Health Technology Assessment 2006; Vol. 10: No. 47

Systematic reviews of clinical decision tools for acute abdominal pain

JLY Liu, JC Wyatt, JJ Deeks, S Clamp,
J Keen, P Verde, C Ohmann, J Wellwood,
M Dawes and DG Altman

November 2006

Health Technology Assessment
NHS R&D HTA Programme

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Systematic reviews of clinical decision tools for acute abdominal pain

JLY Liu,[1–3] JC Wyatt,[3–5*] JJ Deeks,[1,6] S Clamp,[7]
J Keen,[7] P Verde,[8,9] C Ohmann,[8] J Wellwood,[10]
M Dawes[11,12] and DG Altman[1]

[1] NHS/Cancer Research UK Centre for Statistics in Medicine, Wolfson College, Oxford University, UK
[2] Department of Public Health, Oxford University, UK
[3] Health Informatics Centre, University of Dundee, UK
[4] National Institute for Health and Clinical Excellence (NICE), London, UK
[5] Department of Primary Health Care, Oxford University, UK
[6] Department of Public Health and Epidemiology, University of Birmingham, UK
[7] Yorkshire Centre for Health Informatics, University of Leeds, UK
[8] Coordination Centre for Clinical Trials and Theoretical Surgery Unit, Heinrich-Heine University Düsseldorf, Germany
[9] Department of Statistics, University of Dortmund, Germany
[10] Department of Surgery, Whipps Cross Hospital, London, UK
[11] NHS R&D Centre for Evidence-Based Medicine, Oxford University, UK
[12] Department of Family Medicine, McGill University, Montreal, Canada

* Corresponding author

Declared competing interests of authors: none

Published November 2006

This report should be referenced as follows:

Liu JLY, Wyatt JC, Deeks JJ, Clamp S, Keen J, Verde P, *et al*. Systematic reviews of clinical decision tools for acute abdominal pain. *Health Technol Assess* 2006;**10**(47).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE and *Science Citation Index Expanded (SciSearch®)* and *Current Contents®*/Clinical Medicine.

NHS R&D HTA Programme

The research findings from the NHS R&D Health Technology Assessment (HTA) Programme directly influence key decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC) who rely on HTA outputs to help raise standards of care. HTA findings also help to improve the quality of the service in the NHS indirectly in that they form a key component of the 'National Knowledge Service' that is being developed to improve the evidence of clinical practice throughout the NHS.

The HTA Programme was set up in 1993. Its role is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined to include all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care, rather than settings of care.

The HTA Programme commissions research only on topics where it has identified key gaps in the evidence needed by the NHS. Suggestions for topics are actively sought from people working in the NHS, the public, service-users groups and professional bodies such as Royal Colleges and NHS Trusts.

Research suggestions are carefully considered by panels of independent experts (including service users) whose advice results in a ranked list of recommended research priorities. The HTA Programme then commissions the research team best suited to undertake the work, in the manner most appropriate to find the relevant answers. Some projects may take only months, others need several years to answer the research questions adequately. They may involve synthesising existing evidence or conducting a trial to produce new evidence where none currently exists.

Additionally, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme is able to commission bespoke reports, principally for NICE, but also for other policy customers, such as a National Clinical Director. TARs bring together evidence on key aspects of the use of specific technologies and usually have to be completed within a short time period.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this monograph was commissioned by the HTA Programme as project number 97/38/03. The contractual start date was in May 2001. The draft report began editorial review in February 2006 and was accepted for publication in March 2006. As the funder, by devising a commissioning brief, the HTA Programme specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme or the Department of Health.

ISSN 1366-5278

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.

Abstract

Systematic reviews of clinical decision tools for acute abdominal pain

JLY Liu,[1–3] JC Wyatt,[3–5*] JJ Deeks,[1,6] S Clamp,[7] J Keen,[7] P Verde,[8,9] C Ohmann,[8] J Wellwood,[10] M Dawes[11,12] and DG Altman[1]

[1] NHS/Cancer Research UK Centre for Statistics in Medicine, Wolfson College, Oxford University, UK
[2] Department of Public Health, Oxford University, UK
[3] Health Informatics Centre, University of Dundee, UK
[4] National Institute for Health and Clinical Excellence (NICE), London, UK
[5] Department of Primary Health Care, Oxford University, UK
[6] Department of Public Health and Epidemiology, University of Birmingham, UK
[7] Yorkshire Centre for Health Informatics, University of Leeds
[8] Coordination Centre for Clinical Trials and Theoretical Surgery Unit, Heinrich-Heine University Düsseldorf, Germany
[9] Department of Statistics, University of Dortmund, Germany
[10] Department of Surgery, Whipps Cross Hospital, London, UK
[11] NHS R&D Centre for Evidence-Based Medicine, Oxford University, UK
[12] Department of Family Medicine, McGill University, Montreal, Canada
* Corresponding author

