Contemporary Themes . . .

Computer aided diagnosis of acute abdominal pain: a multicentre study

I D ADAMS, M CHAN, P C CLIFFORD, W M COOKE, V DALLOS, F T DE DOMBAL, M H EDWARDS, D M HANCOCK, D J HEWETT, N McINTYRE, P G SOMERVILLE, D J SPIEGELHALTER, J WELLWOOD, D H WILSON

Abstract

A multicentre study of computer aided diagnosis for patients with acute abdominal pain was performed in eight centres with over 250 participating doctors and 16 737 patients. Performance in diagnosis and decision making was compared over two periods: a test period (when a small computer system was provided to aid diagnosis) and a baseline period (before the system was installed). The two periods were well matched for type of case and rate of accrual.

The system proved reliable and was used in 75·1% of possible cases. User reaction was broadly favourable. During the test period improvements were noted in diagnosis, decision making, and patient outcome. Initial diagnostic accuracy rose from 45·6% to 65·3%. The negative laparotomy rate fell by almost half, as did the perforation rate among patients with appendicitis (from 23·7% to 11·5%). The bad management error rate fell from 0·9% to 0·2%, and the observed mortality fell by 22·0%. The savings made were estimated as amounting to 278 laparotomies and 8516 bed nights during the trial period—equivalent throughout the National Health Service to annual savings in resources worth over £20m and direct cost savings of over £5m.

Computer aided diagnosis is a useful system for improving diagnosis and encouraging better clinical practice.

St James's University Hospital, Leeds LS9 7TF
I D ADAMS, MD, consultant in charge, accident and emergency department
M CHAN, AFIMA, project officer
F T DE DOMBAL, MD, FRCS, reader in clinical information science

Royal Hampshire County Hospital, Winchester
P C CLIFFORD, MD, FRCS, senior surgical registrar
D J HEWETT, FFCM, MBCS, district medical officer

Middlesbrough General Hospital, Middlesbrough
W M COOKE, FRCS, consultant surgeon

Whipps Cross Hospital, Leytonstone, London E11 1NR
V DALLOS, FRCP, consultant in charge, accident and emergency department
J WELLWOOD, FRCS, consultant surgeon

Friarage Hospital, Northallerton, North Yorkshire
M H EDWARDS, FRCS, consultant surgeon

Sunderland District General Hospital, Sunderland
D M HANCOCK, FRCS, consultant surgeon

Royal Free Hospital, London NW3 2QG
N McINTYRE, MD, FRCP, professor of medicine

Royal Sussex County Hospital, Brighton
P G SOMERVILLE, FRCS, consultant surgeon

MRC Biostatistics Unit, Cambridge
D J SPIEGELHALTER, PHD, statistician

General Infirmary, Leeds
D H WILSON, FRCS, consultant in charge, accident and emergency department

Correspondence to: Dr de Dombal.

Introduction

Several studies have already been performed using the acute abdominal pain program first reported in 1972 from Leeds.[1] Each has shown an improvement in diagnostic accuracy (McAdam, unpublished report),[2-4] and it has been claimed that such systems would have considerable impact if introduced more widely. This multicentre trial was therefore set up to test the hypothesis that the acute abdominal pain program could be transferred to various types of hospital; that it could be used by doctors with no previous experience of microcomputers; and that clinical and financial benefit would result.

Patients and methods

STRUCTURE OF TRIAL

The project was carried out in eight participating centres under the direction of a project leader (usually a consultant surgeon) in each centre. Data analysis was undertaken by the clinical information science group of the University of Leeds. The eight centres where a computer system was installed and the effects assessed were chosen to represent a range of National Health Service activity, ranging geographically from the north east to the south coast and including both urban and rural populations.

TRIAL DESIGN

In four (mode A) centres we tried to replicate earlier experiments at Airedale and Leeds. In these hospitals data from a baseline period of about one year (immediately preceding installation of the system or a prototype) were compared with data from a test period of two years.

In the remaining four (mode B) centres the purpose of the experiment was more complex; we also wanted to assess the various contributions made by data collection forms, computer analysis, and feedback of performance to individual doctors. Therefore the 112 doctors concerned were divided into four groups. One group used structured data collection forms; the second used forms and personal computers; the third used forms and received feedback about their performance but did not use personal computers; and the fourth group used forms and personal computers and received feedback.

