Annals of the Royal College of Surgeons of England (1992) *vol.* 74, 40–46

How does computer-aided diagnosis improve the management of acute abdominal pain?

J Wellwood MChir FRCS
Consultant Surgeon

S Johannessen
Research Assistant

Whipps Cross Hospital, Leytonstone, London E11 1NR

D J Spiegelhalter PhD
Statistician

MRC Biostatistics Unit, Cambridge

Key words: Decision making, computer aided; Abdominal pain; Abdomen, acute

The introduction of standardised data-collection forms and computer-aided diagnosis has been found to be associated with improved diagnosis and management of patients with acute abdominal pain. The mechanism by which such benefits accrue has been the subject of some controversy. Detailed analysis of 5193 patients from one hospital shows that the major benefit from such diagnostic aids was the accurate early diagnosis of non-specific abdominal pain by senior house officers in the accident and emergency department; this in turn led to fewer admissions and fewer operations with negative findings. Clinical data about patients with acute abdominal pain should be recorded on structured information sheets by junior doctors and early positive diagnosis should be encouraged before decisions affecting the patient's management are made. Improved computer support may confer further benefits.

Computer-assisted diagnosis of acute abdominal pain has a long history *(1–4)*, and the extent of current interest is reflected in a major European initiative *(5)*. A multicentre trial of a computer-aided diagnostic system has been reported *(6,7)* in which data were pooled from eight hospitals, involving over 250 doctors and 16 737 patients. More than 30% of these data were collected at Whipps Cross Hospital, whose results reflected those of the overall study *(6)*. In view of repeated expressions of doubt as to the mechanism by which the system produced the observed clinical benefits *(8–12)*, we have reanalysed the Whipps Cross data in order to determine the process by which the diagnostic aids affected clinical performance.

Correspondence to: Mr James Wellwood MChir FRCS, Consultant Surgeon, Medical Education Centre, Whipps Cross Hospital, Leytonstone, London E11 1NR

Methods

Trial design

The project was carried out in the accident and emergency department in which eight senior house officers (SHOs) were on 6-month placements. Baseline data were collected prospectively during the first 5 months when the performance of all eight SHOs working with no diagnostic aids was recorded without their knowledge (1021 patients). For the subsequent 24 months of the study, SHOs were randomly allocated to the use of different diagnostic aids using the 'biased coin' technique *(13)* in order to balance sex and career intention as far as possible in groups. For 6 months (1182 patients) the trial compared four SHOs using standardised data collection forms with four who had no diagnostic aids ('controls'). During the subsequent 12 months of the study (1991 patients) four of the SHOs were provided with standardised data collection forms only and four were additionally provided with the computer. During the final 6 months of the study (999 patients) all eight SHOs were

Table I. Accuracy, reliability and confidence of computer diagnosis

Diagnostic accuracy of computer

	Discharge diagnosis of patient data entered on computer			
	Appendicitis (n = 113)	*NSAP (n = 820)*	*Other on computer (n = 145)*	*Other not on computer (n = 158)*
% in which computer diagnosis matches discharge diagnosis	73%	77%	59%	0%

Reliability and confidence of computer diagnosis

	Top diagnosis given by computer		
	Appendicitis (n = 180)	*NSAP (n = 765)*	*Other (n = 291)*
% in which computer diagnosis matches discharge diagnosis	46%	82%	38%
Mean probability given by computer to top diagnosis	87%	88%	86%

provided with standardised data collection forms, the computer and monthly analyses of their performance. Thus a total of 40 accident and emergency SHOs were studied in the whole trial, which included 5193 patients.

All patients with acute abdominal pain were seen by SHOs in the accident and emergency department. Data collection forms required the SHO to state an initial diagnosis. Doctors who had been allocated to use the computer either entered the data themselves or gave the structured information sheet to the project assistant (SJ) who entered the data at the first available opportunity. The computer printout, which listed nine diagnoses and assigned a percentage probability to each, was not always ready by the time the final diagnosis was made before admission or discharge of the patient.

