The British Journal of Radiology, 70 (1997), 1085–1098

Review article

Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image

P J A ROBINSON, FRCP, FRCR

Department of Clinical Radiology, St James's University Hospital, Leeds LS9 7TF, UK

Abstract. The performance of the human eye and brain has failed to keep pace with the enormous technical progress in the first full century of radiology. Errors and variations in interpretation now represent the weakest aspect of clinical imaging. Those interpretations which differ from the consensus view of a panel of "experts" may be regarded as errors; where experts fail to achieve consensus, differing reports are regarded as "observer variation". Errors arise from poor technique, failures of perception, lack of knowledge and misjudgments. Observer variation is substantial and should be taken into account when different diagnostic methods are compared; in many cases the difference between observers outweighs the difference between techniques. Strategies for reducing error include attention to viewing conditions, training of the observers, availability of previous films and relevant clinical data, dual or multiple reporting, standardization of terminology and report format, and assistance from computers. Digital acquisition and display will probably not affect observer variation but the performance of radiologists, as measured by receiver operating characteristic (ROC) analysis, may be improved by computer-directed search for specific image features. Other current developments show that where image features can be comprehensively described, computer analysis can replace the perception function of the observer, whilst the function of interpretation can in some cases be performed better by artificial neural networks. However, computer-assisted diagnosis is still in its infancy and complete replacement of the human observer is as yet a remote possibility.

Introduction

A century ago the doctors and scientists working with the then new-fangled Röntgen rays were predicting that they would soon be able to produce images of internal body structures using "a comparatively short exposure, say 5 or 10 min" [1]. Setting up the apparatus and the patient and developing the plates took considerably longer, so a whole day's work would typically produce only a handful of radiographs for interpretation. Even then, it soon became apparent that reading the images could give rise to error and to disputes. Mr W H Brown, speaking at the Leeds and West Riding Medico-Chirugical Society in 1901 on "Skiagraphy *versus* the Practitioner" referred to X-rays causing disputes between patients and their doctors [2], whilst in the same year Dr J B Hall pointed out to the Bradford Medico-Chirugical Society that with appropriate modifications of technique, a knowledgeable operator could produce in the skiagrams almost any degree of deformity he chose. He concluded that "X-rays could be dangerous in the hands of unscrupulous persons, and the profession should take steps to protect itself" [3].

Received 9 May 1997 and accepted 7 July 1997.

Growth in both the range and the volume of imaging procedures has accelerated enormously in the last two decades. It is now commonplace for staging computed tomography in an oncology patient to include 60 or 70 images, whilst a complex MR procedure may produce 200 or more images for interpretation. Simpler examinations such as chest radiography have been speeded up by filmless technology, considerably increasing the throughput of patients. The reporting radiologist is bombarded from all sides with increasing volumes of more and more complex images. It is now commonplace for the interpretion of clinical images to take longer than the process of acquiring them.

Guidelines published recently by professional bodies on both sides of the Atlantic recommend that all imaging procedures should include an expert opinion—by way of a written report or comment—from a radiologist [4, 5]. "Opinion" may be defined as "a conclusion arrived at after some weighing of evidence, but open to debate and suggestion" [6], so by implication the expert's report is not expected to be incontrovertible. Although technology has made enormous progress in the last century, there is no evidence for similar improvement in the performance of the human eye and brain. Recent analyses of errors in general

radiology [7] and in barium enema reporting [8] show that the same errors are being made now as in earlier decades. Moreover, every new development, every new technique, brings with it new opportunities for getting things wrong.

Distinguishing between error and variation

Error implies a mistake—in the context of image reporting, an incorrect interpretation. In order for a report to be erroneous, it follows that a "correct" report must also be possible. Because of the subjective nature of image interpretation, the definition of what is erroneous then becomes a matter of "expert opinion". The observer makes an error if he or she fails to reach the same conclusion that would be reached by a group of expert observers. Therefore, errors can only arise in cases where the correct interpretation is not in dispute. Degrees of error can exist, depending upon how clearly the abnormality is shown on the images or upon the extent, severity or size of the lesion.

Variation between observers may be due to error on the part of one observer, but also includes cases in which there is a genuine difference of opinion about the correct interpretation. A framework for describing the relationship between degrees of abnormality, doubt and certainty, error and variation, is shown in Figure 1. Here, the degree of severity, size or extent of the lesion is expressed on one axis with the probability that an abnormality is present described along a different axis. Cases in which major abnormalities can be recognized with a high degree of certainty (A) could be regarded as "easy" cases. In some cases a minor abnormality may be shown very clearly (B), for example, a tiny pleural effusion shown on CT, whilst in others the observer may have difficulty in deciding whether even a major abnormality is present (C), for example, an impacted fracture of femoral neck, or may have doubt about the significance of a clearly visible finding. Finally, there are cases in which minor degrees of abnormality are perceived with low levels of certainty (D) and, in practice, we tend to blur the distinction between these two attributes. For example, a report on a chest radiograph saying "the heart is slightly enlarged" literally means "yes, the heart is definitely enlarged but only to a slight degree". What we may intend to convey is that "there is a possibility that the heart is enlarged but this is by no means certain". It is with this latter type of case that differences of opinion must be regarded as acceptable, since neither of two conflicting reports would be regarded by the "weight of expert opinion" as being wrong. Somewhere between the clear-cut error and the inevitable difference of opinion is an arbitrary division defining the limit of professional acceptability. The position of this watershed between error and acceptable variation is defined by the performance standards of radiologists as expressed through their professional bodies and as interpreted by the courts in negligence and malpractice case law.

How widespread is observer variation?

The interpretation of radiographs is particularly accessible as a topic for studying observer variation. Unlike the physical examination of patients and findings at endoscopy or surgery, evidence of the examination remains available for subsequent scrutiny. Even so, there is a substantial body of literature covering observer variation in the taking of histories, clinical examination, endoscopy and the examination of pathology specimens. Within radiology, observer performance has been assessed in barium studies and other contrast procedures, in angiography, ultrasound, CT, MRI and radionuclide imaging. In the interests of brevity, this paper focuses primarily on the interpretation of the Röntgen image.

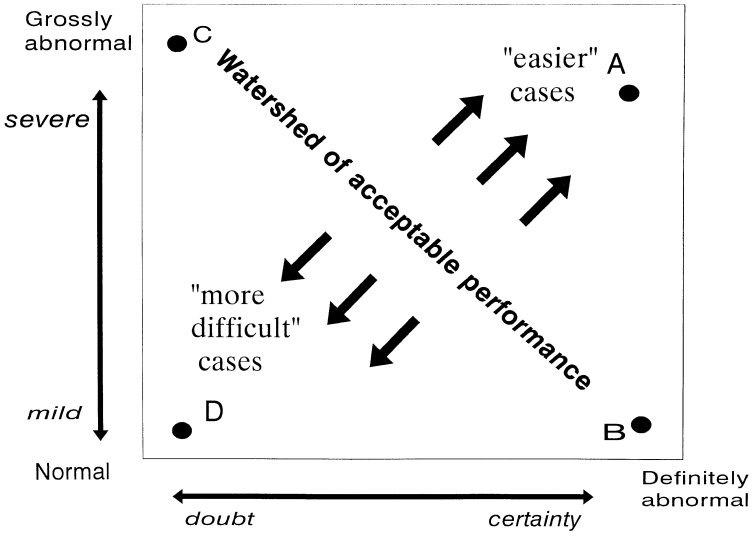

Figure 1. A proposed relationship between severity and certainty in recognizing abnormalities. "Errors" occur when "easy" cases (*e.g.* A) are wrongly interpreted; different readings of "difficult" cases (*e.g.* D) represent acceptable "observer variation". See text for more details.