Objectives: To review for acute abdominal pain (AAP), the diagnostic accuracies of combining decision tools (DTs) and doctors aided by DTs compared with those of unaided doctors. Also to evaluate the impact of providing doctors with an AAP DT on patient outcomes, clinical decisions and actions, what factors are likely to determine the usage rates and usability of a DT and the associated costs and likely cost-effectiveness of these DTs in routine use in the UK.
Design: Electronic databases were searched up to 1 July 2003.
Review methods: Data from each eligible study were extracted. Potential sources of heterogeneity were extracted for both questions. For the accuracy review, meta-analysis was conducted. Among studies comparing diagnostic accuracies of DTs with unaided doctors, error rate ratios provided estimates of the differences between the false-negative and false-positive rates of the DT and unaided doctors' performance. Pooled error rate ratios and 95% confidence intervals (CIs) for false-negative rates and false-positive rates were computed. Metaregression was used to explore heterogeneity.
Results: Thirty-two studies from 27 articles, all based in secondary care, were eligible for the review of DT accuracies, while two were eligible for the review of the accuracy of hospital doctors aided by DTs. Sensitivities and specificities for DTs ranged from 53 to 99% and from 30 to 99%, respectively. Those for unaided doctors ranged from 64 to 93% and from 39 to 91%, respectively. Thirteen studies reported false-positive and false-negative rates for both DTs and unaided doctors, enabling a direct comparison of their performance. In random effects meta-analyses, DTs had significantly lower false-positive rates (error rate ratio 0.62, 95% CI 0.46 to 0.83) than unaided doctors. DTs may have higher false-negative rates than unaided doctors (error rate ratio 1.34, 95% CI 0.93 to 1.93). Significant heterogeneity was present. Two studies compared the diagnostic accuracies of doctors aided by DTs to unaided doctors. In a multiarm cluster randomised controlled trial ($n = 5193$), the diagnostic accuracy of doctors not given access to DTs was not significantly worse (sensitivity 28.4% and specificity 96.0%) than that of three groups of aided doctors (sensitivities of 42.4–47.9%, and specificities of 95.5–96.5%, respectively). In an uncontrolled before-and-after study ($n = 1484$), the sensitivities and specificities of aided and unaided doctors were 95.5% and 91.5% ($p = 0.24$) and 78.1% and 86.4%

($p < 0.001$), respectively. The metaregression of DTs showed that prospective test-set validation at the site of the tool's development was associated with considerably higher diagnostic accuracy than prospective test-set validation at an independent centre [relative diagnostic odds ratio (RDOR) 8.2; 95% CI 3.1 to 14.7]. It also showed that the earlier in the year the study was performed the higher the performance (RDOR 0.88, 0.83 to 0.92), that when developers evaluated their own DT there was better performance than when independent evaluators carried out the study (RDOR = 3.0, 1.3 to 6.8), and that there was no evidence of association between other quality indicators and DT accuracy. The one eligible study of the impact study review, a four-arm cluster randomised trial ($n = 5193$), showed that hospital admission rates of patients by doctors not allocated to a DT (42.8%) were significantly higher than those by doctors allocated to three combinations of decision support (34.2–38.5%) ($p < 0.001$). There was no evidence of a difference between perforation rates ($p = 0.19$) and negative laparotomy rates in the four trial arms ($p = 0.46$). Usage rates of DTs by doctors in accident and emergency departments ranged from 10 to 77% in the six studies that reported them. Possible determinants of usability include the reasoning method used, the number of items used and the output format. A deterministic cost-effectiveness comparison demonstrated that a paper checklist is likely to be 100–900 times more cost-effective than a computer-based DT, under stated assumptions.

Conclusions: With their significantly greater specificity and lower false-positive rates than doctors, DTs are potentially useful in confirming a diagnosis of acute appendicitis, but not in ruling it out. The clinical use of well-designed, condition-specific paper or computer-based structured checklists is promising as a way to improve impact on patient outcomes, subject to further research.

Contents

Glossary and list of abbreviations

Technical terms and abbreviations are used throughout this report. The meaning is usually clear from the context, but a glossary is provided for the non-specialist reader. In some cases, usage differs in the literature, but the term has a constant meaning throughout this review.

Glossary

Acute abdominal pain (AAP) The presentation of previously undiagnosed abdominal pain lasting one week or less prior to a clinical encounter in primary or secondary care.[1]

Algorithm A process for carrying out a complex task broken down into simple decision and action steps. In the clinical context, it is a form of clinical decision tool with pathways linking clinical findings, decision boxes and action boxes. Synonym: flowchart.

Appendectomy American term for appendicectomy.

Appendicectomy Operation in which the appendix is removed.

Benefit–cost ratio A decision index used in cost–benefit analysis, which is the "sum of discounted benefits divided by the sum of discounted costs [B/C]. Any project with a B/C greater than 1 is potentially acceptable, and the higher the ratio, the better."[2]

Burden of disease See Cost of illness studies.

Checklist A type of clinical decision tool: a form listing one or more items of patient data to be collected before, during or after an encounter; can be paper or computer based.

Cholecystectomy Surgical removal of the gallbladder.

Clinical actions Data collection (e.g. examination of the patient) and recording for clinical purposes (rather than research), diagnosis, test ordering, referral, admission, discharge, prescribing, ordering of procedures (e.g. surgery), giving of prognosis, etc.

Clinical decision tools Decision tools used during clinical encounters between the patient and the doctor.

Clinical decisions Decisions about any clinical action.

Clinical practice guideline A document that is sometimes abbreviated as a clinical algorithm, which gives patient management advice, often informed by relevant evidence and group decision-making.

Computed tomography (CT) A method to produce images of body organs by using a computer to combine a set of two-dimensional X-ray images into a three-dimensional image.[3,4] The traditional view is that CT only has a limited role in the diagnosis of AAP patients.[5] However, Rao and colleagues demonstrated that CT is highly accurate diagnostically for AAP patients with suspected acute appendicitis (93–98% for sensitivity and specificity).[6] The study also indicated that routine CT of the appendix in these patients results in improved health outcomes.