PATIENTS

During the baseline period 4075 cases were studied and in the test period 12 662 cases were studied: a total of 16 737 patients suffering from acute abdominal pain of less than one week's duration at presentation to hospital was seen (table I). The age and sex distribution was similar in the various centres, as was the breakdown by diagnostic category. These variables, together with the accrual rate, were similar also in each centre during baseline and test periods.

TABLE I—*Total numbers of patients studied in the trial broken down by trial period and hospital*

Centre or hospital	Patients in baseline period	Patients in test period	Total patients
Brighton	0*	525	525
Leeds (General Infirmary)	600	2759	3359
Leeds (St James's Hospital)	657	2796	3453
Middlesbrough	305	619	924
Northallerton	297	638	935
Sunderland	310	823	1133
Whipps Cross	1611	3583	5194
Winchester	295	919	1214
Total	4075	12662	16737

*Due to lack of histopathological evidence in baseline period.

METHODS OF COMPUTER USE

An attempt was made to use a common method of operating the computer aided system.[1] For the purposes of data gathering by individual doctors a specific data collection form was provided. Each computer (most commonly Commodore PET 8032 or Apple IIe) was provided with software based on programs already developed by the Leeds group,[5] information from each new patient being compared (via a probabilistic analysis using Bayes's theorem) with data on 6000 patients from 13 countries.[6]

EVALUATION

Each patient presenting with acute abdominal pain was followed until discharge from hospital or death.[1] Details concerning diagnostic predictions and patient progress were forwarded to the coordinating centre for analysis using the Amdhal VM7 computer of the University of Leeds with the SPSS/X statistical package. Each set of patient data was checked independently by at least two people. When differences in interpretation were noted the data were returned to the project leader in the relevant centre for review. In addition, range, logic, and random data checks were instituted.

The evaluation of aids for making clinical decisions has been the subject of lively debate.[7] The statistical issues raised in this study are discussed in a separate footnote. Many categories for evaluation are self explanatory; some, however, warrant additional clarification.

Initial diagnostic accuracy measured the diagnostic accuracy of the first doctor to see each case (house surgeon in four centres, senior house officer in the rest). An accurate diagnosis occurred when this initial diagnosis matched the final (discharge) diagnosis, according to criteria decided in advance by consensus view of the project leaders. When no diagnosis at all was made this was counted as an error.

Accuracy after investigation compared (using the same criteria) the diagnosis arrived at by the surgical team after investigation and consultation with the final diagnosis.

Bad diagnostic error was a concept introduced to reflect the reality that not all errors were of equal gravity. A bad diagnostic error was an initial diagnosis which suggested a non-surgical condition (or no initial diagnosis at all) in a patient whose condition was eventually found to warrant emergency surgery.

Bad management error was an actual decision which delayed necessary surgery for 24 hours (or otherwise placed the patient's life at risk).

Perforated appendix rate measured the proportion of patients with genuine appendicitis whose appendix had already perforated when the abdomen was opened.

Negative laparotomy was one where no condition warranting emergency surgery was found at operation. Usually a histopathologically normal appendix was removed.

National Health Service costings are notoriously difficult to measure, as "freed resources" may in practice lie idle and be wasted. This problem is discussed in detail elsewhere (McIntyre *et al*, unpublished report). Some general costs relevant to the present project, however, are, firstly, average total revenue cost per inpatient day, which is the cost of keeping a single patient in hospital for a single day. This is estimated at around £85 for acute hospital beds during 1983-4 (though higher for acute surgical beds). Secondly, direct costs within these overall average total costs are those of actual resources, such as investigations or laparotomy, which are consumed by use and hence saved by non-use.

Results

DIAGNOSTIC ACCURACY

Initial diagnostic accuracy improved from 45·6% during the baseline period to 65·3% during the test period (table II). Accuracy after investigation also improved (from 57·9% to 74·2%), and the rate of bad diagnostic error fell from 6·3% to 2·7%. This improvement was partly explained by a significant decrease (from 14·0% to 2·0%) in the number of cases where no diagnosis at all was made by the first doctor to see the patient ($p<0·001$). The proportion of firm but wrong initial diagnoses, however, also fell significantly (from 44·0% to 31·3%; $p<0·001$). Moreover, an initially incorrect diagnosis was more likely to be corrected by consultation and review during the test period than during the baseline. Despite diversity of locality and admitting procedures, there was considerable consistency between the trends shown in each centre in the trial (fig 1).