Studies carried out 50 years ago showed that the chest radiographs of tuberculosis suspects were read differently by different observers in 10–20% of cases [9]. In the 1970s it was found that most lung cancers detected on screening radiographs were visible in retrospect on previous films [10, 11]. 10–30% of breast cancers are thought to be missed on screening mammograms [12, 13]. In reviewing barium enemas it was found that the "average observer" missed 30% of visible lesions [14].

Causes and consequences of error and variation

Common experience in radiology suggests that many errors are of little or no significance to the patient, and some significant errors remain undiscovered. Even so, errors are undesirable and professional audit should include error review [15, 16]. Risk management strategies aimed at reducing errors have recently highlighted requirements for "…development of and adherence to high standards of practice…" and "…insistence on technically satisfactory images…", avoiding excessive workload, ensuring the availability of previous examinations and using a suitable viewing environment [17].

As errors which come under the most intense scrutiny are those resulting in litigation, these cases should provide a useful illustration of the causes of error. In a recent review of 20 years' experience of medical litigation, Berlin [18] found that 70% of legal cases arising within radiology departments were due to alleged diagnostic errors. Berlin classified the causes of radiological "misses" as inadequate technique, perceptual errors, lack of knowledge and errors of judgement [19–22]. Analysing 437 errors detected in routine radiological practice in the 1960s, Smith [23] categorized about half as "under-reading" and a further quarter as failures of judgment in interpretation. In a more recent analysis of 182 radiological errors, Renfrew et al [7] found perceptual misses to be the most common problem. Other causes included limitations of technique, misleading or incomplete clinical data, unavailability of previous studies or reports, false positive errors (over-calls), misinterpretation of perceived findings, and misses due to the phenomenon of "satisfaction of search" in which subtle findings are more likely to be overlooked if overt abnormalities are also present [24]. Experimental studies recording the eye movements of radiologists searching chest radiographs for faint nodules found that about 30% of misses were due to incomplete scanning, 25% to failure of recognition, and 45% to wrong decisions [25]. The same group also showed that prolonged attention to a specific area on the radiograph ("visual dwell") increased both false negative and false positive errors [26]. However, reducing the viewing time for chest radiographs to less than 4 s also increased the miss rate [27].

Whilst individual errors are a cause of real concern in the care of the patients involved, a second major consequence of observer error and variation arises in the inappropriate assessment of new techniques. When there are several causes of error or variation in a system, the total error is equal to the square root of the sum of the squares of the component errors [28].

$$\text{Error}_{\text{total}} = \sqrt{(\text{error}_{\text{technique}})^2 + (\text{error}_{\text{observer}})^2}$$

As will be seen later, observer variation can be very considerable, so that comparative studies using only a single observer, or a single observer for each of two techniques, would need to show a very substantial difference between the two techniques to outweigh the probable effect of observer variation. Although the variability in observer performance was highlighted almost half a century ago [9], only in the last few years have studies on the clinical validity of new methods included multiple observers in their methodology [29].

The measurement of error and variation

In order to develop strategies for improving performance, we need to know how often errors and differences of opinion occur in different clinical contexts. In the simplest case, we can measure error by posing a set of test images and counting how often the observer calls the correct interpretation. This type of experiment allows us to measure the proportion of positive cases correctly called (sensitivity), the proportion of negative cases correctly called (specificity), and the proportion of correct calls in the whole sample (accuracy). The proportion of error is then (1 − accuracy). Expressing accuracy in this way does not distinguish between false positive and false negative calls. Even if sensitivity and specificity are used, this approach does not allow any weighting for the severity of individual errors and can only be applied when there is a clear-cut positive/negative interpretation, *i.e.* disease present or disease absent. With most biological systems there is overlap or gradation between normal and abnormal—a grey area—and in ordinary experience we would classify these cases as being "possibly abnormal" or perhaps "probably abnormal". Forcing a positive/negative response requires an arbitrary threshold which could differ between individual observers. Changing this decision threshold would alter sensitivity, specificity and accuracy.

An approach was developed to overcome this same problem in radar signal detection by allowing graded responses—degrees of certainty—rather than yes/no answers. Plotting the frequency of true positive calls against false positive calls at different levels of probability produces the receiver operating characteristic (ROC) curve, first applied to radiology by Lusted [30] and developed by Swets [31], Metz [32] and others. A perfect set of observations would produce a rectangular ROC curve encompassing an area of "1" when plotted in a unit square. The more accurate a set of observations, the larger the area under the ROC curve would be and the more it will approach unity. The measured area under the ROC curve (A_z) is a more resilient and reproducible indicator of accuracy than the simple sensitivities and specificities calculated from positive/negative calls, although the ROC method is more time-consuming and requires a degree of training in the observers. When used to compare the relative success of different imaging techniques both ROC and sensitivity/specificity methods share a major disadvantage—they test the observers as well as the techniques. Many radiological publications, even quite recently, have concluded erroneously that differences in results could be attributed to differences between techniques, when the effect of observer variation has been ignored.

A major limitation of enquiries into the effectiveness of clinical diagnostic methods is the difficulty in obtaining satisfactory verification of the diagnosis. Any measurement of accuracy requires comparison of the test method with a reference standard and the validity of the test results is dependent on the veracity of the reference standard. If no independent standard is available, as is often the case in clinical practice, the accuracy with which a cognitive task such as film reading is carried out might be approximated by comparison with a reference observer who is believed to be a "good performer". One way of doing this is to use the kappa statistic, described by Cohen et al [33], which is an indicator of the degree of agreement between two sets of observations of the same data. Kappa is defined as $\kappa=(p_o-p_c)/(1-p_c)$ where p_o is the observed level of agreement and p_c is the level of agreement attributable to chance. The level of agreement between observations can be applied either to a single observer interpreting the data on two separate occasions, or to two observers. Modifications of the kappa statistic allow weighting for the magnitude or significance of individual discrepancies [34] comparisons of multiple observers [35] and comparisons of multiple samples [36], but the applications and limitations of these statistical methods still need further exploration.

Reproducibility or accuracy?

Broadly, the ROC curve measures accuracy whilst kappa measures reproducibility. Clearly, an observer cannot sustain a high level of accuracy without being reproducible, so a low level of reproducibility is incompatible with a high level of accuracy. Since observer variation is largely idiosyncratic, it follows that convergence of multiple observers' opinions by expert consensus should provide a better reference standard than the use of a single expert, and the consensus method is now widely adopted in radiological evaluations. However, reproducible results are not necessarily accurate—all observers could agree on a finding and all of them could be wrong.

Factors influencing the magnitude of observer variation

In trying to understand the cause of error and variation we may enquire to what extent variations are intrinsic to individual observers, and to what extent they are determined by the nature of the task being performed. We could also ask how each of the potential causes of error: faulty technique, faulty perception, faulty knowledge and faulty judgment, impacts on the performance of various types of task. In scientific studies we would expect to use standard techniques and ensure that the observers had an adequate level of knowledge at the outset, unless these particular aspects were being tested. In studies using "expert observers" it is implicit that perception and judgment are the elements being tested. The tasks used in studies comparing observer variation in radiology generally fall into one of three categories: measurement, scoring or grading, and diagnosis.