Conjoint analysis According to Phillips and colleagues, this "is an approach to measuring preferences (utilities) that estimates both overall preferences for a good or service as well as preferences for its specific attributes."[7]

Cost–benefit analysis An economic evaluation that considers the costs and effects

continued

Glossary *continued*

of at least two alternative courses of action, where costs and benefits are measured in money terms. One decision index used is the benefit–cost ratio.

Cost-effectiveness analysis An economic evaluation that considers the costs and effects of at least two types of possible intervention where the effects are measured in natural units (e.g. life-years gained, cases of illness or morbidity avoided, procedure or outcome prevented or enhanced). The decision index is cost per unit of effect.

Cost of illness studies An assessment of the full burden of a disease in terms of what costs are involved, when they occur and where. Useful for identifying research priorities. Synonym: burden of disease study.

Cost studies Detailed assessment of the costs of developing, implementing, supporting and training people to use an intervention, e.g. a decision tool.

Cost–utility analysis An economic evaluation that considers the costs and effects of at least two types of possible healthcare interventions where the effects are not the same but are measured in units of utility or satisfaction (e.g. quality-adjusted life-years). The decision index is cost per quality-adjusted life-year.

Decision support system A type of clinical decision tool: a computer system that uses two or more items of patient data to generate case-specific or encounter-specific advice; includes prognostic model.

Decision tool A knowledge resource that supports decision-making about an individual patient by a health worker, the patient themselves or someone else concerned about them.

Decision tree A framework that represents alternative decisions, outcomes arising from those decisions and the probabilities of occurrence of those outcomes.

Delphi method "Iterative circulation to a panel of experts of questions and responses that are progressively refined in light of responses to each round of questions … The aim is to reduce the number of viable options or solutions, perhaps to arrive at a consensus judgement on an issue or problem, or a set of issues or problems, without allowing anyone to dominate the process".[8]

Diagnostic odds ratio The ratio of the odds of a positive test result in a patient with disease compared with a patient without disease.

Discount rate If the effects of a programme take place over time, adjustments need to be made to maintain the comparability of net benefits (i.e. benefits minus costs) over different periods. The discount rate weights future years to make benefits and costs comparable over time. The rationale for and the determination of the discount rate are discussed in standard economics texts covering economic evaluation, including Barron and colleagues.[2]

Economic evaluation A study in which both the costs and effects of alternative health interventions or programmes are analysed and compared. Cost-effectiveness analysis, cost–utility analysis and cost–benefit analysis are economic evaluations, whereas cost studies and cost-of-illness studies are not.

Economies of scale Decreases in operating costs resulting from an increase in the scale of operations of a hospital or clinic.[9]

Economies of scope Savings in operating costs of related healthcare activities/services/systems as a result of their joint provision by the same healthcare organisation.[9]

False-negative rate The probability that the test or tool will give negative results in cases with the disease (e.g. appendicitis).

False-positive rate The probability that the test or tool will give positive results in cases without the disease (e.g. non-specific abdominal pain).

continued

Glossary *continued*

Funnel plot A graphical device used to detect publication bias and other small-study effects, constructed by plotting the measure of effect (e.g. relative risk or risk difference) of each study included in a meta-analysis against a measure of its precision (e.g. inverse of the standard error or sample size). In the absence of publication bias and other small-study effects, the shape of the plot will be similar to that of an inverted funnel. In the presence of publication bias, the plot will have an asymmetrical shape. For more details, see Sterne and colleagues.[10]

Incremental cost-effectiveness ratio The ratio of the difference between the costs of two healthcare-related programmes to the difference in effectiveness between them. Common measures of effectiveness include life-years gained and quality-adjusted life-years gained.

Knowledge base A store of knowledge represented explicitly so that a computer can search and reason with it automatically.

Knowledge-based system (expert system) A computer decision support system with an explicit knowledge base and separate reasoner program that uses this to give advice or interpret data, often patient data.

Laparoscopy "A technique by which the contents of the abdomen may be examined, biopsies taken, and minor surgical procedures carried out, by the insertion of a tube through a small hole made in the abdominal wall."[3] Laparoscopy is increasingly used in operations to remove the appendix.

Laparotomy "A general term applied to any operation in which the abdominal cavity is opened."[3]

Likelihood ratio positive/negative The probability that a given test result would be expected in a patient with the target disorder compared with the probability that the same result would be expected in a patient without the target disorder.

Medical informatics The study and application of methods to improve the management of patient data, medical knowledge, population data and other information relevant to patient care and community health. Unlike some other definitions of medical informatics (e.g. Shortliffe and colleagues[11]), this definition puts the emphasis on information management rather than technology.[12]

Negative appendicectomy rate Incidence rate of appendicectomies in which the excised appendix was found to be normal.

Negative laparotomy rate Incidence rate of laparatomies in which the excised appendix was found to be normal and no other abnormality was identified in the abdominal cavity.

Net present value Sum of benefits minus the sum of costs discounted to convert projected future costs/benefits into the present value.[2]

Nominal group technique Also called the expert panel. "Uses a highly structured meeting to gather information from relevant experts (usually nine to twelve in number) about a given issue. The technique consists of two rounds in which panellists rate, discuss and then re-rate a series of items or questions."[13]

Non-specific abdominal pain Acute abdominal pain that is not due to an identified serious cause.

Patient data Information about an individual patient and potentially relevant to decisions about her current or future health or illness. Patient data should be collected using methods that minimise systematic and random error.