The final or discharge diagnosis was obtained after surgery or at discharge. For descriptive purposes all diagnoses have been compressed into three groups: appendicitis, non-specific abdominal pain (NSAP) and all other organic causes.

Statistical methods

The diagnostic and predictive accuracy of the computer and the SHOs has been tabulated for each of the modes of decision support provided (none, forms alone, forms + computer, forms + computer + feedback). Variation between doctors has been taken into account by carrying out nested analysis of variance *(14)* when making comparisons between different types of aid provided. Admissions and operations have been analysed with respect to the true disease of the patient and the initial diagnostic opinion of the SHO. Statistical significance in comparing proportions was assessed by χ^2 tests; total numbers in tables do not always exactly match due to occasional missing data.

Results

Patient population

There were 5193 patients in the Whipps Cross trial. Appendicitis was the final diagnosis in 324 (6%) of the patients and NSAP was the final diagnosis in 3612 (70%). A further 494 patients had other diagnoses which were included on the computer's main option, but 763 (15% of all patients) had a diagnosis which was not on the main computer program and which therefore could not have been diagnosed correctly by the computer. The proportion of patients with each diagnosis showed no evidence of varying between modes of decision support ($P = 0.35$).

Compliance: how often were the forms and computer used?

Forms should have been completed for 3583 patients and were actually completed in 2330 (65%). The computer was used by the SHO in 981 (50%) of the 1980 cases for which it had been allocated and in 772 (39%) its opinion was available to the SHO before the initial diagnosis was made.

Diagnostic accuracy, reliability and confidence of the computer

Table I compares the diagnostic accuracy and reliability of the computer for patients with acute appendicitis,

NSAP, and 'other' diagnoses. The diagnostic accuracy of the computer is the percentage of all computer diagnoses that match the final discharge diagnosis. The diagnostic accuracy for the 113 patients who had acute appendicitis was 73%. The similar figure was 77% for NSAP and 59% for other causes of abdominal pain on the computer. Overall accuracy on all cases was 64%.

An alternative means of viewing the computer diagnosis is by assessing predictive accuracy, or reliability, which is the proportion of assertions (shown here in italics) that turn out to be correct *(15)*. In Table I it can be seen that for 180 diagnoses of *appendicitis* claimed by the computer the predictive accuracy was subsequently found to be only 46%. The predictive accuracy when claiming *NSAP* was 82% and when asserting any other specific diagnosis was 38%. It can be seen, therefore, that when the computer made any diagnosis other than *NSAP* its reliability was $<50\%$.

Table I also shows the means of the probabilities given by the computer to its top diagnosis for each diagnostic group. The computer was 'very certain' of its diagnosis of acute appendicitis (mean probability 87%), but was subsequently found to be correct in only 82 (46%) of cases. Figure 1 further emphasises this over-confidence of the computer system. A total of 1236 computer opinions have been grouped according to the probability given to the top diagnosis, and plotted against the frequency with which that diagnosis was actually correct.