Examples of measurement tasks include the derivation of Cobb angles in scoliosis [37–39] and measuring anatomical angles across joints [40–44]. With this type of task the technique for acquiring the radiographs is standard, the landmarks are clearly visible, and the method of measurement is well defined. The differences between observers should therefore be small. However, subtle changes may be less reproducibly measured than severe changes; Goldberg et al [45] found "excellent" agreement between observers evaluating primary Cobb angles in scoliosis, but much less agreement in measuring secondary Cobb angles because "…small angles (<20 degrees) were often not noticed." Furthermore, faulty technique in acquiring the images may undermine the validity of measurements which appear to be precise and reproducible. The measurement of varus angle at the knee in patients with bow-legs can be performed with precision, but changing the degree of rotation of the femur when the radiographs are

obtained produces differences which may be much greater than the measurement error [46]. The acetabular index, an indicator for hip dysplasia, was found to vary considerably if the degree of flexion/extension of the hip was not carefully reproduced at each successive examination [47] and may not be reliable as a single radiological measurement [48]. Nevertheless, many studies of this type of task show that the measurement error between different observers is fairly small and resistant to technical variables, for example, Robb et al [49] found the acetabular "tear drop" to be clearly visible on 93/100 routine pelvic radiographs, and were easily able to recognize coronal or sagittal tilt of 5° or more. Where intraobserver variation has also been measured, it is invariably less than the variation between observers, although this difference may not reach statistical significance if the measurement error itself is small.

The second type of task typically combines judgments on individual image features with numerical weighting, aimed at reducing opportunities for error. Examples include classifying fractures [50–53], grading the severity of a condition or a disease [40, 54], scoring or ranking studies [55, 56] and identifying the presence or severity of a condition by adding together points for different image features [57–61]. Conclusions from such studies vary enormously. Nance et al [60] found "…highly consistent and reproducible results…" between observers scoring the features of rheumatoid arthritis on hand radiographs. The recognition of specific features of bronchiolitis on chest radiographs showed "acceptable" agreement between observers in a study by Coblenz et al [61]. Thomsen et al [53] found "acceptable" levels of agreement for the overall classification of ankle fractures by four observers, but agreement was "poor" for staging of supination/adduction and supination/eversion. Herrs-Neilsen et al [62], using a scoring system for osteoporosis, found "satisfactory" interobserver agreement in diagnosing wedge compression fractures, but poor reproducibility for endplate infractions. Several studies found low levels of agreement between observers in the classification of fractures of the femoral neck or proximal humerus [50–52, 63] with one of them concluding that "neither…classification is sufficiently reproducible to allow meaningful comparison of different studies" [52]. In classifying the shape of the acromion, Haygood et al [64] found only moderate interobserver agreement for radiographs and poor agreement for MR images, questioning the value of the previous system of interpretation. Testing previously reported methods for determining the presence of loosening in hip prostheses, McCaskie et al [65] concluded that "…these methods cannot provide reliable data." Typical findings from this type of study show greater observer variation than with measurement studies, but again the variation within the same observer is less than that between different observers.

The third type of task requires the observer to diagnose the presence or absence of a condition without explicitly quantifying the individual features in the image which lead to the diagnostic conclusion—a "pattern recognition" task. Examples include the diagnosis of pneumonia [66], loosening of hip prostheses [67], osteopenia [68], spondylitis [69], and sacroiliitis [70, 71]. Not surprisingly, this type of study produces much more observer variation than studies of measurement reproducibility. Sargeant et al [58] found that in only seven out of 60 cases of subsequently confirmed intussusception was the diagnosis correctly made on abdominal radiographs by all of three observers. Shaw et al [72] found that three observers agreed on the interpretation of chest radiographs in children with malignant disease in only 29% of cases. Harris lines are widely used to establish the chronology of stress incidents during growth, but Macchiarelli et al [73] found poor interobserver reproducibility for both radiographs and pathological sections. Observer variation is much greater with some image features than with others, and with some diagnoses than with others. Assessing the loosening of hip prostheses, observers agreed about the femoral component much more often than about the acetabular component [67]. In chest radiographs of patients with suspected aortic injury, observers agreed more often on the presence of mediastinal widening and obscuration of the aortic knuckle than on the presence of several other signs [57]. Reviewing plain abdominal films, Markus et al [74] found some features to be unreliable whilst others were consistently reported. Also, it appears that observers agree more often in "easy" cases—advanced disease or gross classifications—than in "difficult" cases. In a multiobserver comparison between dual energy X-ray absorptiometry (DXA) and lumbar spine radiography, agreement was much better in cases of severe osteopenia confirmed by DXA than in mild cases [68]. Reviewing chest radiographs in pneumonoconiosis, Maranelli et al [75] found that the worst levels of observer agreement were associated with the most subtle cases. In a forensic dental study [76] 17 observers were asked to match up 31 pairs of dental radiographs and to describe each case as easy, moderate or difficult. Most of the errors were made on "difficult" cases. In the tibia vara study mentioned above [40], the level of agreement between observers was best for early and late stages and worst for intermediate stages of disease. Studies of lung scintigraphy for pulmonary embolism similarly show better levels of agreement for clear-cut high probability or normal cases

than for the intermediate or indeterminate categories. In these latter examples it is not the subtlety of disease which provides the difficulty but the arbitrary nature of the discrimination between different intermediate categories.

Review of these studies suggests two broad conclusions—firstly, that the variation between observers is always greater than the variation within a single individual's performances; and secondly, that the frequency of errors and the magnitude of observer variation both increase in proportion to the "difficulty" of the task. But what constitutes a "difficult" call? In some cases it is distinguishing between the normal range and early stages of disease—the borderlines of normality—whilst in other cases the presence of abnormality is not in doubt but criteria for indicating the severity or significance of the findings are highly subjective, and the reproducibility of judgment calls is poor. It now becomes clear that there is some circular logic here—the "easy" cases may be defined as those we usually agree on, whilst the "difficult" cases, by definition, are those in which expert opinion is divided.

Even when reproducible interpretations are obtained, they may have little relevance to clinical or pathological evidence of disease. Hand radiographs in rheumatoid arthritis were found to show highly reproducible abnormalities, but the findings were unrelated to clinical indicators of disease activity [77]. Reviewing radiographs of the sacro-iliac joints, Rothschild et al [71] found a high level of false positives and concluded that "… radiological techniques therefore have major limitations for the assessment of sacro-iliac disease…" Yao et al [78] found that even when there was agreement between observers in assessment of acromial shape, there was no correlation with the presence of impingement.

A final point is revealed by the language used to discuss findings and present conclusions. Studies which use pre-determined criteria for interpretation by different observers test both the methods and the observers. Where results are unsatisfactory it is typical for the authors to conclude that the methods are inadequate, rather than that there is anything wrong with the observers. This raises the question of the importance of training and experience.

The influence of training and experience

Most studies comparing the performance of different groups of observers have concluded that inexperienced or untrained observers are less successful in interpreting radiographs than are experienced radiologists. Such studies include comparisons of accident and emergency (A/E) doctors' readings of bone [79, 80] and abdominal radiographs [81], chest radiograph readings by various groups of physicians and medical students [54, 66], readings of A/E radiographs by non-medical observers [82, 83] and the radiographic identification of unknown human remains [84]. However, when the interpretations of radiologists have been compared with those of physicians or surgeons who were experienced in the context of the cases being examined, they have found little difference in performance. Clinical haematologists performed as well as radiologists in reading spine films of myeloma patients [85] and orthopaedic surgeons and radiologists produced concordant reports in trauma patients [86].

All these clinical studies share the limitation that there is no incontrovertible verification of the true reading of each case—no gold standard. Since training in medical specialties is largely a matter of aiming to achieve a level of performance indistinguishable from that of recognized experts in the field, convergence is inevitable. Improved reproducibility might indicate improved accuracy, but could also be explained by a convergence of the individual readings without a true improvement in accuracy. Throughout the history of medicine, new discoveries have usually been met by concerted hostility from established experts, and it is likely that at least some of our radiological teaching is based on illogical and ill-conceived judgments arising from inadequate knowledge. In the absence of conclusive proof of the presence or absence of disease in most of our test situations, the extent to which improved reproducibility simply means that we all make the same mistakes, rather than making different ones, remains uncertain.