DECISION MAKING

Perforated appendix rate—During the baseline period 114 out of 479 inflamed appendices (23·7%) were perforated before removal as opposed

TABLE II—*Diagnostic performance (%) in baseline and test periods (all data combined)*

	Baseline (4075 cases)	Test period (12 662 cases)	Significance
Initial diagnostic accuracy	45·6	65·3	p<0·001
Accuracy after investigation	57·9	74·2	p<0·001
Bad surgical errors	6·3	2·7	p<0·001
Adjustments from wrong to right*	26·9	37·0	p<0·001

*Percentage of cases with wrong initial diagnosis where this was subsequently altered to correct diagnosis by consultation, investigation, and review.

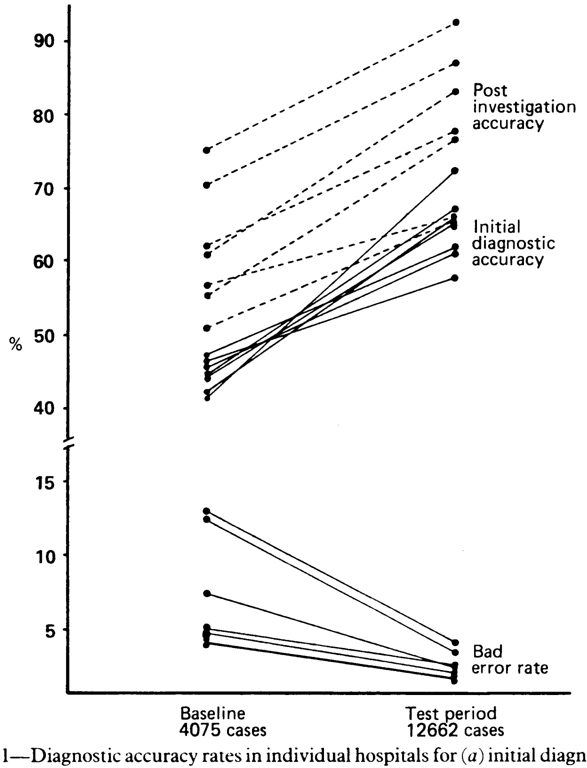


FIG 1—Diagnostic accuracy rates in individual hospitals for (*a*) initial diagnosis by first doctor (reflecting educational aspect); (*b*) accuracy after investigation (reflecting performance of unit); and (*c*) bad surgical error rate (reflecting impact on patient). Note close consistency between hospitals. There was no baseline for the Royal Sussex County Hospital, Brighton.

to 166 out of 1611 (11·5%) during the test period (p <0·001) (fig 2). The fall in the perforation rate for appendicitis was associated with the finding that patients with appendicitis came to surgery earlier during the test period.

Negative laparotomies—The number of laparotomies with negative findings fell from 313 cases a year during the baseline period to 174 cases a year during the test period. Among patients presenting directly to the surgical wards, the negative laparotomy rate fell from 96 out of 382 patients with non-specific abdominal pain (25·2%) during the baseline period to 66 out of 643 (10·4%) during the test period (p<0·001). In hospitals where patients were admitted to hospital via the accident and emergency department during the baseline period 2411 patients with non-specific abdominal pain presented and 228 (9·5%) came to surgery. During the test period 7161 presented and 399 (5·6%) came to surgery (p<0·001).

Bad management errors during the baseline period were made in 0·9% of the patients. During the test period this rate of error fell to 0·2% (p<0·001).

Mortality fell from 1·20% during the baseline period to 0·92% during the test period (p=0·34). Mortality associated with a bad management error (though not common) fell by four fifths (from 0·17% to 0·04%) (p=0·04).

USE OF RESOURCES

Admission rates—When the computer was placed in the accident and emergency department, and this was the department that normally accepted acute admissions, fewer patients with acute abdominal pain were admitted to hospital (table III).