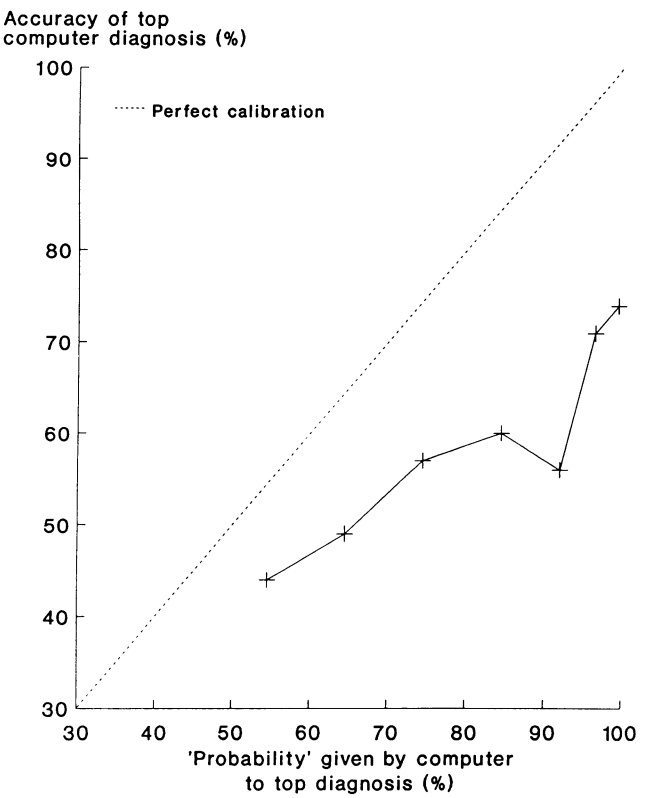

Figure 1. The reliability of the 'probabilities' provided by the computer: for example, the computer diagnosis was correct in only 65 of 116 cases (56%) for which the computer had given its diagnosis a 'probability' of between 90% and 95%.

This 'calibration plot' *(16,17)* shows that any diagnosis given less than 95% probability is correct in only 50–60% of cases. In the extreme, 433 patients (35% of all computer diagnoses) received a probability of 99% for their top computer diagnosis; however, only 74% of these diagnoses were correct.

Effect of (a) the provision and (b) the use of standardised data collection forms

Figure 2 contrasts the initial diagnostic accuracy (%) of SHOs who were given no diagnostic aids with that of SHOs provided with standardised data collection forms both when the latter were available but not used, and when they were available and also used. It is clear that the SHOs benefited when the forms were available even when they were not used. However, the increased consistency of the performance by the SHOs when they used the forms is clearly shown.

How do the diagnostic aids increase initial diagnostic accuracy?

Figure 3 shows the diagnostic accuracy of the SHOs receiving different diagnostic support, distinguishing incorrect diagnoses from diagnoses not given: there is no particular reduction in the production of incorrect diagnoses made. Between 70% and 80% of patients with appendicitis were correctly diagnosed by the SHO with or without diagnostic aids and the observed rise with the use of decision aids is not statistically significant ($P=0.66$) when correction is made for changing identity of the SHOs. There was, however, a significant improvement ($P<0.001$) in the proportion of patients with NSAP who were correctly diagnosed instead of being given no diagnosis at all. Overall, the initial accuracy rose from 48% for those SHOs provided with no decision aids, to 69% for those with forms, computer and performance feedback. However, four-fifths of this improvement was due to NSAPs being correctly diagnosed, in place of no diagnosis being given.

The accuracy of the computer diagnosis for each diagnostic group is also shown in Fig. 3 (data from Table I). This shows that the SHOs were slightly better than the computer in the diagnosis of appendicitis, and approached the computer's accuracy in the diagnosis of patients with NSAP.

The reliability of the diagnoses made by the SHOs is plotted in Fig. 4, which shows the number of occasions when a diagnosis claimed by the SHO proved susbsequently to be correct. The computer reliabilities are also plotted (*see* Table I). There was only a small increase in the reliability of judgements made by SHOs when diagnostic aids were available. Furthermore, the reliability of any diagnosis made by either the SHOs or the computer was low with the exception of the diagnosis of NSAP. Any other diagnosis made, whether by SHO or computer, had less than a 50% chance of being correct when compared with the final diagnosis.