Percutaneous cholecystostomy "A minimally invasive procedure [to treat acute cholecystitis] that can benefit patients with serious comorbidity who are at high risk from major surgery."[14]

Perforated appendix rate Incidence rate of appendicectomies in which the appendix was perforated. Perforations can occur because of delays in operating on an AAP patient.

continued

Glossary *continued*

Peritoneum A serous membrane that covers organs in the abdomen and lines the abdominal cavity. The peritoneum secretes fluid to lubricate the membrane and organs in the abdomen.[3]

Peritonitis Inflammation of the peritoneum due to bacteria or acid entering the abdominal cavity, often caused by perforation of the appendix or colon, whose contents are contaminated.

Primary care UK general practice, US family practice or community-based clinics elsewhere.

Prognostic model/tool A form of decision support system that calculates the probability of an outcome from two or more items of patient data.

Publication bias The publication or non-publication of research findings, depending on the nature and direction of the results.

Radiology Radiology is not used routinely in AAP patients, but may be valuable when the diagnosis is uncertain, e.g. when "the cause of right-sided abdominal pain is obscure".[4]

Receiver operating characteristic (ROC) plot A plot of the true-positive rate (sensitivity) against the false-positive rate (1 – specificity), used either to display results from a set of studies or for illustrating how results vary with choice of cut-point.

Relative diagnostic odds ratio A measure of differences in observed diagnostic ability between two tests/groups/types of study, calculated by dividing the DOR from one study by the DOR from the other.

Reminder A type of clinical decision tool that reminds a doctor about some item of patient data or clinical knowledge relevant to an individual patient that they would be expected to know. Can be paper based or computer based.

Secondary care Ambulatory care, outpatients or A&E (casualty) departments or ward referrals.

Serious cause of acute abdominal pain One for which hospital admission is judged necessary, but surgery is not necessarily needed (e.g. acute pancreatitis).

Serum amylase A special investigation in which raised levels of serum amylase could indicate a number of conditions in AAP patients: mesenteric infarct, acute pancreatitis, peptic ulcer and acute cholecystitis. AAP patients should not be routinely tested for serum amylase.[15]

Small-study effects The tendency of smaller or less precise studies in a meta-analysis to show larger treatment effects. Causes include publication and other reporting biases, trial quality and heterogeneity between trials. For a more detailed discussion, see Sterne and colleagues.[10]

Summary ROC curve An approach for meta-analysis of diagnostic accuracy that involves fitting a curve through the points on an ROC curve that would arise if the differences in studies arose owing to differences in threshold.

Telemedicine The use of any electronic medium to mediate or augment clinical consultations. It can be simultaneous (e.g, telephone, videoconference) or store and forward (for example, an e-mail with an attached image). (Excluded from this review for reasons to be explained in main text.)

Test set Data used to validate the performance of a decision tool after it has been developed, assessed and fine-tuned with training set data.

Training set Data used to develop and fine-tune the performance of a new decision tool.

Ultrasonography Test in which "direct high frequency sound waves … are passed through the body. Part of this sound echoes back from the tissues and is recorded at skin level by a transducer."[1] It can be used as a treatment (e.g. to break up kidney stones) or a diagnostic test.[16] For the purposes of this monograph, the focus is on ultrasonography as a diagnostic test for AAP conditions.

continued

Glossary *continued*

Urinalysis The assessment of urine, to ascertain levels of normal and abnormal contents, involving the use of chemical and physical tests and microscopy.[15] According to Clamp,[15] the routine employment of urine tests, such as rapid testing using urinary dipstick, is not recommended. Urinalysis is helpful for the diagnosis of patients who are suspected of having urinary tract infection and renal colic.[15]

White blood cell count The routine use of this test is not needed for all AAP patients.[15] The test is helpful for the diagnosis of AAP patients with suspected acute appendicitis, but whose diagnosis is uncertain after clinical interview and physical examination. High white blood cell levels beyond a certain threshold ($13 \times 10^9\ \mathrm{l}^{-1}$) may indicate acute appendicitis, but other conditions could also account for a raised count, e.g. diverticulitis, peritonitis, perforated peptic ulcer and cholecystitis.[15]

List of abbreviations

A&E	accident and emergency
AAP	acute abdominal pain
AI	artificial intelligence
ANOVA	analysis of variance
ART	adaptive resonance therapy
AUROC	area under the receiver operating characteristic curve
BP	Back Propagation
CADA	Computer Assisted Diagnostic and Audit database
CI	confidence interval
CONSORT	Consolidated Standards of Reporting Trials
DOR	diagnostic odds ratio
DSS	decision support system
DT	decision tool
FN	false negative
FP	false positive
LVQ	Learning Vector Quantisation
NSAP	non-specific abdominal pain
PDA	personal digital assistant
R&D	research and development
RAPT	Review of Abdominal Pain Tools
RCT	randomised controlled trial
RDOR	relative diagnostic odds ratio
ROC	receiver operating characteristic
SD	standard deviation
SE	standard error
SHO	senior house officer
SOM	Self-Organising Map
SROC	summary receiver operating characteristic
STARD	Standards for Reporting of Diagnostic Accuracy
TN	true negative
TP	true positive

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices in which case the abbreviation is defined in the figure legend or at the end of the table.

Executive summary

Background

Making accurate decisions for patients with acute abdominal pain (AAP) is difficult. To avoid missing seriously ill patients, many undergo unnecessary surgery, with negative laparotomy rates of 25%. However, delays can lead to 20% perforation rates. Many conditions cause AAP and no single clinical finding or test is both specific and sensitive. Many decision tools (DTs) combining two or more findings have been developed to aid AAP management, but no consensus exists on their appropriateness for clinical use.