Some light is cast on this question by studies in which nodules were artificially superimposed on normal chest radiographs, so that the "truth" was not in doubt, and by studies in which radiographs were interpreted either with or without the patients' clinical data. Reassuringly, the task of correctly determining the presence or absence of faint superimposed nodules on chest radiographs, mimicking the detection of early lung cancer, was performed more accurately by experienced radiologists than by trainees [87]. The availability of clinical data is generally regarded as being helpful in the interpretation of radiographs. Several authors have demonstrated improved performance when relevant clinical data were available to the reader [88–90] although others have argued that clinical details are unhelpful in lesion detection [91, 92]. The question of whether clinical clues help or hinder searches for specific abnormalities remains contentious. Some experimental evidence indicates that unguided search produces better results than guided search [93, 94], but these studies revolve around small differences in observer performance in the detection of very subtle lesions, tasks which

may by our previous definition be described as "difficult", so it may be inappropriate to apply their findings to routine clinical radiology. In another recent study [95], five experienced radiologists reviewed a sample of radiographs in which the diagnosis was "known", first without and then with the relevant clinical data. Availability of clinical data did not significantly improve accuracy but led to a substantial reduction in interobserver variation. When clinical data were available, the observers all tended to make errors on the same cases, rather than making different errors. This "distracting" effect of clinical data may be related to the phenomenon of "satisfaction of search" referred to earlier. The threshhold for detecting a specific abnormality is lowered by a prior suspicion of its likelihood, whilst the threshhold for detecting an unexpected lesion is raised, once another abnormality has been seen. In routine practice, we need clinical data to ensure firstly that the most appropriate procedure is carried out, and secondly to avoid spending time and effort searching for findings which would be irrelevant in the clinical context. These requirements heavily outweigh any potential advantage of eliminating bias in the observer by withholding relevant clinical data.

Circumstantial evidence that improved reproducibility includes at least an element of increased accuracy can be gained by scrutinizing the variation between inexpert observers. In one recent study, the variation between observers who were critical care physicians or anaesthetists was much greater than that between radiologists looking at the same chest radiographs [54]. In detecting fibrosing alveolitis on CT, inexperienced readers showed greater intraobserver variation than did experienced observers [96]. Whilst improved reproducibility does not necessarily equate to improved accuracy, low levels of reproducibility could not be associated with high levels of accuracy.

What then are the attributes of the expert observer which are learned through training and experience? Firstly, experimental data suggest that perception can be improved by disciplined strategies of visual search [97]. A second major element of learned experience is the ability to recognize new cases based on memories of previous experience. This is not just the gestalt phenomenon of recalling identical appearances, as in the recognition of faces, but also includes elements of analysis and synthesis so that visual features of similarity in images can be detected [98]. A third factor is the acquisition of new knowledge and the execution of those mental processes which make new knowledge available for recall in appropriate circumstances, so that the observer can approach new images with at least some knowledge of conditions which he or she has never seen before. The final and perhaps most critical element of "learned expertise" is the ability of the observer to understand the context of the diagnostic examination—to know what to look for in the images, and why [99].

Assistance from computers

The ubiquitous microchip is available to help us in almost everything we do. In the current context two questions arise—can computer acquisition or processing of the radiographic image contribute to improving the performance of the human observer, and can we expect computers to replace human observers in the analysis of image data?

Digital acquisition and/or display

Since the introduction of digital technology for acquiring X-ray images, conventional film has been superseded by digital hard copy images first in angiography, then in fluoroscopy, and most recently in "plain film" radiography. The take-up of digital acquisition technology is now delayed more by financial and cultural obstacles than by limitations of its performance. Numerous studies compared the detection of specific features and overall "image quality" of digitally acquired *versus* conventional analogue radiographs, while others have compared hard copy reporting with reading of "soft copy" images displayed on monitor screens. The typical findings were that low contrast structures are better seen on digital image presentations whereas fine structural detail, exemplified by the early changes of interstitial lung disease on chest radiographs, is better seen on conventional radiographs [100–104]. However, most of these studies used single groups of observers and are limited in that the training and experience of the radiologists used as observers was largely derived from analogue films, whereas the experiment tests both the digital technology and the observers' ability. In one of the few studies using multiple groups of observers, Kido et al [104] found that the difference between conventional chest radiographs and storage phosphor images was insignificant when the performance of a mixed group of experienced and inexperienced observers was considered. Subdividing their observers, it became clear that the experienced radiologists performed better with the analogue films whereas the trainees' performance was the same with both types of image—the difference between the observers was greater than the difference between the technologies. This factor was emphasized in another recent study in which observers' performance in viewing conventional radiographic images or light microscopy appearances was strongly correlated with years of

experience, whereas when digital displays of the same data were viewed by the same observers, performance was negatively correlated with experience [105], *i.e.* the inexperienced observers were better than the experts when the data were presented in an unfamiliar format. In some cases, therefore, digital acquisition of the radiographic image produces increased accuracy of detection, in some cases reduced accuracy, and in other cases no significant difference [106]. From first principles it seems unlikely that the introduction of digital radiography and soft copy displays for reporting will have much effect on the underlying causes of error which were described above. The obvious exception—that errors arising from incorrect exposure of radiographs will be much less likely with digital technology—is counterbalanced by the new opportunity for technical error which is inherent in the new methods.

Image enhancement and computer-assisted image analysis

Post-processing of digitized radiographs or digitally acquired image data offers some intriguing possibilities. Some early techniques focused on improving fine structural detail by spatial filtration for edge enhancement (*e.g.* unsharp masking) whilst others aimed at improving the visibility of low contrast objects by varying the available grey levels according to the degree of exposure in different areas of the image (*e.g.* histogram equalization). These methods often improved sensitivity at the expense of increased false positives. More recently, methods for direct analysis of image data by computers programmed to search for specific features, for example, microcalcifications in mammograms and nodular or interstitial densities on chest radiographs, have compared the performance of radiologists interpreting the images with and without computer assistance. Using ROC analysis, Chan et al [107] found a significant improvement in the accuracy of the observers' performance when computer analysis was available, whilst Kegelmeyer et al, using computer-assisted diagnosis (CAD) for detection of spiculated breast lesions, found an improvement in sensitivity with no loss in specificity [108]. Similar improvements have been found when radiologists interpreted chest images with CAD-generated analyses for detecting consolidation, nodules, or interstitial disease [109–111]. Using a variety of "rubber sheet warping" to create accurate registration of sequential chest radiographs, Defazio et al have recently shown that temporal subtraction produced a very significant improvement in the detectability of small or ill defined areas of consolidation [112].

Direct computer analysis of images

Clinical studies have so far concentrated on the analysis of image features which can be readily defined. Examples include measurement of the area of the hip joint space [113], metacarpal morphometry automated on digital images [114], bone age estimation from hand and wrist radiographs [115], and the diagnosis of hyperparathyroidism by detection of sub-periosteal erosions in hand radiographs [116]. Not surprisingly, all these studies confirm that automation of the decision making reduces the scope for variation until, with a totally automatic method, variation is abolished. Although absence of variation does not necessarily mean absence of error, and a limitation of these studies is that the reference standard for accuracy was derived from a group of expert observers, differences between the expert consensus and the computer analysis results were so small as to be probably insignificant.