TABLE III—*Patterns of admission to hospital overall and for patients with non-specific abdominal pain. Consideration restricted to three major sites with computer in accident and emergency department where normal mode of admission is via accident and emergency department**

Centre	Baseline period		Test period		Probability
	Patients presenting	No (%) admitted	Patients presenting	No (%) admitted	
		All patients			
Leeds (General Infirmary)	600	472 (78·6)	2759	1506 (54·6)	p<0·001
Leeds (St James's Hospital)	660	442 (67·0)	2808	1503 (43·3)	p<0·001
Whipps Cross	1611	689 (42·8)	3583	1332 (37·2)	p<0·001
Total	2871	1603 (55·8)	9150	4341 (47·4)	p<0·001
		Patients with non-specific abdominal pain			
Leeds (General Infirmary)	304	228 (75·0)	1536	533 (34·7)	p<0·001
Leeds (St James's Hospital)	415	197 (47·5)	1808	530 (29·3)	p<0·001
Whipps Cross	1085	312 (28·8)	2481	469 (18·9)	p<0·001
Total	1804	737 (40·9)	5825	1532 (26·3)	p<0·001

*In other centres normal mode of admission was direct from the general practitioner to surgical wards. Hence there was no reduction in percentage admitted at these sites.

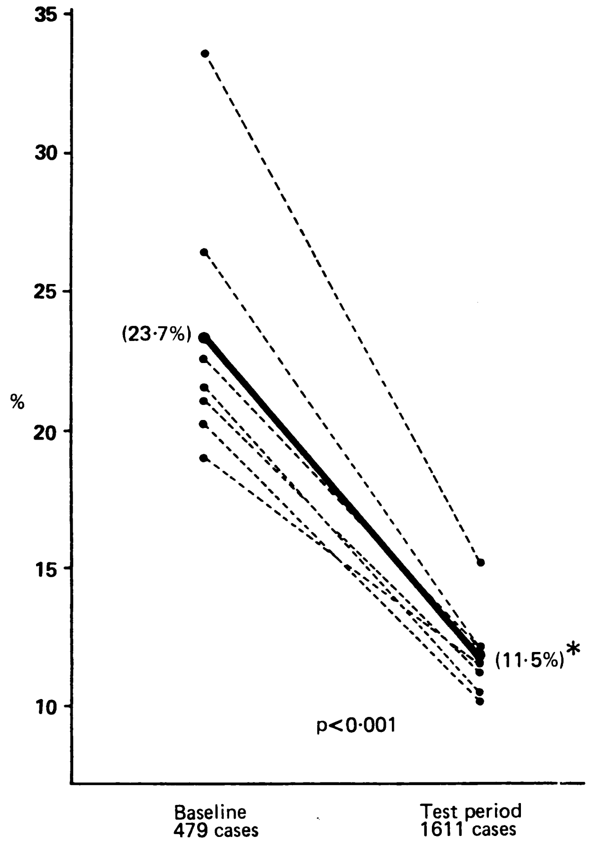


FIG 2—Perforation rates for acute appendicitis cases during baseline and test periods, both for individual hospitals and (solid line) overall for trial.

Rates of stay in hospital were reduced. For patients with non-specific abdominal pain the mean stay fell from 4·0 days during the baseline period to 3·3 during the test period (p<0·001). For patients with appendicitis it fell from 6·7 days to 5·4 days (p<0·001). The saving in terms of hospital bed nights resulting from all these trends was calculated at 4258 bed nights per year: this was equivalent to 8516 bed nights in these hospitals during the course of the trial.

Special investigations—Some reduction was noted in the use of special investigations. Overall during the baseline period 2·4 tests per patient were performed. During the test period this figure fell to 2·0 per patient.

Financial implications—Despite difficulties in measuring the financial savings, clearly the trends described above resulted in considerable savings. Taking values provided by the economic adviser's office of the Department of Health and Social Security as a guide,[9] the average total revenue cost saved in association with the project was estimated at £748 000 over two years. On a national basis (were these results to be repeated more widely) this would imply an average total revenue saving of £23m a year to the National Health Service. When direct cost savings were combined—for example, savings associated with fewer laparotomies, shorter stay, fewer investigations—the estimate of costs saved by the project were £210 000 over two years. Nationally these values implied a direct cost saving of about £5m a year to the National Health Service. Against these figures should be set the cost of the system (£2500 hardware plus £500 per year maintenance and service) and in some centres the salary of a project assistant, usually part time, whose cost averaged £3000 per centre per year.

