the studies read by technologists. This lesion was detected by all of the experienced radiologists, 13 (93%) of the trained radiologists, and 12 (80%) of the technologists (Table 3). The secondary lesion was a medium polyp in 35 of the studies read by experienced radiologists, 33 of the studies read by radiologists, and 38 of the studies read by technologists. This lesion was detected by 13 (37%) experienced radiologists, seven (21%) trained radiologists, and six (16%) technologists.

Prior Experience Levels

We performed a subset analysis to compare experienced readers whose a priori experience exceeded 1000 studies (four individuals) with those whose experience did not exceed 1000 studies (six individuals) and found no significant difference: Rates for detection of cancer, large polyps, and medium polyps were 100%, 56%, and 43%, respectively, for readers with the most experience and 86%, 58%, and 53%, respectively, for readers with the least experience. False-positive rates were also similar (80% for readers with the most experience vs 84% for readers with the least experience).

Discussion

Unsurprisingly, prior experience enhances performance. Investigators in a prior study found the average area under the receiver operating characteristic curve was 0.80 for the most experienced readers and 0.77 for the least experienced readers (6); this is a small difference in relative terms. Highly experienced individuals agree that specific and supervised training is a prerequisite for acceptable performance (7). Moreover, they specified that such training should involve interpretation of 40–50 endoscopically validated studies. However, there is little evidence to support or refute this recommendation. Our data show that the overall sensitivity of novice observers trained with this scheme is significantly inferior to that of experienced observers. Both groups of trained observers detected approximately 70% of cancers, whereas experienced observers detected 92% of can-

Table 2

Summary of Lesion Detection Rates according to Observer Experience

	All Lesions		Cancer		Large Polyps		Medium Polyps	
Observer Group	Seen	Missed	Seen	Missed	Seen	Missed	Seen	Missed
Experienced radiologists	116 (66)	60 (34)	47 (92)	4 (8)	54 (57)	40 (43)	15 (48)	16 (52)
Trained radiologists	85 (51)	82 (49)	34 (71)	14 (29)	42 (47)	48 (53)	9 (31)	20 (69)
Trained technologists	87 (47)	97 (53)	40 (73)	15 (27)	38 (39)	59 (61)	9 (28)	23 (72)

Note.—Data are numbers of lesions. Data in parentheses are percentages.

Table 3

Observer Accuracy and Sensitivity for All Lesions and When Six Difficult-to-Detect Polyps Were Excluded

A: Accuracy and Sensitivity for Each Observer Group

Observer Group	Overall Accuracy	Overall Accuracy Excluding Difficult Cases	Sensitivity	Sensitivity Excluding Difficult Cases
Experienced radiologists	74.2	83.7	65.5	85.3
Trained radiologists	66.6	76.9	50.7	69.7
Trained technologists	63.2	72	47.3	63.5

B: Difference in Accuracy and Sensitivity between Observer Groups

Group Comparison	Difference in Overall Accuracy	*P* Value	Difference in Overall Accuracy Excluding Difficult Cases	*P* Value	Difference in Sensitivity	*P* Value	Difference in Sensitivity Excluding Difficult Cases	*P* Value
Experienced radiologists vs trained radiologists	7.6 (1.2, 14.3)	.017	6.8 (0.5, 13.1)	.035	14.9 (4.3, 25.2)	.007	15.6 (5.7, 25.8)	.004
Experienced radiologists vs trained technologists	11.0 (0.5, 17.7)	.003	11.7 (5.2, 17.9)	.001	18.2 (8.3, 28.3)	.002	21.9 (6.2, 16.9)	.001
Trained radiologists vs trained technologists	3.4 (−3.4, 10.1)	.33	4.9 (−1.8, 11.7)	.16	3.4 (−7.2, 13.7)	.52	6.1 (−0.8, 10.7)	.31

Note.—Unless otherwise indicated, data are percentages. Data in parentheses are 95% confidence intervals. Accuracy refers to the correct classification of patients with and without lesions, whereas sensitivity refers to detection of cancer and polyps only.