Objectives

The study aimed to answer the following questions.

1. What are the diagnostic accuracies of DTs and doctors aided by DTs compared with those of unaided doctors?
2. What is the impact of providing doctors with an AAP DT on patient outcomes, clinical decisions and actions?
3. What factors are likely to determine the usage rates and usability of a DT?
4. What are the associated costs and likely cost-effectiveness of these DTs in routine use in the UK?

Methods

Data sources

MEDLINE, EMBASE, CINAHL, INSPEC CENTRAL, SIGLE and HEALTH-CD were searched for empirical English-language studies. Searches were conducted to 1 July 2003.

Study selection (inclusion criteria)

For question 1, the criteria for eligible studies included:

- Unselected patients with AAP were recruited consecutively or randomly sampled from a primary or secondary care setting.
- Patients had previously undiagnosed AAP lasting for 7 days or less from onset.
- The study reported accuracies of AAP DTs, with or without comparisons to unaided doctors' decisions.
- An adequate reference standard was described.
- Sensitivity and specificity could be calculated.

For question 2, the criteria for eligible studies included:

- The study was a randomised controlled trial (RCT) or quasi-RCT.
- The patients were the same as for question 1.
- Evaluations were conducted of the impact of AAP DTs, compared with unaided doctors' decisions.
- The study reported some measure of impact on patient outcomes, clinical decisions or actions.

Data extraction

Data from each eligible study were extracted. For question 1, this included patient characteristics, type of DT, healthcare setting, and the accuracy of DTs and unaided doctors' decisions. For question 2, this included outcomes, clinical decisions and actions for patients of doctors aided or unaided by a DT. Potential sources of heterogeneity were extracted for both questions.

Data synthesis

For the accuracy review, meta-analysis was conducted. Among studies comparing diagnostic accuracies of DTs with unaided doctors, error rate ratios provided estimates of the differences between the false-negative and false-positive rates of the DT and unaided doctors' performance. Pooled error rate ratios and 95% confidence intervals (CIs) for false-negative rates and false-positive rates were computed. Metaregression was used to explore heterogeneity.

Results

Question 1

Thirty-two studies from 27 articles, all based in secondary care, were eligible for the review of DT accuracies, while two were eligible for the review of the accuracy of hospital doctors aided by DTs. Sensitivities and specificities for DTs ranged from 53 to 99% and 30 to 99%, respectively. Those for unaided doctors ranged from 64 to 93% and 39 to 91%, respectively. Thirteen studies reported false-positive and false-negative rates for both DTs and unaided doctors, enabling a direct comparison of their performance. In random effects meta-analyses, DTs had significantly lower false-positive rates (error rate ratio 0.62, 95% CI 0.46 to 0.83) than unaided doctors. DTs may have higher false-negative rates than unaided doctors (error rate ratio 1.34, 95% CI 0.93 to 1.93). Significant heterogeneity was present.

Two studies compared the diagnostic accuracies of doctors aided by DTs to unaided doctors. In a multiarm cluster RCT ($n = 5193$), the diagnostic accuracy of doctors not given access to DTs was not significantly worse (sensitivity 28.4% and specificity 96.0%) than that of three groups of aided doctors (sensitivities of 42.4–47.9%, and specificities of 95.5–96.5%, respectively). In an uncontrolled before-and-after study ($n = 1484$), the sensitivities and specificities of aided and unaided doctors were 95.5% and 91.5% ($p = 0.24$) and 78.1% and 86.4% ($p < 0.001$), respectively.

The metaregression of DTs showed that:

- prospective test-set validation at the site of the tool's development was associated with considerably higher diagnostic accuracy than prospective test-set validation at an independent centre [relative diagnostic odds ratio (RDOR) 8.2; 95% CI 3.1 to 14.7]
- the earlier in the year the study was performed the higher the performance (RDOR 0.88, 0.83 to 0.92)
- when developers evaluated their own DT there was better performance than when independent evaluators carried out the study (RDOR = 3.0, 1.3 to 6.8)
- there was no evidence of association between other quality indicators and DT accuracy.

Question 2

The one eligible study of the impact study review, a four-arm cluster randomised trial ($n = 5193$), showed that hospital admission rates of patients by doctors not allocated to a DT (42.8%) were significantly higher than those by doctors allocated to three combinations of decision support (34.2–38.5%) ($p < 0.001$). There was no evidence of a difference between perforation rates ($p = 0.19$) and negative laparotomy rates in the four trial arms ($p = 0.46$).

Question 3

Usage rates of DTs by doctors in accident and emergency departments ranged from 10 to 77% in the six studies that reported them. Possible determinants of usability include the reasoning method used, the number of items used and the output format.

Question 4

A deterministic cost-effectiveness comparison demonstrated that a paper checklist is likely to be 100–900 times more cost-effective than a computer-based DT, under stated assumptions.

Conclusions

Implications for healthcare

- With their significantly greater specificity and lower false-positive rates than doctors, DTs are potentially useful in confirming a diagnosis of acute appendicitis, but not in ruling it out.
- The clinical use of well-designed, condition-specific paper or computer-based structured checklists is promising as a way to improve impact on patient outcomes, subject to further research.

Recommendations for research

This review uncovered important evidence gaps. The authors' research recommendations include the following:

- Primary research to compare paper-based checklists with computer-based tools exploring the type/format that maximises patient benefit.
- Empirical research to identify the determinants of successful DTs, to provide more evidence to support the development of clinically useful tools.
- More general systematic reviews (across a range of diseases or tools) to assess (1) factors that make DTs more acceptable to doctors and patients and (2) the relative clinical value of paper checklists versus computer-based tools.