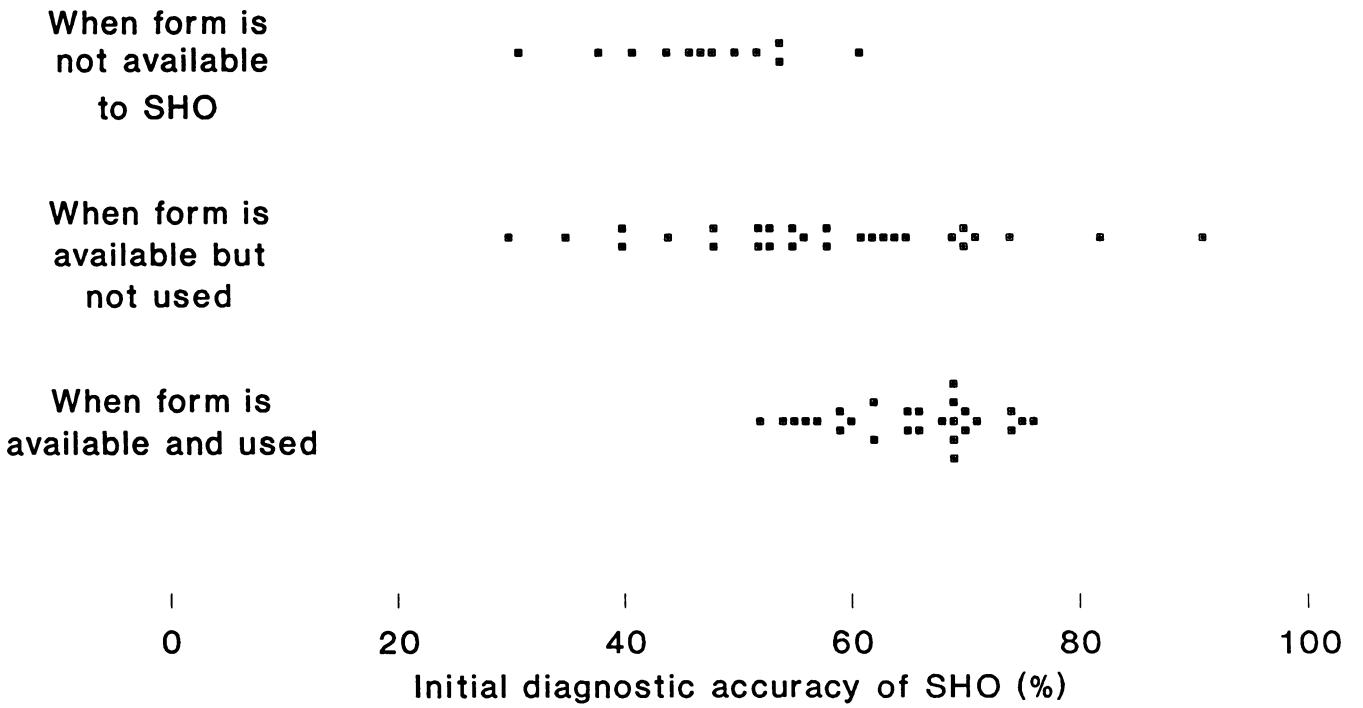

Figure 2. The initial diagnostic accuracy of each SHO for whom standardised data collection forms were *not* available and for each SHO for whom they were available. The accuracy for this latter group is given separately for the occasions when the available forms were not used and for the occasions when they were used.

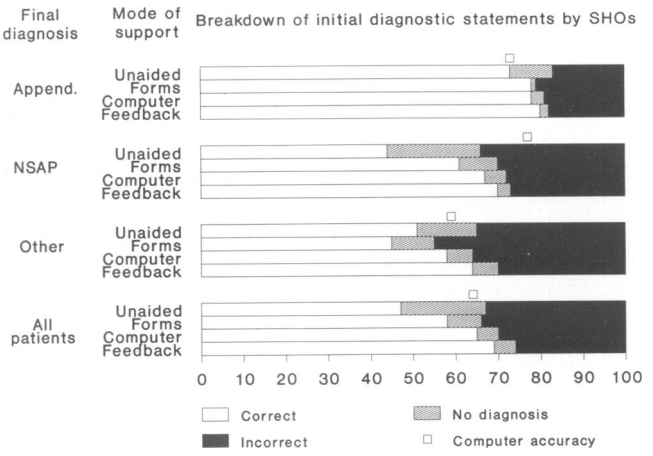

Figure 3. The proportion of cases in which the SHO's initial diagnosis was correct, not given or incorrect. Data are shown for each final diagnostic category and for each mode of decision support available. ('Feedback' indicates forms, computer and monthly feedback of the doctors' performance were all available.)