Figures 3, 4

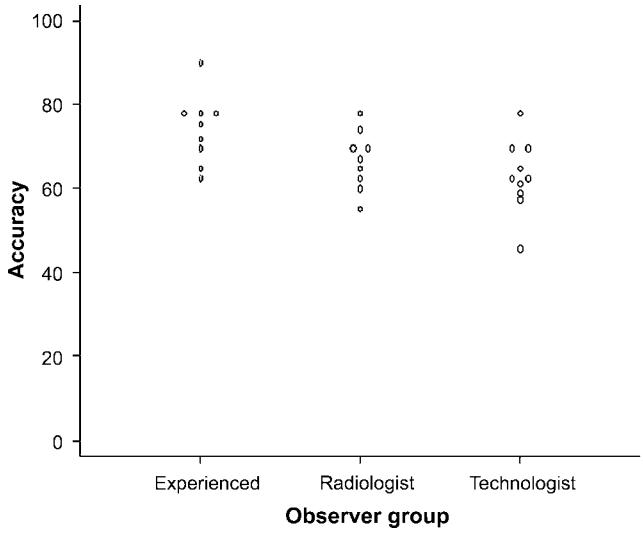


Figure 3: Graph shows overall observer accuracy. Mean values were 74.2% for experienced observers, 66.6% for trained radiologists, and 63.2% for technologists.

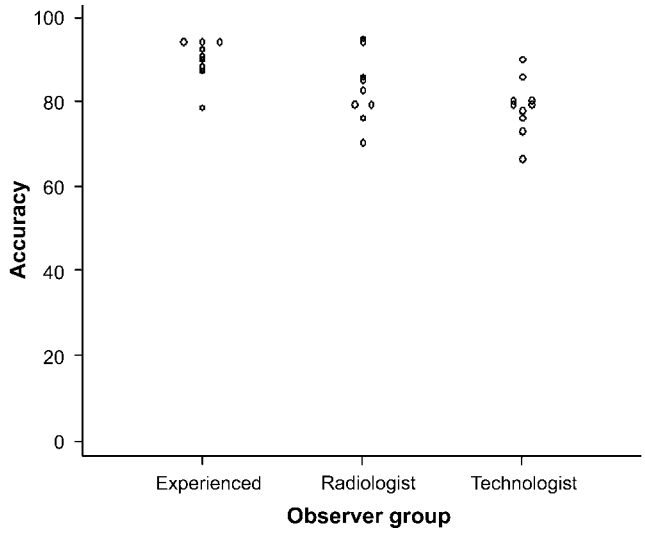


Figure 4: Graph shows observer accuracy after six difficult-to-detect lesions were excluded. Mean values were 83.7% for experienced observers, 76.9% for trained radiologists, and 72.0% for technologists.

cers. This discrepancy seems to suggest that the training we administered was inadequate.

However, it could be argued that our proposition that trainees should achieve the competence of experienced observers is flawed. A distinction can be made between "best achievable" and "acceptable" performance. On average, subspecialist radiologists perform better than their generalist peers because subspecialists are able to make decisions on the basis of prior experience (13). Whether all radiologists interpreting CT colonographic studies need to be as capable as those with extensive experience is a question to be answered by the wider radiologic community. The answer will depend on whether the examination is performed by generalists or subspecialists. A number of observations will indicate this. First, subspecialization has affected radiology since the 1920s (14); since then, it has become more prevalent for a number of reasons, not the least of which is that it is thought to benefit patients (15). The ultimate position of CT colonography as a specialist examination is thus more likely today than it was previously. A parallel may be drawn with a barium enema examination, which is widely considered a general examination despite compelling evidence that interpretation of barium enema studies is best handled by those with extensive experience (16). Furthermore, since the inception of CT colonography, the diagnostic performance of this modality has been compared with that of colonoscopy, which is reportedly a more effective test than a barium enema examination (17). Comparisons between skilled colonoscopists and less-skilled colonographers damage the reputation of CT colonography (18).