Chapter 1
Background to study

Introduction

Acute abdominal pain (AAP) is a common problem in secondary care. Although more than 1000 causes of AAP exist,[16] over 80% of cases in secondary care can be explained as acute appendicitis (26%), non-specific abdominal pain (NSAP) (50%) and acute cholecystitis (8%) alone [see *Table 1* for a detailed breakdown of the case-mix seen in surgical and accident and emergency (A&E) departments in the UK).[1] Some AAP conditions (acute appendicitis, small bowel obstruction and perforated peptic ulcer) require emergency hospital admission and surgery. Around 50% of AAP patients in secondary care have a non-specific cause and are said to have NSAP. In primary care, less than 10% of AAP patients have serious disease requiring surgery and the case-mix of AAP is different from that in secondary case: enteritis, gastritis, dyspepsia and dysmenorrhoea constitute over 90% of cases.[1] GPs must assess which patients with AAP need referral to the hospital. Among patients seen at A&E, casualty officers need to decide which patients should be admitted to hospital, and once admitted, those who require surgery need to be identified.

The epidemiology of AAP

AAP exerts a considerable burden on the health services. Among the many causes of AAP, acute appendicitis alone accounts for the most common surgical operation in the West.[17,18] *Stedman's Medical Dictionary* defines the 'acute abdomen' as "any serious acute intra-abdominal condition (such as appendicitis) attended by pain, tenderness, and muscular rigidity, and for which emergency surgery must be considered."[19] For empirical research, there is a need to be more precise about what is meant by 'acute'. A widely accepted definition of AAP is adopted here: the presentation of previously undiagnosed abdominal pain lasting for 1 week or less before a clinical encounter in primary or secondary care.[1] This definition distinguishes AAP from chronic or recurrent abdominal pain.

There is some evidence that the AAP case-mix in the UK is changing over time. In the 1950s, perforated peptic ulcer was the second most common cause of surgery in AAP patients behind appendicitis, but its incidence has decreased substantially in recent decades and may now be rarer than the 1991 figures in *Table 1*.[1,20–22] A 25-year study of emergency surgical admissions indicates a decrease in admissions for acute appendicitis and intestinal obstruction from 1974 to 1998 and an increase in admissions relating to gallstones, NSAP and diverticular diseases.[23] The Oxford Record Linkage Study also suggest a drop in hospital admission rates for acute appendicitis, but no decrease for diseases with signs and symptoms that resemble it.[24]

TABLE 1 *Causes of AAP in patients admitted direct to a surgical ward, admitted via a UK A&E department and in a worldwide sample*[1]

Cause of AAP	Admissions direct to a UK surgical ward *n* (% of total)	Admissions via a UK A&E department *n* (% of total)	Worldwide sample of hospital admissions *n* (% of total)
NSAP	279 (50.5)	532 (53.7)	3,507 (34.0)
Appendicitis	145 (26.3)	187 (18.9)	2,895 (28.1)
Cholecystitis	42 (7.6)	69 (7.0)	1,005 (9.7)
Small bowel obstruction	20 (3.6)	31 (3.1)	423 (4.1)
Perforated peptic ulcer	17 (3.1)	27 (2.7)	253 (2.5)
Pancreatitis	16 (2.9)	16 (1.6)	302 (2.9)
Diverticular disease	11 (2.0)	13 (1.3)	151 (1.5)
Urinary tract problems	–	55 (5.5)	–
Gynaecological problems	–	49 (4.9)	413 (4.0)
Others	22 (4.0)	12 (1.2)	1,371 (13.3)
	552 (100.0)	991 (100.0)	10,320 (100.0)

Acute appendicitis

Acute appendicitis refers to the acute inflammation of the appendix. The exact causal pathway is not known, but theories include possible roles played by diet, genetic factors, infection and other factors in the aetiology of appendicitis.[25] In England and Wales, the annual incidence proportion is 60,000 out of a population of 53.1 million.[17] A decrease in the incidence of acute appendicitis in the UK has been reported.[17] In Britain, the lifetime individual risk of acute appendicitis is 7%, similar to American figures of 8.6% for males and 6.7% for females.[17,26] In a population-based study of childhood deaths from appendicitis in England and Wales from 1963 to 1997, mortality rates had fallen from 1.06 per 1000 discharges in 1963–1967 to 0.16 per 1000 discharges in 1993–1997.[27] Possible explanations for the decreases include changing distributions of childhood diseases, risk factor exposures and better hygiene, but the exact reasons are uncertain.

Surgical removal of the appendix (known as appendicectomy or appendectomy) is the standard treatment for people with suspected appendicitis. Delays in surgery can result in perforation of the appendix; 1.7% of patients with a perforated appendix die, compared with 0.3% of those without perforation (relative mortality risk 5.7).[17]

Complications of surgery include wound infection (in 5–33% of patients) and abscess (2% of patients).[17] The Oxford Record Linkage Study showed that men had emergency appendicectomies more often than women, while negative appendicectomy rates were higher for women than for men.[24] According to the 17-year study, positive appendicectomy rates have declined while negative appendicectomy rates have remained steady.