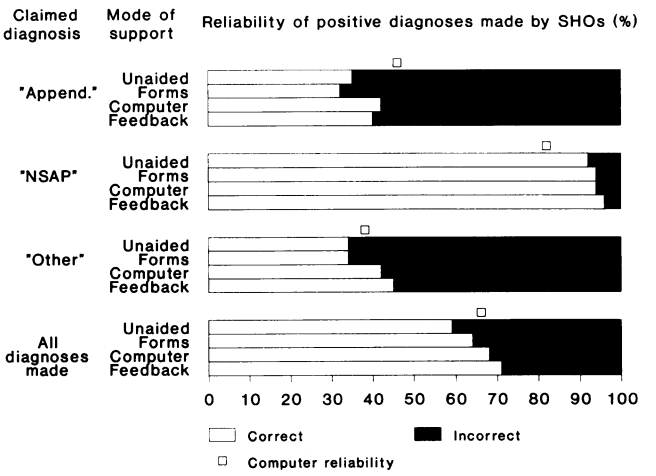

Figure 4. The proportion of initial diagnoses made by the SHOs which were correct. Data are shown for each claimed diagnosis and for each mode of decision support.

Alterations in management

Admissions

Of all the patients in the study, 39% (2021/5193) were admitted to hospital. Row (a) of Table II shows a significant fall in the proportion of patients who were admitted when diagnostic aids were used: without diagnostic aids 43% of patients were admitted but using the structured forms, computer and monthly feedback of performance the proportion of patients who were admitted fell to 34% ($P<0.001$). Rows (b) and (c) of Table II show the patients divided into those with NSAP and all others. The proportion of all NSAP patients who were admitted when no diagnostic aids had been provided was 29% and this fell to 16% with the use of forms, the computer and monthly feedback of performance. This fall is highly significant ($P<0.0001$). The proportion of patients with all other diagnoses who were admitted did not change significantly with the use of diagnostic aids. The fall in admission rates with the use of diagnostic aids can therefore be attributed wholly to the admission of fewer patients with NSAP.

Greater clarification is provided by further dividing the NSAP group of patients in row (c) into those patients

Table II. Analysis of patients admitted

	Diagnostic aids provided				
	None	*Forms*	*Forms + computer*	*Computer + feedback*	
(a) Patients admitted	43% (689/1610)	38% (610/1598)	39% (380/986)	34% (342/999)	($P<0.001$)
(b) All non-NSAP patients	75% (369/495)	79% (359/453)	78% (250/319)	74% (232/314)	($P=0.19$)
(c) All NSAP patients	29% (320/1115)	29% (251/1145)	20% (130/667)	16% (110/685)	($P<0.001$)
(d) All NSAP patients diagnosed *correctly* by SHO	12% (57/495)	8% (57/697)	9% (39/443)	8% (37/476)	($P=0.15$)
(e) All NSAP patients *not diagnosed* by SHO	42% (263/620)	43% (194/448)	41% (91/224)	35% (73/209)	($P=0.20$)

whose NSAP had been diagnosed correctly by the SHO in the accident and emergency department (row (d)) and those patients with NSAP whose diagnosis had not been made in the accident and emergency department (row (e)). The number of patients admitted when the SHO had diagnosed NSAP correctly in the accident and emergency department is fairly constant, varying between 12% and 8%, while admission rates for those patients not diagnosed correctly by the SHO in the accident and emergency department is again fairly constant, but varied between 42% and 35%. The highly significant fall in the proportion of all NSAP patients who were admitted is therefore almost wholly attributable to more correct initial diagnoses made by the SHO in the accident and emergency department. Of all NSAP patients who were either undiagnosed or incorrectly diagnosed by the SHO, 41% were subsequently admitted, whereas only 9% of those correctly identified by the SHO were admitted.