It may be possible to stratify acceptable performance contingent on the clinical setting. For example, it has been argued that for mammography, the highest aptitude is necessary for screening because patients are asymptomatic and lesions are often difficult to detect (19). The same principle might apply to CT colonography. Like mammography, colonography may be used to examine symptomatic patients (who actually constitute the largest group that undergoes this procedure in research studies). Symptomatic colonic lesions tend to be larger and easier to detect than asymptomatic lesions; thus, it may be possible that less interpretative skill is needed to detect symptomatic lesions. This hypothesis is supported by our findings, which show that detection rates increased in line with lesion size for all observer groups. However, while cancers were the easiest index lesions for the trained observers to detect,

Table 4

Ability to Detect a Secondary Lesion in Patients with More than One Lesion

	All Lesions		Large Polyps		Medium Polyps	
Observer Group	Seen	Missed	Seen	Missed	Seen	Missed
Experienced radiologists	27 (55)	22 (45)	14 (100)	0 (0)	13 (37)	22 (63)
Trained radiologists	20 (43)	27 (57)	13 (93)	1 (7)	7 (21)	26 (79)
Trained technologists	18 (34)	35 (66)	12 (80)	3 (20)	6 (16)	32 (84)

Note.—Data are numbers of observations. Data in parentheses are percentages. Seven patients had a second lesion (ie, a lesion other than the index lesion).

whether a potential patient or health-policy maker would be satisfied with a 70% average detection rate is a subject for wider debate.

It was not the aim of our study to investigate the performance characteristics of CT colonography. Rather, we aimed to determine the performance of novice observers relative to that of experienced observers after novice observers had undergone training with a schedule that was in line with proposed guidelines (7). With this approach, no aspect of individual aptitude is taken into account. It is inevitable that some individuals will outperform others despite similar professional backgrounds and training. Our data revealed considerable overlap in individual performance between all groups. Notably, there were two trained radiologists and one trained technologist whose accuracy exceeded the mean accuracy achieved by the experienced observers. Conversely, the accuracy of one experienced observer was below the mean accuracy achieved by both trained groups, even after difficult-to-detect lesions were excluded. Our data suggest that competence might be achieved by certain talented individuals after they complete a training program based on 50 validated studies. Merely completing such training is insufficient to guarantee competency, and it is self-evident that competent individuals will need to be identified in some other way, possibly with an examination. Again, this is a subject for wider debate. It should be noted that because our sample data set was relatively small, the observed variability between observers will likely exceed the real variability because of sampling error. A larger study would likely reveal performance that regressed toward the mean value for each group. Because of this, it would be unwise to overemphasize the performance of individuals in the present study.

Considering aptitude further, on average, we found no difference between the trained radiologists and the trained technologists, despite the radiologists' relative wealth of interpretative experience with CT. Also, the range of individual abilities was similar between these two groups. This suggests that the paradigm for interpretation of CT colonographic studies differs from that for interpretation of routine CT studies; thus, radiologists may not have an intrinsic advantage (unless we also consider the detection of extracolonic lesions, which we chose not to address). This may be explained by the fact that one organ is being examined for one disease (ie, neoplasia); therefore, an extensive medical knowledge base confers no substantial advantage. Furthermore, the skills required for colonic navigation are different from those used to interpret conventional CT studies, and interpretation takes longer, with a greater potential for observer fatigue and error (20). Our data possibly support the concept that radiographic technologists may be a valuable resource for interpretation of studies, especially when radiologists are in short supply. This is already the case for interpretation of barium enema studies, and it is a cost-effective measure (21,22).