DIFFERENT SYSTEMS

When forms alone were used initial diagnostic accuracy improved over baseline, from 45·7% to 56·7%. When forms were used and use of personal computers encouraged a further rise in initial diagnostic accuracy was noted (to 64·8%). When all three modes were provided (forms, computer, and feedback) initial diagnostic accuracy was consistently high. All hospitals achieved an average of 68% or higher. Figure 3 illustrates the comparative roles of forms, computers, and feedback. Data relating to bad surgical errors, perforation rates, and admission rates showed similar beneficial

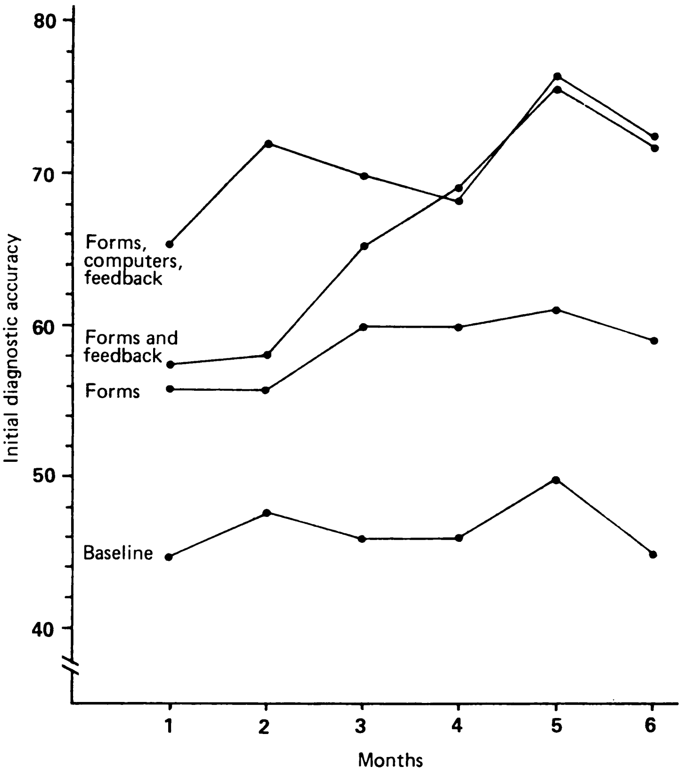


FIG 3—Month by month analysis showing learning curves in mode B hospitals.

trends. Logistic regression analysis showed significant contributions, of about equal size, by forms and computers.

TECHNICAL PERFORMANCE

The "down time" (when participants could not use the computer because of hardware faults) was less than 1%. The most common cause of hardware failure was corruption of disks within the disk drive. The project team were unaware of a time when the system could not be used because of software faults. In 5739 cases the computer system was used and the final diagnosis known. The most probable diagnosis on the relevant computer program matched the final diagnosis in 3911 cases (68·2%). This value excluded cases not run through the computer because the doctor felt the diagnosis to be obvious. When these cases were added computer accuracy rose to over 70%.

USER REACTION

Compliance—According to the trial design, the doctor was required to complete a structured data collection form for 12 610 patients. Forms were completed for 9751 of these patients (77·3%). For 7757 patients the doctor was encouraged to obtain immediate computer feedback. The computer was used personally by the doctor in 3451 cases (44·5%) and by the on site research assistant in a further 2298 cases (29·6%). Computer feedback was thus obtained in 75·1% of possible cases.

Survey after use—After the trial 100 users were polled and asked for comments on the system; 76 responded. Most became familiar with the forms in two to three days and with the computer in about a week. Almost all indicated the system had some impact on their clinical practice; 30·3% of respondents, however, said the system was time consuming if the doctor was busy, and four users (5·2% of respondents) disapproved of the system on principle. When asked whether the system (modified by experience) should be more widely available, 94·3% of user respondents and each of the project leaders said yes.

Discussion

Our work clearly confirms the results of earlier studies[1-4] and suggests that the system can work reliably and effectively in routine practice. More accurate diagnosis by the first doctor to see each patient seems to have led to a higher diagnostic accuracy on the part of more senior doctors, fewer inappropriate admissions and operations, speedier operation for those needing it, and a reduction in hospital admission and stay rates.

The number of deaths in each period was rather small, and the reduction during the trial period did not reach significance. When these data are taken in conjunction with the significant reduction in bad management error rates, however, it can be argued that the improved decision making performance observed during the trial was associated with a significant reduction in risk to the patient.