Artificial intelligence

Perhaps the most exciting area of development is the application of artificial intelligence (AI) to the task of image interpretation. Various types of computer-based decision support are undergoing development [117]. Heuristic expert systems using rule-based reasoning are likely to be useful in guiding the appropriate selection of procedures but will probably have only a limited application in image interpretation. Bayesian networks offer a powerful way of representing knowledge about conditional probabilities and their interrelationships in complex clinical situations. Human performance in the assessment of conditional probabilities is known to be rather weak, so Bayesian systems may be most useful in this area, for example determining the relative likelihood of several possible causes of a solitary bone lesion. Case-based reasoning is a complex form of pattern recognition in which the knowledge base is made up of previous cases analysed in detail; this looks like a promising avenue for CAD with complex clinical problems but applications are, as yet, still in their infancy.

The most successful area at present is the use of artificial neural networks (ANNs). Unlike other CAD architectures, neural networks make up their own rules. Training of an ANN is typically by the direct input of image data from cases with established diagnoses. The subsequent processing of new (undiagnosed) cases by the ANN is driven by hidden layers of network elements ("neurons") and one of the disadvantages of ANNs is that the "thought processes" are covert—the explanation for a particular output decision cannot be traced. However ANNs have been shown to be effective

in the interpretation of chest radiographs in neonates [118] and in discriminating between benign and malignant lung nodules [119]. In distinguishing benign from malignant breast lesions an ANN has been found to out-perform expert radiologists by improving specificity whilst maintaining 100% sensitivity [120, 121]. At the time of writing, ANNs for the detection and differential diagnosis of bone lesions have been less successful but still sufficiently encouraging to promote further development.

Reducing observer error—the way forward

As we enter the second century of diagnostic radiology, it is clear that the weakest link in the chain of events which represents clinical imaging is the performance of the observer. A strategy for reducing observer error requires optimization of perception and of interpretation.

Viewing conditions

An understanding of human visual responses [122] suggests appropriate requirements for the physical environment of the reporting area, including luminosity of monitor screens or viewing boxes, ambient lighting levels, and varying viewing distances for detecting image features of different sizes and contrast. Testing of radiologists' visual acuity has been suggested [123], although it would be important to test perception of low contrast objects as well as the usual test of spatial resolution at high contrast.

Training

Many of the studies comparing the performance of observers who have different levels of experience actually rely on "expert consensus" as the final gold standard. It is therefore difficult to disentangle reproducibility from accuracy—what we may be seeing is convergence of observers rather than reduction in error. However, tests using simulated lesions have also shown improvement with experience, and the observation that reproducibility is worse in untrained observers also supports the idea that training is worthwhile. Because even expert observers show a wide variation in performance, an acceptable level of performance can be reached after only a moderate amount of training, even for observers with a non-medical background [124–126]. This applies only to perception and fairly simple categorization tasks—the more complex cognitive elements of reporting have not yet been studied in the same way.

Clinical data and previous films

Although experimental evidence suggests that the perception of low contrast objects is more accurately achieved by free search (*i.e.* with no prior clues as to the site and nature of the suspected lesion) rather than by directed search, the weight of evidence from clinical studies suggests that the availability of relevant clinical information is overwhelmingly more important. There is also good evidence that the availability of previous films (not just previous reports) improves diagnostic accuracy [127].

Dual reading

In the situation of screening for low-contrast objects, for example, early breast cancer on mammograms and early lung cancer on chest radiographs, error rates have varied between 10 and 50%. If these misses were truly random for each observer, dual reading should reduce the proportion of misses from $1/FN$ to $1/FN^2$ where FN is the percentage of false negative calls for each observer. Results obtained in practice largely bear this out. Dual reading of mammograms increases the number of lesions found by about 10–15% [128–130] whilst dual reading of chest radiographs was shown long ago to improve substantially the pick-up rate of pulmonary tuberculosis [131–133]. Dual reading of barium enemas also reduces perception errors [14, 134]. In a study of CT in lymphoma dual reading was found to have a greater impact on staging of disease than the use of contiguous rather than interval slices [135]. Evidence from screening mammography suggests that dual reading is also cost-effective [136]. Medical tradition accepts the "second opinion" as a useful resource in difficult cases, but what remains to be decided is the value of a second opinion in those cases which are not thought to be difficult. When considering the diagnostic efficacy of new techniques, future studies should include multiple observers in order to show how the costs and benefits of new technologies compare with the costs and benefits of dual or multiple observers. It may be more effective and cheaper to employ extra observers than to buy new technology.

Standardization

It would be an exceptional radiologist who could produce accurate bone age estimates without the use of a reference atlas, and using similar sets of reference images might reduce interobserver and intraobserver variation in other contexts, for example osteoarthritis of the knee [137]. Similarly, structured terminology such as the BI-RADS system for mammography reporting [138], allows easier comparison of complex reports and offers

an avenue towards eventual mechanization. So far, studies comparing the content of radiology reports have focused on specific descriptive and diagnostic elements. Nuances of probability, uncertainty, emphasis, significance and correlation with clinical data are more difficult to define. Attempts to capture the full richness of the information content of radiology reports in terms of symbolism, semiotics and semantics are still in their infancy [139]. Artificial intelligence techniques offer several possible routes for exploration.

Image processing and computer-assisted diagnosis

Post-processing techniques such as unsharp masking and histogram equalization may improve the conspicuity of subtle lesions but are unlikely to improve the performance of observers trained on conventional analogue images. The new generation of radiologists, trained with digital images, is more likely to benefit from computer assistance. More sophisticated CAD using search algorithms for specified image features have already been shown to improve accuracy of interpretation and the combination of human observer with computer assistance seems to be additive. The intelligent work station of the future will offer to the observer not only digital images with the capability of enhancing those features thought to be relevant to the clinical problem, but will also pre-process the images to offer diagnostic suggestions and will make available such reference data (other images of similar conditions, text descriptions, references etc.) as will help the observer to reach an accurate interpretation [140, 141].

Eliminating the observer

As pointed out by Jaffe [142] the observer contributes to radiographic interpretation only in those areas where the physical attributes of image abnormalities are not definable. Where the physical properties have been identified clearly enough to allow reliable detection and distinction from normal structures, the observer could be eliminated. Automated feature detection, combined with neural networks for decision making, can already improve on the performance of single human observers in breast cancer detection [120, 121, 143]. Further developments of these and similar approaches will surely replace human observers in some interpretive tasks. However, many radiological procedures produce complex images, and human perception is surprisingly adept at detecting low level signals in a noisy background. The integration of multiple image features, comparison with banks of similar images stored in memory, evaluation of interrelationships within the image data and with associated clinical data, and the other cognitive aspects of image interpretation are tasks at which AI devices have so far been less successful. Radiologists will not be replaced in the next decade or two, but could help themselves by contributing to the development of CAD and AI systems for assisting and improving their performance.