Although our primary aim was to assess the relative performance of experienced and trained observers, we should explore the reasons behind the overall detection rate of only 57% of large polyps, which lags behind that in some studies (23) and meta-analyses (24,25). This was undoubtedly influenced by the disproportionately high percentage of flat adenomas (a third of large polyps were flat, and one was invisible on CT images, even in retrospect), and it may not translate to series that are more representative of the general population. The findings of large series in which dye-spray colonoscopy was used suggest that 13%–15% of large adenomas are flat (26,27). Ironically, the higher percentage of flat adenomas in our study was a result of our attempts to make the data set reflect conditions in everyday practice. We prevented investigators from submitting only their best studies by stipulating that studies be accrued in a chronologically consecutive fashion. Some contributing centers had ongoing research relating to hereditary cancer, which increased the prevalence of flat lesions in our study. The consequence of this was twofold: Most obviously, detection rates were reduced. Also, flat adenomas diminished our power to discriminate between groups because they present a challenge to all observers (28). However, they can be detected if observers are careful in their interpretation (28); for example, one experienced observer identified two flat adenomas. The proportion of flat adenomas should be reported in future studies of CT colonography.

Our study did have limitations. We originally intended that all participating centers would contribute studies obtained in 10 patients; however, not all centers did this. Three centers did not contribute any studies because they could not satisfy protocol stipulations.

Although the data set was designed to reflect what might be expected in a fecal occult blood test screening program, it was by necessity a simulation and can be regarded as a convenience sample. Assumptions for the bootstrap analysis best suit a random sample. For example, cancers detected at screening are in an earlier stage than those that are detected in patients who present with symptoms (10–12). Conversely, adenomas detected with the fecal occult blood test are larger than those in asymptomatic patients (10–12). We have discussed the difficulties posed by the proportion of flat adenomas.

Reading conditions were, by necessity, artificial. Image interpretation induces fatigue (6), and practitioners are currently unlikely to read 20 studies per day. However, this was a pragmatic necessity for this study, and this paradigm has been adopted successfully in other high-profile studies that have involved large numbers of observers from several centers (6). Our original intention was for observers to use their preferred software platform, but difficulties uploading studies prevented this. Instead, we assembled the data set onto laptop computers that could be transported easily to each center. These computers had high-resolution screens, and the software used a two-dimensional approach, with a three-dimensional approach available for problem solving; this was the preferred method of analy-

sis for the majority of experienced readers at the time of the study. All normal software functions were preserved on the computers. Because some readers had been trained to use another platform locally, we ensured that the software used in this study was easy to learn, and the study supervisor was available at all times to help, if necessary. While there is some evidence that the type of software platform used does not influence accuracy (6), it is possible that accuracy may have improved if a primary three-dimensional approach had been available (23). However, it should be stressed that we aimed to investigate the relative performance of observers and not the confounding effect of the software platform. Whether the reading platform used has a differential effect on experienced observers versus trained readers clearly merits further research. The use of laptop computers also meant that study loading times were longer than those of a workstation, and this may have frustrated some readers.

Investigators have examined the effect of implementing an identical training schedule for novice observers, with use of a teaching file and test set (29); however, we decided to leave the patient selection and training schedule largely to the discretion of the local trainer (beyond stipulations relating to the number of studies and length of training) because we thought this would better reflect current teaching practice. As a result, differences in performance potentially could be explained by variations in the quality of local training, which are precisely what occur in residency programs in general. For example, some trainers may have emphasized the importance of careful soft-tissue reading when looking for flat lesions, whereas other trainers may not have stressed this point. Whether an identical training scheme and materials administered via a training course are superior to more prolonged but less standardized local training is a subject that needs further investigation.

We have already stated that because our data set was relatively small, observed variability between readers may have been increased. Also, not all observers read the same studies to prevent recall bias due to interpretation of studies obtained at an observer's own center; however, we did balance the prevalence of abnormalities across all data sets so that they would remain comparable.