Acute cholecystitis

Acute cholecystitis is a condition in which the gallbladder becomes acutely inflamed, the most common cause of which is gallstones.[14] Acute cholecystitis is the third most common diagnosis (after NSAP and appendicitis). In the West, the lifetime prevalence of gallstones in adults is about 10%, of whom 80% show no symptoms. One to three per cent of people who have gallstones and experience symptoms go on to develop acute cholecystitis per year. The diagnosis of acute cholecystitis is obtained from the patient's symptoms (constant pain in the right upper quadrant of the abdomen for more than 12 hours and tenderness in the right upper quadrant) and inflammation [reflected by fever, white blood cell count and C-reactive protein higher than normal]. Inflammation indicates possible peritonitis. Treatment options include the following:[14,28]

- fasting, intravenous fluids and analgesia as initial treatments
- cholecystectomy at any time after admission, a common procedure to treat acute cholecystitis today
- percutaneous cholecystostomy for patients who are too ill or unfit to undergo an operation, a rare procedure nowadays. Four randomised controlled trials (RCTs) showed no evidence of a difference "between early (within 72 hours) and delayed cholecystectomy … in rates of intraoperative or postoperative complications"[28]
- emergency surgery to treat patients who have perforated gallbladder or gangrenous cholecystitis, an uncommon treatment option today.

Other diseases requiring surgery

Emergency surgery is almost always needed for perforated peptic ulcer and intestinal obstruction.[1,16] Since the prevalence of peptic ulcer has declined substantially, perforations have become less common.[20–22] For other conditions, such as acute cholecystitis, acute pancreatitis or diverticular disease,[14,29,30] the choices are less clear-cut. Conservative measures are usually used before surgery is considered, unless evidence of perforation or blockage exists.

Non-specific abdominal pain

NSAP refers to AAP that is not attributable to an identified cause. It is the most common diagnosis among secondary care patients with AAP and one of the most common reasons for hospital admissions.[1] A study on the incidence of NSAP in UK children suggests that it consists of a diverse set of conditions with different and probably multifactorial aetiologies.[31] Psychological factors are one set of associated factors.[32] The condition tends to be self-limiting. In most cases, recovery is spontaneous, patients are discharged and no cause is ever found.[1] However, the literature in this area is scarce. A systematic review on the health status of discharged patients with NSAP is therefore needed.

Gynaecological disorders

Gynaecological disorders that cause AAP in women admitted to secondary care include pelvic inflammatory disease (also called salpingitis), unsuspected abortion, ovarian disease and ectopic pregnancy.[16] Of these conditions, ectopic

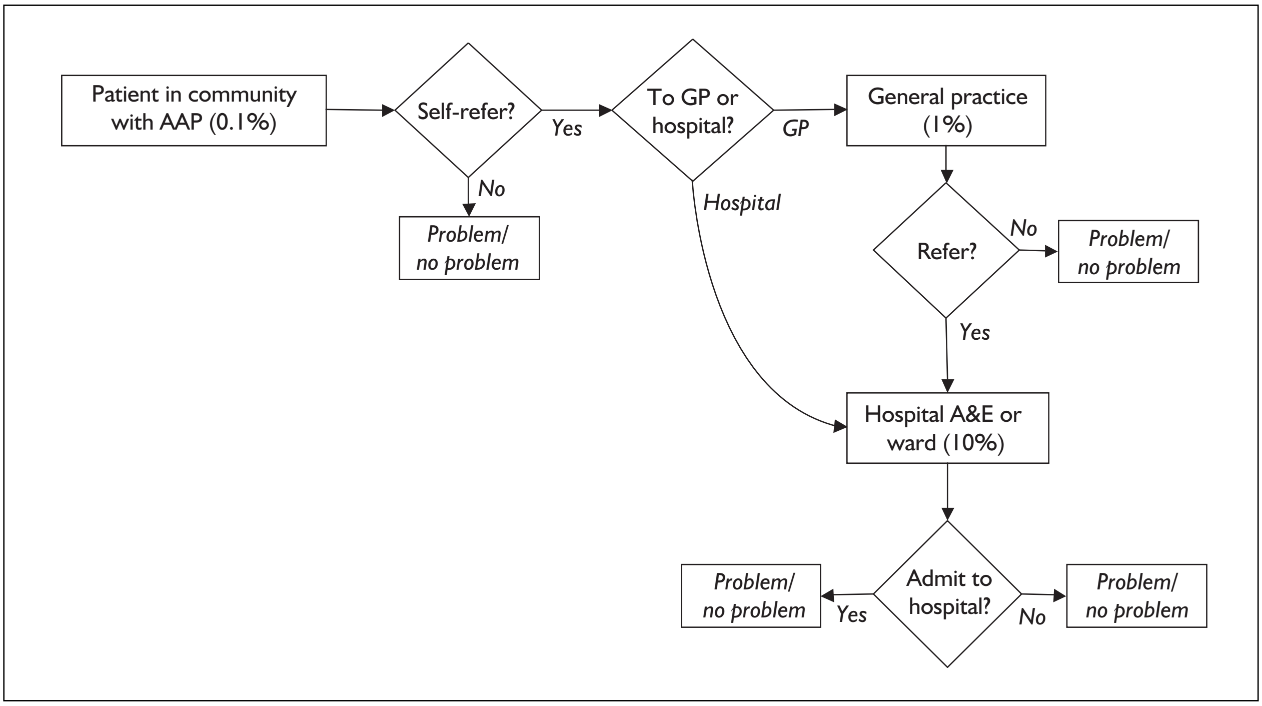

FIGURE 1 *Model of management of patients with AAP. Illustrative figures in parentheses are approximate risks of serious causes of AAP. Problem/no problem: perspectives differ, and include the patient, the professional and the healthcare system.*

pregnancy is probably the most serious, happens in about 1% of pregnancies and causes 11% of pregnancy-associated deaths. The initial diagnosis is often wrong.[5] Major ovarian disease and ectopic pregnancy are normally diagnosed during an operation and confirmed through histopathology, which is also used to diagnose incomplete abortion.[1] Laparoscopy is commonly used to diagnose pelvic inflammatory disease.[5]