Demonstrable savings in National Health Service resources also appear to have been associated with the system's introduction. These savings are open to differences of interpretation—for example, they may be expressed in financial terms or in terms of release of facilities for other purposes, such as reducing waiting lists for elective surgery. It seems reasonable to conclude, however, that benefits clearly outweighed costs by a wide margin.

Doctors' reactions to the system were mixed, though most were broadly in favour. The major drawback appeared to be the time taken to use the system when the house surgeon or senior house officer was busy. This was partly associated with the use of currently obsolescent and rather slow computers. A project assistant, who could help at such times, follow up cases, and feed back the results of treatment, proved valuable in several centres.

It would be a major disadvantage if doctors using computer aided decision support systems were to "fall under the spell" of the computer and lose their humanity or their own ability to make decisions. This seems not to have happened in the present trial. The few doctors antagonistic to the system ignored it; the remainder appear to have used the system responsibly. It would also be facile to conclude that the improvement observed was entirely due to the computer feedback, as one major consequence of the "package" provided, of which the computer was merely a part, was that it created a climate in which the inexperienced doctor was stimulated and motivated towards doing the work correctly. Such philosophical argument over the computer's role is important, albeit speculative. Beyond dispute, however, is the improvement in decision making observed among doctors who used the system.

The present results are closer to the findings at Airedale District Hospital (McAdam, unpublished report) and Bangour District General Hospital[3] than to those of the original studies.[1] This is understandable, as the original study dealt with registrars and our study (like Airedale and Bangour) with more junior doctors. The present results, however, represent much more realistically what could be achieved by routine use of the system in the National Health Service.

Overall, the project team concludes that these results suggest that more widespread use by the National Health Service of this system (and by implication similar systems in other clinical areas) is desirable and that inexperienced doctors should be encouraged to make use of such systems. They are no substitute for consultant opinion but are valuable as a combination of special investigation tools, postgraduate educational devices, and, above all, stimuli to good clinical practice.

We thank the many people and agencies whose support made this project possible; the project assistants in the coordinating centre (May Chan, Susan Clamp, Colin Williamson) and participating hospitals (Carol Bates, Margaret Heptinstall, Marilyn Martin, Ann Lawson, Janet Laithwaite, Sue Johannessen); and data processing assistants (George Chan, Linda Chan, and Angela Softley). We thank the World Organisation of Gastroenterology, who made available a database of information about 6000 patients, and the University of Leeds for access to the Amdahl VM7 SPSS/X package.

The work was made possible by funding and support from several agencies, including the Office of Health Economics and the information technology branch of the Department of Health and Social Security, the Northern, Yorkshire, and North East Thames regional health authorities, district health authorities in areas served by participating centres, and the World Organisation of Gastroenterology. The project was administered by Dr B Stone of the University of Leeds industrial services.

Finally, and particularly, we thank our clinical colleagues, both those who contributed to the development of the system and those whose experience is reported here.

Statistical footnote

There are several important aspects in which trials of this type differ from classical multicentre therapeutic studies.[7] Firstly, it is not reasonable to allocate randomly patients to intervention or non-intervention groups; such a device would be very difficult to organise and would entail an unnatural use of the system, but, more importantly, the object of such an educational intervention is the doctor rather than the patient. With doctors as the "experimental unit" some adjustment is necessary to the p values associated with tests on patient statistics; this design effect is often ignored[8] but is fairly simple to implement.[9 10]

Secondly, the use of a baseline control group required special attention to ensure that patient mix and accrual were the same as in the test period and that no changing outside factor had influenced the results. The awareness of being studied can improve behaviour (the Hawthorne effect) but the staggered design in mode B hospitals corrected for this.

Finally, the significance of pooled results over centres was calculated by means of the stratification technique for pooling clinical trials described by Collins *et al* (1985).[11]

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 Wilson DH, Wilson PD, Walmsley GL, *et al*. Diagnosis of acute abdominal pain in the accident and emergency department. *Br J Surg* 1977;**64**:250-4.

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

4 Boom R. Improvement of internist's diagnostic performances by systematic computer evaluations. *Proceedings of Medinfo '80*. Amsterdam: N Holland, 1980:1760.

5 Wilson PE, Horrocks JC, Yeung CK, *et al*. Simplified computer-aided diagnosis of acute abdominal pain. *Br Med J* 1975;ii:73-5.

6 Bouchier IAD, de Dombal FT. Studies co-ordinated by the research committee of the World Organisation of Gastro-Enterology. *Scand J Gastroenterol* 1984;**19**(suppl 95).