Surgical operations

There was no significant reduction in the total number of surgical operations undertaken when diagnostic aids were introduced (Table III, row (a)). However, when we divide patients into NSAP and others (rows (b) and (c)), we find a clear fall in the proportion of patients with NSAP who underwent surgical operations from 3.4% to 1.8%. However, if we express the operations in row (c) *as a proportion of all NSAP patients admitted* (row (d)), we find there is no significant drop. Thus, we have identified that the reduction in surgical operations was brought

Table III. Proportion of patients undergoing surgical operation

	Diagnostic aids provided				
	None	*Forms*	*Forms + computer*	*Computer + feedback*	
(a) Patients undergoing surgical operation	14% (224/1610)	14% (217/1598)	15% (146/986)	12% (121/999)	($P=0.36$)
(b) Proportion of non-NSAP patients undergoing surgery	38% (186/495)	43% (193/453)	43% (136/319)	35% (104/314)	($P=0.08$)
(c) Proportion of all NSAP patients undergoing surgery	3.4% (38/1115)	2.1% (24/1145)	1.5% (10/667)	1.8% (12/685)	($P=0.03$)
(d) Proportion of admitted NSAP patients undergoing surgery	12% (38/320)	9% (24/251)	8% (10/130)	11% (12/110)	($P=0.59$)
(e) Laparotomies with negative findings	23% (30/130)	19% (24/124)	14% (12/85)	17% (12/72)	($P=0.39$)

about by fewer patients with NSAP being admitted to the hospital, and this resulted from better initial diagnosis by the SHOs in the accident and emergency department. Row (e) of Table III shows that this brought about a reduction in laparotomies that were negative from 23% to 17%, although the figures do not achieve statistical significance.

Discussion

The design of the trial of computer-aided diagnosis at Whipps Cross Hospital permitted a detailed examination of the mechanism by which the benefits from a computer-aided diagnostic system could occur. The system is intended as a general educational tool and the random allocation of doctors to modes of decision support appeared the correct experimental design. The compliance of the SHOs was reasonably good (65% of forms filled in: the SHO using the computer himself in 50% of cases).

Although the diagnostic accuracy was reasonable, the reliability of any specific claimed diagnosis was poor. If anything other than NSAP was claimed there was less than 50% chance of being right. This would not be so unsettling if the computer could recognise when it is unsure and give a reasonable assessment of its 'doubt'. However, 'probability' statements were over-confident, and for diagnoses when the computer stated it was essentially certain (99%) it was wrong in over 25%. There are two possible reasons for this: firstly, the incidence rates assumed in the program do not match those at Whipps Cross, and using the 'correct' incidences would increase the 'probability' given to NSAP, and decrease that to appendicitis. Since the majority of diagnoses made were NSAP, this adjustment could make the over-confidence even worse. The second reason is the unwarranted assumption of the independence of symptoms within disease, and this is rectifiable by only slightly more sophisticated statistical modelling using standard logistic regression analysis (*16,17*).

The accuracy of the initial diagnosis made by the SHOs improved dramatically during the study and we have shown that this was largely due to more rather than better diagnoses being made, particularly on patients with NSAP. The standard data collection form encouraged the making and entering of an initial diagnosis. The SHOs' performance had similar characteristics to the computer, with poor reliability especially when claiming a diagnosis of one of the less common diseases. Most doctors provided with decision aids did better when they used forms than when they did not and the overall impression is that using the forms tended to narrow down the inevitable disparities between SHOs. This pattern should be expected since early diagnosis of patients with NSAP (70% of all patients in the Whipps Cross study) brought SHOs up to a common standard.