In conclusion, experienced observers asked to interpret CT colonographic studies performed significantly better on average than did novice observers who were trained with 50 endoscopically validated studies. However, individual performance is variable, and some trainees may outperform some experienced radiologists. On average, we found no performance difference between trained radiologists and trained radiographic technologists, which suggests that prior interpretation of conventional abdominal CT studies may not be of benefit for interpretation of CT colonographic studies.

Acknowledgments: The investigators are grateful to Voxar and E-Z-Em for providing workstations and CT colonography interpretation software.

European Society of Gastrointestinal and Abdominal Radiology (ESGAR) CT Colonography Study Group Investigators: Principal investigator: Steve Halligan, FRCR (University College London). Trial coordinator and data manager: David Burling, FRCR (St Mark's Hospital, London, England). Writing committee: Steve Halligan, FRCR; David Burling, FRCR, Wendy Atkin, PhD, and Clive Bartram, FRCR (St Mark's Hospital); Helen Fenlon, MD (Mater Misericordiae University Hospital, Dublin, Ireland); Andrea Laghi, MD (La Sapienza, Rome, Italy); and Jaap Stoker, MD (Amsterdam Medical Centre, Amsterdam, the Netherlands). Statisticians: Douglas G. Altman, DSc (Cancer Research UK/NHS Centre for Statistics in Medicine, Wolfson College, Oxford, England) and Paul Bassett, BSc (Statistical Consultant, Ruislip, Middlesex, England). ESGAR liaison: Roger Frost, FRCR (Salisbury NHS Trust, Salisbury, England). Study readers and local coordinators: Stuart Taylor, FRCR (University College London); Clive Bartram, FRCR, Lesley Honeyfield, DCR, and Melinda De Villiers, DCR (St Mark's Hospital); David Nicholson, FRCR, Velauthan Rudralingham, FRCR, and Lisa Renaut, DCR (Hope Hospital, Salford, England); Clive Kay, FRCR, Andy Lowe, FRCR, and Jane Williams-Butt, DCR (Royal Infirmary, Bradford, England); Jasper Florie, MD, and Martin Poulus (Academic Medical Center, Amsterdam, the Netherlands); Victor Van der Hulst, MD (Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands); Philippe Lefere, MD, Jesse Marrannes, and Guido Dessey, MD (Stedelijk Ziekenhuis, Roeselare, Belgium); Helen Fenlon, MD, Alan O'Hare, MD, and Shane Foley (Mater Misericordiae University Hospital, Dublin, Ireland); Emmanuele Neri, MD, Paola Vagli, MD, and Benedetta Politi (University of Pisa, Pisa, Italy); Riccardo Iannaccone, MD, Filipo Mangiapane, MD, and Sante Ori (La Sapienza, Rome, Italy); and Teresa Gallo, MD, Giulia Nieddu, MD, Saverio Signoretta, and Daniele Regge, MD (Candiolo Oncologic Hospital, Turin, Italy).

References

1. Kalish GM, Bhargavan M, Sunshine JH, Forman HP. Self referred whole-body imaging: where are we now? Radiology 2004;233:353–358.
2. Illes J, Fan E, Koenig BA, Raffin TA, Kann D, Atlas SW. Self-referred whole-body CT imaging: current implications for health care consumers. Radiology 2003;228:346–351.
3. Burling D, Halligan S, Taylor SA, Usiskin S, Bartram CI. CT colonography practice in the United Kingdom: a national survey. Clin Radiol 2004;59:39–43.
4. Halligan S, Taylor SA, Burling D. Virtual colonoscopy [letter]. JAMA 2004;292:432.
5. Ferrucci J, Barish M, Choi R, et al. Virtual colonoscopy [letter]. JAMA 2004;292:431–432.
6. Johnson CD, Toledano AY, Herman BA, et al. Computerized tomographic colonography: performance evaluation in a retrospective multicenter setting. Gastroenterology 2003;125:688–695.
7. Soto JA, Barish MA, Ferrucci JT. CT colonography interpretation: guidelines for training courses [abstr]. In: Radiological Society of North America scientific assembly and annual meeting program. Oak Brook, Ill: Radiological Society of North America, 2004; SSQ09-07.
8. Taylor SA, Halligan S, Burling D, et al. CT colonography: effect of experience and training on reader performance. Eur Radiol 2004;14:1025–1033.
9. Rockey DC, Paulson E, Niedzwiecki D, et al. Prospective comparison of colon imaging tests: a determination of the relative sensitivity of air contrast barium enema, computed tomographic colonography, and colonoscopy. Lancet 2005;365:305–311.
10. Hardcastle JD, Chamberlain JO, Robinson MH, et al. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. Lancet 1996;348:1472–1477.
11. Kronborg O, Fenger C, Olsen J, Jorgensen