7 Spiegelhalter DJ. Evaluation of clinical disease aids, with an application to a system for dyspepsia. *Statistics in Medicine* 1983;2:297-316.

8 Pozen MW, D'Agostino RB, Selker HP, *et al*. A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. *N Engl J Med* 1984;**310**:1273-8.

9 Cornfield J. Randomisation by group; a formal analysis. *Am J Epidemiol* 1978;**108**:100-2.

10 Cochran WG. *Planning and analysis of observational studies*. New York: Wiley, 1983.

11 Collins R, Yusuf S, Peto R. Overview of randomised trial of diuretics in pregnancy. *Br Med J* 1985;**290**:17-23.

(*Accepted 1 July 1986*)

Lesson of the Week

Rest pain and leg ulceration due to syphilitic osteomyelitis of the tibia

M WALZMAN, A A H WADE, S M DRAKE, A M C THOMAS

Many patients with peripheral vascular disease have rest pain and intractable ulceration of the skin. These findings may also be present in late syphilis. We report on a patient with proved peripheral vascular disease, treated by bypass surgery, whose pain was caused by syphilitic osteomyelitis of the tibia.

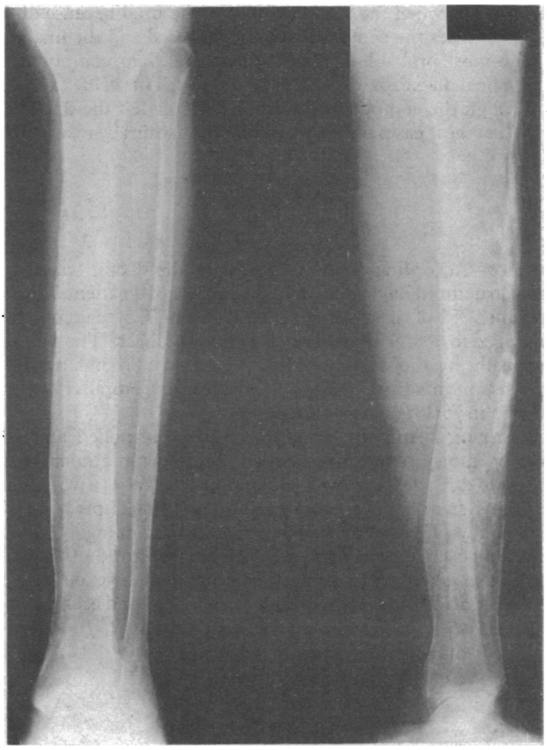
Late syphilis should be considered in patients with leg ulceration and rest pain

Case report

An 81 year old woman had initially presented seven years previously with intractable ulceration of the left leg and foot. She had two ulcers, one on the dorsum of the foot and one on the lower leg, about 3 cm in diameter, with well defined edges but not punched out to any depth and lacking a classic "wash leather" base.[1] No pulses were palpable below the femur on vascular examination. She complained of a continuous deep seated pain in the leg.

Chemical sympathectomy was performed in an attempt to improve the cutaneous circulation. The ulceration persisted, and amputation was considered but was refused by the patient. Femoral arteriography showed a superficial femoral artery block with good run off from the popliteal artery, and femoropopliteal bypass grafting was performed. This failed to improve the leg symptoms and eventually ceased to be patent.

The patient was referred for an orthopaedic surgical opinion because of the finding of extensive areas of mixed osteoporosis and sclerosis in the tibia (figure). Serological tests for syphilis were performed and yielded a positive result for the Venereal Diseases Reference Laboratory test (titre 1/8), *Treponema pallidum* haemagglutination antibody, and fluorescent treponemal antibody (absorption). There was no evidence of neurosyphilis so



Anteroposterior and lateral radiographs, showing mixed lytic and sclerotic syphilitic osteomyelitis of the tibia.

Departments of Genitourinary Medicine and Orthopaedic Surgery, Coventry and Warwickshire Hospital, Coventry CV1 4FH

M WALZMAN, MD, MRCOG, registrar in genitourinary medicine
A A H WADE, MRCOG, consultant in genitourinary medicine
S M DRAKE, MRCP, consultant in genitourinary medicine
A M C THOMAS, FRCS, FRCSED, registrar in orthopaedic surgery

Correspondence to: Dr Wade.