References

1. Anon. Lancet (8 Feb) 1896;i:1079.
2. Anon. Lancet (26 Jan) 1901;i:251.
3. Anon. Lancet (2 Feb) 1901;i:382.
4. American College of Radiology. ACR standard for communication: diagnostic radiology. Reston, VA: American College of Radiology, 1995.
5. Royal College of Radiologists. Statement on reporting in Departments of Clinical Radiology. London: RCR, 1995.
6. Longman dictionary of the English language. England: Longman, 1984.
7. Renfrew RL, Franken EA, Berbaum KS, Weigelt FH, Abu-Yousef MM. Error in radiology: classification and lessons in 182 cases presented at a problem case conference. Radiology 1992; 183:145–50.
8. Brady AP, Stevenson GW, Stevenson I. Colorectal cancer overlooked at barium enema examination and colonoscopy: continuing perceptual problem. Radiology 1994;192:373–8.
9. Garland LH. On the scientific evaluation of diagnostic procedures. Radiology 1949;52:309–28.
10. Heelan RT, Flehinger BJ, Melamed MR, Zaman MB, Perchick WB, Caravelli JF, et al. Non-small-cell lung cancer: results of the New York screening programme. Radiology 1984;151:289–95.
11. Muhm JR, Miller WE, Fontana RS, Sanderson DR, Uhlenhopp MA. Lung cancer detected using a screening program using four-month chest radiographs. Radiology 1983;148:609–15.
12. Baines CJ, Miller AB, Wall C, McFarlane DV, Simor IS, Jong R, et al. Sensitivity and specificity of first screen mammography in the Canadian National Breast Screening Study: a preliminary report from five centers. Radiology 1986;160:295–8.
13. Martin JE, Moskowitz M, Milbrath JR. Breast cancer missed by mammography. AJR 1979; 132:737–9.
14. Markus JB, Somers S, O'Malley BP, Stevenson GW. Double-contrast barium enema studies: effect of multiple reading on perception error. Radiology 1990;175:155–6.
15. Wheeler PS. Risk prevention, quality assurance, and the missed diagnosis conference. Radiology 1982;145:227–8.
16. Royal College of Radiologists Working Party. Medical audit in radiodiagnosis. London: RCR, 1989.
17. Royal College of Radiologists. Risk management in clinical radiology. London: RCR, 1995.
18. Berlin L, Berlin JW. Malpractice and radiologists in Cook County, IL: trends in 20 years of litigation. AJR 1995;165:781–8.
19. Berlin L. Malpractice Issues in Radiology. The importance of proper radiographic positioning and technique. AJR 1996;166:769–71.

20. Berlin L. Malpractice Issues in Radiology. Perceptual Errors. AJR 1996;167:587–90.
21. Berlin L. Malpractice Issues in Radiology. Errors in judgement. AJR 1996;166:1259–61.
22. Berlin L. Malpractice Issues in Radiology. Possessing ordinary knowledge. AJR 1996; 166:1027–9.
23. Smith MJ. Error and variation in diagnostic radiology. Springfield, IL: Charles C. Thomas, 1967:6.
24. Berbaum KS, Franken EA, Dorfman DD, et al. Satisfaction of search in diagnostic radiology. Invest Radiol 1990;25:133–40.
25. Kundel HL, Nodine CF, Carmody DP. Visual scanning, pattern recognition and decision making in pulmonary nodule detection. Invest Radiol 1978;13:175–81.
26. Kundel HL, Nodine CF, Krupinski EA. Visual dwell indicates locations of false-positive and false-negative decisions. Invest Radiol 1989;24:472–8.
27. Oestmann JW, Greene R, Kushner DC, Bourgouin PM, Linetsky L, Llewellyn HJ. Lung lesions: correlation between viewing time and detection. Radiology 1988;166:451–3.
28. Bevington PR. Data reduction and error analysis for the physical sciences. New York: McGraw Hill, 1969.
29. Obuchowski NA, Zepp RC. Simple steps for improving multiple-reader studies in radiology. AJR 1996;166:517–21.
30. Lusted LB. Signal detectability and medical decision-making. Science 1971;171:1217–9.
31. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979;14:109–21.
32. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720–33.
33. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
34. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
35. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82.
36. Donner A, Klar N. The statistical analysis of kappa statistics in multiple samples. J Clinl Epidemiol 1996;49:1053–8.
37. Pruijs JEH, Hageman MAPE, Keessen W, Van der Meer R, Van Wieringen JC. Variation in Cobb angle measurements in scoliosis. Skeletal Radiol 1994;23:517–20.
38. Peterson MD, Nelson LM, McManus AC, Jackson RP. The effect of operative position on lumbar lordosis: A radiographic study of patients under anaesthesia in the prone and 90–90 positions. Spine 1995;20:1419–24.
39. Loder RT, Urquhart A, Steen H, Graziano G, Hensinger RN, Schlesinger A, et al. Variability in Cobb angle measurements in children with congenital scoliosis. J Bone Joint Surg (B) 1995;77:768–70.
40. Stricker SJ, Edwards PM, Tidwell MA. Langenskiold classification of tibia vara: An assessment of interobserver variability. J Paediat Orthopaed 1994;14:152–5.
41. Seelen JL, Bruijn JD, Hansen BE, Kingma LM, Bloem JL. Reproducible radiographs of acetabular prostheses. A method assessed in 35 patients. Acta Orthop Scand 1994;65:258–62.
42. Saltzman CL, Brandser EA, Berbaum KS, DeGnore L, Holmes JR, Katcherian DA, et al. Reliability of standard foot radiographic measurements. Foot Ankle Int 1994;15:661–5.
43. Resch S, Ryd L, Stenstrom A, Johnsson K, Reynisson K. Measuring hallus valgus: A comparison of conventional radiography and clinical parameters with regard to measurement accuracy. Foot Ankle Int 1995;16:267–70.
44. Tallroth K, Ylikoski M, Landtman M, Santavirta S. Reliability of radiographical measurements of spondylolisthesis and extension–flexion radiographs of the lumbar spine. Eur J Radiol 1994;18:227–31.
45. Goldberg MS, Poitras B, Mayo NE, Labelle H, Bourassa R, Cloutier R. Observer variation in assessing spinal curvature and skeletal development in adolescent idiopathic scoliosis. Spine 1988;13:1371–7.
46. Stricker SJ, Faustgen JP. Radiographic measurement of bowleg deformity: Variability due to method and limb rotation. J Paediat Orthopaed 1994;14:147–51.
47. Portinaro NMA, Murray DW, Bhullar TPS, Benson MKD. Errors in measurement of acetabular index. J Paediat Orthopaed 1995;15:780–4.
48. Broughton NS, Brougham DI, Cole WG, Menelaus MB. Reliability of radiological measurements in the assessment of the child's hip. J Bone Joint Surg (B) 1989;71:6–8.
49. Robb JE, Rymaszewski LA, Bentley HB, Donnan PT. Reliability of the acetabular teardrop as a landmark. Surg Radiol Anat 1991;13:181–5.
50. Kristiansen B, Andersen ULS, Olsen CA, Varmarken JE. The Neer classification of fractures of the proximal humerus. An assessment of interobserver variation. Skel Radiol 1988;17:420–2.
51. Sidor ML, Zuckerman JD, Lyon T, Koval K, Cuomo F, Schoenberg N. The Neer classification system for proximal humeral fractures. An assessment of interobserver reliability and intraobserver reproducibility. J Bone Joint Surg (A) 1993;75:1745–50.
52. Siebenock JA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. J Bone Joint Surg (A) 1993;75:1751–5.
53. Thomsen NOB, Overgaard S, Olsen LH, Hansen H, Nielsen ST. Observer variation in the radiographic classification of ankle fractures. J Bone Joint Surg (B) 1991;73:676–8.
54. Beards SC, Jackson A, Hunt L, Wood Frerk CM, Brear G, et al. Interobserver variation in the chest radiograph component of the lung injury score. Anaesthesia 1995;50:928–32.
55. Warren-Forward HM, Millar JS. Assessment of image quality for chest radiography in the West Midlands. Radiat Prot Dosim 1995;57:171–4.
56. Vehmas T, Tikkenen H, Bondestam S, Holmberg G, Kivisaari A, Knuutila T, et al. Observed lung markings in normal chest roentgenograms. Rofo Fortschr 1993;159:50–3.
57. Burney RE, Gundry SR, Mackenzie JR, Whitehouse WM, Wu SC. Chest roentgenograms in diagnosis of traumatic rupture of the aorta: observer variation of interpretation. Chest 1984;85:605–9.
58. Sargent MA, Babyn P, Alton DJ. Plain abdominal radiography in suspected intussusception: A reassessment. Paediat Radiol 1994;24:17–20.
59. O'Sullivan MM, Lewis PA, Newcombe RG, et al. Precision of Larsen grading of radiographs in assessing progression of rheumatoid arthritis in individual patients. Ann Rheum Dis 1990;49: 286–9.