A number of clinicians (*8–12*) have been mystified as to how improvements in initial diagnostic accuracy by SHOs resulted in changes in admission and operation rates. We feel this has now been explained by revealing the crucial importance of obtaining the correct initial diagnosis, particularly of NSAP. A patient with NSAP who was not initially correctly diagnosed had a fivefold higher chance of subsequent admission than one correctly diagnosed, regardless of decision support available. Thus, more early diagnosis immediately lowers admission rates of NSAPs and, hence, lowers negative operation rates.

The popular image of how computer-aided diagnosis confers benefit appears to be that of a junior clinician making an incorrect diagnosis, consulting the computer and obtaining a correct diagnosis, and then revising his opinion appropriately. Our investigation revealed this scenario occurred in only 22 of the 1985 occasions in which the computer was consulted. Rather, the reason why patients benefited after the introduction of diagnostic aids was probably related to the increased pressure on the SHOs in the accident and emergency department to make a diagnosis and the self-audit engendered by the computer system. The routine use of structured data collection sheets for the recording of details of patients with acute abdominal pain should be seriously considered throughout the NHS, while improvements in computer hardware and software, which we understand have now been developed, may increase the attractiveness of the use of the computer in the accident and emergency department and thereby increase the benefits detected in this study.

References

1 de Dombal FT, Leaper DJ, Staniland JR *et al.* Computer-aided diagnosis of acute abdominal pain. *Br Med J* 1972; **ii**:9–13.

2 Gunn AA. The diagnosis of acute abdominal pain with computer analysis. *J R Coll Surg Edinb* 1976;**21**:170–2.

3 Fenyo G. Routine use of a scoring system for decision-making in suspected acute appendicitis in adult. *Acta Chir Scand* 1987;**153**:545–51.

4 McAdam WAF, Brock BM, Armitage T, Davenport P, de Dombal FT. Twelve years' experience of computer-aided diagnosis in a district general hospital. *Ann R Coll Surg Engl* 1990;**72**:140–46.

5 de Dombal FT. The European Community concerted action on objective medical decision making in patients with acute abdominal pain. A preliminary report. *Theor Surg* 1990;**5**:112–17.

6 Adams ID, Chan M, Clifford PC *et al.* Computer aided diagnosis of acute abdominal pain: a multicentre study. *Br Med J* 1986;**293**:800–804.

7 Clifford PC, Chan M, Hewett DJ. The acute abdomen: management with microcomputer aid. *Ann R Coll Surg Engl* 1986;**68**:182–4.

8 Hall PN. Letter. *Br Med J* 1986;**293**:1025.

9 Paterson-Brown S, Eckersley JRT, Dudley HAF. Letter. *Br Med J* 1986;**293**:1025–6.

10 Driscoll P, Vincent C. Letter. *Br Med J* 1986;**294**:56.

11 Sutton GC. How accurate is computer aided diagnosis? *Lancet* 1989;**2**:905–8.

12 Sutton GC. Computer-aided diagnosis: a review. *Br J Surg* 1989;**76**:82–5.

13 White SJ, Freedman LS. Allocation of patients to treatment groups in a controlled clinical study. *Br J Cancer* 1978;**37**:849–57.
14 Snedecor GW, Cochran WG. *Statistical Methods*, 7th Edition. Iowa: Iowa State University Press, 1980.
15 Galen RS, Gambino SR. *Beyond Normality: the Predictive Value and Efficiency of Medical Diagnoses*. New York: Wiley, 1975.
16 Spiegelhalter DJ. Statistical methodology for evaluating gastrointestinal symptoms. *Clin Gastroenterol* 1985;**14**:489–515.
17 Seymour DG, Green M, Vaz FG. Making better decisions: construction of clinical scoring systems by the Spiegelhalter–Knill-Jones approach. *Br Med J* 1990;**300**:223–6.