OD, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. Lancet 1996;348:1467–1471.

12. Mandel JS, Bond JH, Church TR, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. N Engl J Med 1993;328:1365–1371.

13. Halligan S. Subspecialist radiology. Clin Radiol 2002;57:982–983.

14. Alderson PO. A balanced subspecialization strategy for radiology in the new millennium. AJR Am J Roentgenol 2000;175:7–8.

15. Capp MP. Subspecialization in radiology. AJR Am J Roentgenol 1990;155:451–454.

16. Halligan S, Marshall M, Taylor SA, et al. Observer variation in the detection of colorectal neoplasia on double contrast barium enema: implications for colorectal cancer screening and training. Clin Radiol 2003;58:948–954.

17. Rex DK, Vining D, Kopecky KK. An initial experience with screening for colon polyps using spiral CT with and without CT colography (virtual colonoscopy). Gastrointest Endosc 1999;50:309–313.

18. Cotton PB, Durkalski VL, Pineau BC, et al. Computed tomographic colonography (virtual colonoscopy): a multicenter comparison with standard colonoscopy for detection of colorectal neoplasia. JAMA 2004;291:1713–1719.

19. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224:861–869.

20. Johnson CD, Harmsen WS, Wilson LA, et al. Prospective blinded evaluation of computed tomographic colonography for screen detection of colorectal polyps. Gastroenterology 2003;125:311–319.

21. Culpan DG, Mitchell AJ, Hughes S, Nutman M, Chapman AH. Double contrast barium enema sensitivity: a comparison of studies by radiographers and radiologists. Clin Radiol 2002;57:604–607.

22. Brown L, Desai S. Cost-effectiveness of barium enemas performed by radiographers. Clin Radiol 2002;57:129–131.

23. Pickhardt PJ, Choi JR, Hwang I, et al. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. N Engl J Med 2003;349:2191–2200.

24. Sosna J, Morrin MM, Kruskal JB, Lavin PT, Rosen MP, Raptopoulos V. CT colonography of colorectal polyps: a metaanalysis. AJR Am J Roentgenol 2003;181:1593–1598.

25. Halligan S, Altman DG, Taylor SA, et al. CT colonography in the detection of colorectal polyps and cancer: systematic review, meta-analysis, and proposed minimum data set for study level reporting. Radiology 2005;237:893–904.

26. Rembacken BJ, Fujii T, Cairns A, et al. Flat and depressed colonic neoplasms: a prospective study of 1000 colonoscopies in the UK. Lancet 2000;355:1211–1214.

27. Suzuki N, Talbot IC, Saunders BP. The prevalence of small, flat, colorectal cancers in a Western population. Colorectal Dis 2004;6:15–20

28. Fidler JL, Johnson CD, MacCarty RL, Welch TJ, Hara AK, Harmsen WS. Detection of flat lesions in the colon with CT colonography. Abdom Imaging 2002;27:292–300.

29. Fidler JL, Fletcher JG, Johnson CD, et al. Understanding interpretative errors in radiologists learning computed tomography colonography. Acad Radiol 2004;11:750–756.