60. Nance EP Jr, Kaye JJ, Callahan LF, Carroll FE, Winfield AC, Earthman WJ, et al. Observer variation in quantitative assessment of rheumatoid arthritis. Part I: scoring erosions and joint space narrowing. Invest Radiol 1986;21:922–7.
61. Coblentz CL, Babcook CJ, Alton D, Riley BJ, Norman G. Observer variation in detecting the radiologic features associated with bronchiolitis. Invest Radiol 1991;26:115–8.
62. Herrs-Nielsen VA, Podenphant J, Martens S, Gotfredsen A, Juel-Riis B. Precision in assessment of osteoporosis from spine radiographs. Eur J Radiol 1991;13:11–4.
63. Frandsen PA, Andersen E, Madsen F, Skjodt T. Garden's classification of femoral neck fractures: an assessment of inter-observer variation. J Bone Joint Surg (B) 1988;70:588–90.
64. Haygood TM, Langlotz CP, Kneeland JB, Iannotti JB, Williams GR Jr, Dalinka MK. Categorization of acromial shape: interobserver variability with MR imaging and conventional radiology. AJR 1994;162:1377–82.
65. McCaskie AW, Brown AR, Thompson JR, Gregg PJ. Radiological evaluation of the interfaces after cemented total hip replacement: Interobserver and intraobserver agreement. J Bone Joint Surg (B) 1996;78:191–4.
66. Young M, Marrie TJ. Interobserver variability in the interpretation of chest roentgenograms of patients with possible pneumonia. Arch Int Med 1994;154:2729–32.
67. Kramhoft M, Gehrchen PM, Bodtker S, Wagner A, Jensen F. Inter and intraobserver study of radiographic assessment of cemented total hip arthroplasties. J Arthroplasty 1996;11:272–6.
68. Jergas M, Uffmann M, Escher H, Gluer CC, Young KC, Grampp S, et al. Interobserver variation in the detection of osteopenia by radiography and comparison with dual X-ray absorptiometry of the lumbar spine. Skeletal Radiol 1994;23:195–9.
69. Davies AM, Fowler J, Tyrrell PNM, Millar JS, Leahy JF, et al. Detection of significant abnormalities on lumbar spine radiographs. Br J Radiol 1993;66:37–43.
70. Yazici H, Turunc M, Ozdogan H, Yurdakul S, Akinci A, Barnes CG. Observer variation in grading sacroiliac radiographs might be a cause of "sacroiliitis" reported in certain disease states. Ann Rheum Dis 1987;46:139–45.
71. Rothschild BM, Poteat GB, Williams E, Crawford WL. Inflammatory sacroiliac joint pathology: evaluation of radiologic assessment techniques. Clin Exp Rheumatol 1994;12:267–74.
72. Shaw NJ, Hendry M, Eden OB. Interobserver variation in interpretation of chest X-rays. Scott Med J 1990;35:140–1.
73. Macchiarelli R, Bondioli L, Censi L, Hernaez MK, Salvadei L, Sperduti A. Intra- and interobserver concordance in scoring Harris lines: a test on bone sections and radiographs. Am J Phys Anthropol 1994;95:77–83.
74. Markus JB, Somers S, Franic M, Moola C, Stevenson GW. Interobserver variation in the interpretation of abdominal radiographs. Radiology 1989;171:69–71.
75. Maranelli G, Lovisatti L, Gaffuri E. Interobserver variations of chest radiograph readings for pneumoconiosis. Med Lav 1988;79:187–93.
76. Ekstrom G, Johnsson T, Borman H. Accuracy among dentists experienced in forensic odontology in establishing identity. J Forensic Odont Stomatol 1993;11:45–52.
77. Jaeckel WH, Cziske R, Kuehn T, Schulz H, Jacobi E. Validity and objectivity of radiographic parameters as outcome criteria in rheumatoid arthritis. Z Rheumatol 1990;49:151–4.
78. Yao L, Lee H-Y, Gentili A, Shapiro MM. Lateral downsloping of the acromion: a useful MR sign? Clin Radiol 1996;51:869–72.
79. Vincent CA, Driscoll PA, Audley RJ, Grant DS. Accuracy of detection of radiographic abnormalities by junior doctors. Arch Emerg Med 1988;5:101–8.
80. Wardrope J, Chennels PM. Should all casualty radiographs be reviewed? Br Med J 1985; 290:1638–40.
81. Suh RS, Maglinte DDT, Lavonas EJ, Kelvin FM. Emergency abdominal radiography: discrepancies of preliminary and final interpretation and management relevance. Emerg Radiol 1995;2:315–8.
82. Renwick I, Butt WP, Steele B. How well can radiographers triage X-ray films in Accident and Emergency Departments. Br Med J 1991;302:568–9.
83. Loughran CF. Reporting of accident and emergency radiographs by radiographers: A study to determine the effectiveness of a training programme. Br J Radiol (Congress Suppl) 1994;67:93.
84. Hogge JP, Messmer JM, Doan QN. Radiographic identification of unknown human remains and interpreter experience level. J Forensic Sci 1994;39:373–7.
85. Browman GP, Markman S, Thompson G, Minuk T, Chirawatkul A, Roberts RS. Assessment of observer variation in measuring the radiographic vertebral index in patients with multiple myeloma. J Clin Epidemiol 1990;43:833–40.
86. Turen CH, Mark JB, Bozman R. Comparative analysis of radiographic interpretation of orthopedic films: is there redundancy? J Trauma Inj Infect Crit Care 1995;39:720–1.
87. Christensen EE, Murry RC, Holland K, Reynolds J, Landay MJ, Moore JG. The effect of search time on perception. Radiology 138;1981:361–5.
88. Aideyan UO, Berbaum K, Smith WL. Influence of prior radiologic information on the interpretation of radiographic examinations. Acad Radiol 1995;2:205–8.
89. Doubilet P, Herman PG. Interpretation of radiographs: Effect of clinical history. AJR 1981; 137:1055–8.
90. Schreiber MH. The clinical history as a factor in roentgenogram interpretation. JAMA 1963; 185:399–401.
91. Good BC, Cooperstein LA, DeMarino GB, Miketic LM, Gennari RC, Rockette HE, et al. Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation? AJR 1990; 154:709–12.
92. Eldevik OP, Dugstad G, Orrison WW, Haughton VM. The effect of clinical bias on the interpretation of myelography and spinal computed tomography. Radiology 1982;145:85–9.
93. Swensson RG, Hessel SJ, Herman PG. Radiographic interpretation with and without search: visual search aids the recognition of chest pathology. Invest Radiol 1982;17:145–51.
94. Swensson RG, Hessel SJ, Herman PG. The value of searching films without specific preconceptions. Invest Radiol 1985;20:100–7.