Received 15 April 1991

Notes on books

Surgery of the Shoulder edited by M Post, B F Morrey and R J Hawkins. 386 pages, illustrated, Mosby Yearbook, St Louis. 1990. £67.50. ISBN 0 8151 6745 8

The Proceedings of a Symposium held in New York City in 1989. 84 contributions, each illustrated but somewhat sparsely referenced. Folio format, art paper, spacious layout.

Atlas of Adult Cardiac Surgery by William A Gay Jr. 189 pages, illustrated. Churchill Livingstone, New York. 1990. £87.50. ISBN 0 443 08598 6

Full-page clear illustrations on the right and brief supporting text on the left, combine to make the pages of this new atlas attractive to the eye as well as imparting much craft knowledge.

Decision Making in Surgical Sepsis edited by R L Nichols, N E Hyslop and J G Bartlett. 389 pages, illustrated. B C Dekker Inc., Philadelphia. 1991. £40.00. ISBN 1 55664 053 6

This book is unusual in that it consists mainly of algorithms. There are nineteen sections each containing numerous chapters on every aspect of surgical sepsis. On the right-hand side of the open page is an algorithm showing the decision-making process that should be adopted in the management of the particular problem, while the left-hand page gives a succinct supporting text. Clear and elegant layout.

Current Practice of Surgical Stapling edited by M M Ravitch, F M Steichen and R Welter. 324 pages, illustrated. Lea & Febiger, Philadelphia. 1991. £51.00. ISBN 0 8121 1328 4

The Proceedings of an International Symposium on stapling held in Luxembourg in 1988. Many well-known names in the field contribute to the 51 presentations. The early papers relate to history, principles and instrumentation. The later chapters relate to individual operations in the chest and abdomen.

Surgical Pathophysiology edited by A O Aasen and Bo Risberg. 452 pages, illustrated. Harwood Academic Publishers, Chur. 1990. $95.00. ISBN 3 7186 5038 X

Endotoxins, free oxygen radicals, ARDS, complement activation, intestinal peptides and multiple-organ failure are just some of the many subjects covered in this authoritative book from Scandinavia.

Complications of Cancer Management by P N Plowman, T J McElwain and Anna T Meadows. 504 pages, illustrated. Butterworth-Heinemann, Oxford. 1991. £98.00. ISBN 0 7506 1341 6

Most of the chapters relate to the morbidity that may be associated with radiotherapy or chemotherapy in the treatment of various types of malignant disease. There is a small section on the complications of surgery. Extensively illustrated and referenced, this book will serve as a useful reference volume and will prove of interest to all who practise oncology, be it medical or surgical.

Clinical and Experimental Approaches to Dermal and Epidermal Repair edited by A Barbul and others. 497 pages, illustrated. Wiley Liss, New York. 1991. No price given. ISBN 0 471 56075 8

The Proceedings of a Symposium on tissue repair held in Florida in 1990. Surgeons interested in the physiology and pathology of wound healing and tissue repair will find this book of interest.

Constipation in Childhood by Graham Clayden and Ulfur Agnarsson. 123 pages, illustrated, paperback. Oxford University Press, Oxford. 1991. £9.50. ISBN 0 19 262027 4

Although written by two paediatricians there is much in this book to interest those surgeons who from time to time see children with chronic constipation. Until now very little information has been readily available on this most common of problems. This slim volume written in an easy style and containing much common sense deserves wide circulation and sale.

Recent Advances in the Epidemiology and Prevention of Gallstone Disease edited by L Capocaccia, G Ricci, F Angelico, M Angelico, A F Attili and L Lalloni. 165 pages, illustrated. Kluwer Academic Publishers, Dordrecht. 1991. £46.50. ISBN 0 7923 0994 4

The Proceedings of an International Workshop held in Rome in 1989. Twenty-two contributions mainly from Italy and the United States summarise the non-surgical aspects of gallstone diease.