95. Tudor GR, Finlay D, Taub N. An assessment of inter-observer agreement and accuracy when reporting plain radiographs. Clin Radiol 1997;52:235–8.
96. Collins CD, Wells AU, Hansell DM, Morgan RA, McSweeney JE, DuBois RM, et al. Observer variation in pattern type and extent of disease in fibrosing alveolitis on thin section computed tomography and chest radiography. Clin Radiol 1994;49:236–40.
97. Kundel HL, LaFollette PS Jr. Visual search patterns and experience with radiological images. Radiology 1972;103:523–8.
98. Neisser U. Cognitive psychology. New York: Appleton–Century–Crofts, 1967.
99. Robinson PJ. The nature of image reporting. In: Paterson A, Price R, editors. Current topics in radiography. London: W B Saunders, 1996:70–82.
100. Krupinski EA, Maloney K, Bessen SC, Capp MP, Graham K, Hunt R, et al. Receiver operating characteristic evaluation of computer display of adult portable chest radiographs. Invest Radiol 1994;29:141–6.
101. Dobbins JT III, Rice JJ, Beam CA, Ravin CE. Threshold perception performance with computed and screen–film radiography: implications for chest radiography. Radiology 1992;183:179–87.
102. Dawood RM, Craig JOMC, Todd-Pokropek A, Porter AW, Highman JH, Cunningham DA, et al. Clinical diagnosis from digital displays: results and conclusions from the St Mary's evaluation project. Br J Radiol 1994;67:1–10.
103. Baker JA, Floyd CE, Lo JY, Ravin CE. Observer evaluation of scatter subtraction for digital portable chest radiographs. Invest Radiol 1993;28:667–70.
104. Kido S, Ikezoe J, Takeuchi N, Kondoh H, Tomiyama M, Jokoh T, et al. Interpretation of subtle interstitial lung abnormalities: Conventional *versus* storage phosphor radiography. Radiology 1993;187:527–33.
105. Krupinski EA, Weinstein RS, Rozek LS. Experience related differences in diagnosis from medical images displayed on monitors. Telemed J 1996;2:101–8.
106. Van Heesewijk HPM, Van der Graaf Y, De Valois JC, Vos JA, Feldberg MAM. Chest imaging with a selenium detector versus conventional film radiography: a CT-controlled study. Radiology 1996;200:687–90.
107. Chan HP, Doi K, Vyborny CJ, Schmidt RA, Metz CA, Lam KL. Improvement of radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990;25:1102–10.
108. Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW, Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994;191:331–7.
109. Difazio MC, MacMahon H, Xu X, Doi K. Effect of time interval difference images on detection accuracy in digital chest radiology. Radiology 1994;193:288.
110. Kobayashi T, Xu X-W, MacMahon H, Metz CE, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology 1996;199:843–8.
111. Katsuragawa S, Doi K, Yoshida H, MacMahon H, Ishida T. Texture analysis based on wavelet transform for detection and characterisation of interstitial diseases on chest radiographs. Radiology 1995;197(P):292.
112. Difiazio MC, MacMahon H, Xin-Wei Xu BS, Tsai P, Shiraishi J, Armato SG III, et al. Digital chest radiography: effect of temporal subtraction images on detection accuracy. Radiology 1997;202:447–452.
113. Conrozier T, Tron AM, Balblanc JC, Mathieu P, Piperno M, Fitoussi G, et al. Measurement of the hip joint space using computerized image analysis. Rev Rheum Eng Ed 1993;60:105–11.
114. Kalla AA, Meyers OL, Parkyn ND, Kotze TjvW. Osteoporosis screening: radiogrammetry revisited. Br J Rheumatol 1989;28:511–7.
115. Cox LA. Preliminary report on the validation of a grammar-based computer system for assessing skeletal maturity with the Tanner–Whitehouse 2 method. Acta Paediat/Int J Paediat (Suppl) 1994;83:84–5.
116. Chang CL, Chan HP, Niklason LT, Cobby M, Crabbe J, Adler RS. Computer-aided diagnosis: detection and characterization of hyperparathyroidism in digital hand radiographs. Med Phys 1993;20:983–92.
117. Kahn CE Jr. Decision aids in radiology. Rad Clin N Am 1996;34:607–28.
118. Gross GW, Boone JM, Greco-Hunt V, Greenberg B. Neural networks in radiologic diagnosis: II. Interpretation of neonatal chest radiographs. Invest Radiol 1990;25:1017–23.
119. Gurney JW, Swenson SJ. Solitary pulmonary nodules: determining the likelihood of malignancy with neural network analysis. Radiology 1995; 196:823–9.
120. Baker JA, Kornguth PJ, Lo JY, Floyd CE Jr. Artificial neural network: improving the quality of breast biopsy recommendations. Radiology 1996;198:131–5.
121. Floyd CE, Lo JY, Yun AJ, Sullivan DC, Kornguth PJ. Prediction of breast cancer malignancy using an artificial neural network. Cancer 1994; 74:2944–8.
122. Campbell FW. The neurosciences, third study programme. London: MIT Press, 1975.
123. Straub WH, Gur D, Good BC. Visual acuity testing of radiologists—is it time? AJR 1991;156:1107–8.
124. Robinson PJ. Plain film reporting by radiographers—a feasibility study. Br J Radiol 1996;69:1171–4.
125. Pauli R, Hammond S, Cooke J, Ansell J. Radiographers as film readers in screening mammography: an assessment of competence under test and screening conditions. Br J Radiol 1996;69:10–14.
126. Alcorn FS, O'Donnell E, Ackerman LV. The protocol and results of training nonradiologists to scan mammograms. Radiology 1971;99:523–9.
127. White K, Berbaum K, Smith WL. The role of previous radiographs and reports in the interpretation of current radiographs. Invest Radiol 1994;29:263–5.
128. Bird RE. Professional quality assurance for mammography screening programmes. Radiology 1990;177:587.
129. Brenner RJ. Medicolegal aspects of breast imaging: variable standards of care relating to different types of practice. AJR 1991;156:719–23.
130. Anderson EDC, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. Clin Radiol 49;1994:248–51.
131. Yerushalmy J, Harkness JT, Kennedy BR. The role of dual reading in mass radiography. Am Rev Tuberculosis 1950;61:443–64.

132. Stradling P, Johnston RN. Reducing observer error in a 70 mm chest radiography service for general practitioners. Lancet 1955;i:1247–50.
133. Hessel SJ, Herman PG, Swensson RG. Improving performance by multiple interpretations of chest radiographs: effectiveness and cost. Radiology 1978;127:589–94.
134. Ott DJ, Gelfand DW, Ramquist NA. Causes of error in gastrointestinal radiology: II barium enema examination. Gastrointest Radiol 1980;5:99–105.
135. Naik KS, Spencer JA, Craven C, McClellan K, Robinson PJ. Computed tomography in staging lymphoma: comparison of contiguous with interval 10 mm slices. Proc American Roentgen Ray Society, Boston, May 1997.
136. Seradour B, Wait S, Jacquemier J, Dubuc M. Dual reading in a non-specialized breast cancer screening programme. Breast 1996;5:398–403.
137. Scott WW Jr, Lethbridge-Cejku M, Reichle R, Wigley FM, Tobin JD, Hoshberg MC. Reliability of grading scales for individual radiographic features of osteoarthritis of the knee: the Baltimore longitudinal study of aging atlas of knee osteoarthritis. Invest Radiol 1993;28:497–501.
138. American College of Radiology. Breast imaging reporting and data system (BI-RADS). Reston, VA: American College of Radiology, 1993.
139. Robinson PJ, Fletcher JM. Clinical coding in radiology. Imaging 1994;6:133–42.
140. Lou SL, Huang HK, Arenson RL. Workstation design. Radiol Clin N Am 1996;34:525–44.
141. Rogers E. A blackboard-based system for diagnostic radiology. Artif Intell Med 1995;7:343–60.
142. Jaffe CC: quoted in Kundel HL, Hendee WR. The perception of radiologic image information: report of an NCI workshop on April 15–16, 1985. Invest Radiol 1985;20:874–7.
143. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CA. Application of neural networks in mammography: applications in decision making in the diagnosis of breast cancer. Radiology 1993;187:81–7.