Children and the elderly

Children and the elderly are subgroups of particular interest, since the distribution of diseases causing AAP in these two groups is different from that in other age groups.[16,33] In children, acute appendicitis (32%) and self-limiting conditions (61%) constitute more than 90% of hospital admissions.[16] About 60–70% of admitted children with AAP in the UK are classified by doctors as having NSAP and discharged, with the remainder undergoing surgery.[31] Intussusception and urinary tract infection are other occasional causes of AAP in children. Although appendicitis occurs less often in younger than in older children, the diagnosis is more difficult in the former.[34,35] Among children with acute appendicitis, only one-third will present with classic symptoms, and the initial diagnosis is often incorrect.[35,36] Between 70 and 100% of initial diagnoses were wrong in very young children (aged 3 years or less). This figure dropped to 12–28% in schoolchildren and less than 15% in teenagers. A prospective cohort study found that computed tomography could greatly decrease and ultrasonography could decrease the misdiagnosis rate in children.[37]

Although appendicitis and cholecystitis are common conditions in the elderly, other possibilities include cancer and mesenteric vascular disease.[16,33] Perforation occurs more often in older AAP patients. Perforation rates in this subgroup are in the range of 55–70%, given problems with delays in diagnosis and difficulties in obtaining the correct diagnosis.[38] Another challenge is the poor memory of many elderly patients, which often makes it difficult to ascertain the initial signs and symptoms and when they began.[39,40]

Current management standards for AAP

Figure 1 shows the typical management and referral pathways for AAP patients through the UK healthcare system. Substantial changes are likely in the risk of serious causes of AAP as patients move through the healthcare system, with those in the community having the lowest risk of a serious problem, compared with higher risk for those visiting the GP, a much higher risk for those admitted to A&E and the highest risk for those in surgical wards.

Diagnosis in primary and secondary care

In primary care, the key decision for GPs is to exclude serious causes requiring emergency treatment. They therefore need highly sensitive tests or signs to rule out such causes.[41] If the duration of AAP is 6 months, they can quite safely exclude appendicitis. If a patient has diarrhoea, that is more likely to indicate gastroenteritis than appendicitis.[42] In the initial phase of acute appendicitis, for example, the sole symptom is often abdominal pain migrating from the umbilical region to the right iliac fossa. About 1% of primary care patients have appendicitis each year and GPs will tend to refer a patient if they suspect at least a 10% probability of acute appendicitis or other condition needing urgent surgery.[43] If symptoms are mild and the GP's assessed probability of a serious condition is only about 2 or 3%, the appropriate initial action is to wait for 24–48 hours and treat the patient with analgesics for pain relief and antibiotics.[42,44] A GP needs to make a decision within 1 or 2 days on whether to refer the patient to A&E or hospital.

The primary concern of a doctor in secondary care is whether or not to operate on an AAP patient, given the risk of perforation and peritonitis should the decision be delayed for too long and also the uncertainty regarding the underlying condition. In the absence of a clinical decision tool, a secondary care doctor's diagnostic strategy for AAP should mainly be based on the clinical interview and physical examination. The following aspects of patient history should be covered in a clinical interview with an AAP patient:[1,5]

- demographic data: age and gender
- pain: site at onset, site now, time since onset, severity and type of pain, progress, radiation, and factors that aggravate or relieve pain
- other symptoms: nausea, vomiting, feeling faint, jaundice, appetite, bowel habit, micturition, and in female patients, gynaecological symptoms
- previous history: similar pain in the past, abdominal surgery in the past, previous indigestion, history of major illness, allergies and drugs.

An AAP patient should also be examined for the following signs:[1,5]

- general examination: mood, pulse, colour, temperature, respiration and blood pressure
- abdominal examination:
  - inspection: movement, scars and distension
  - palpation: tenderness, rebound, guarding, rigidity, Murphy's sign, swellings
- other examinations: rectal examination and in female patients, vaginal examination
- auscultation: bowel sounds.

Detailed accounts of diagnosing AAP conditions can be found elsewhere.[1,5]

Special investigations

Cope warned that "overreliance on laboratory tests and radiological evaluations will very often mislead the clinician, especially if the history and physical examination are less than diligent and complete."[5] Although clinical signs and history play a primary role in the diagnosis of AAP conditions, special investigations and tests can play a part in the decision-making process, particularly when the diagnosis is unclear after a careful interview and physical examination.[15] Some of the AAP decision tools that will be assessed in the systematic reviews in this study make selective use of the results of special investigations. These investigations include laboratory tests such as white blood cell count, C-reactive protein or urine dipstick tests, and imaging procedures such as radiology, ultrasound, laparoscopy or computed tomography, which are outlined and discussed in detail in many sources.[5,16,45,46] An RCT in 1998 strongly suggested that CT is highly accurate in diagnosing acute appendicitis,[6] although subsequent studies and a recent systematic review provided mixed conclusions.[47] The characteristics of these tests are briefly outlined in the Glossary.

Reference standards for diagnosis and definition of causative disease

For acute appendicitis, histopathological examination of the excised appendix is the reference standard for patients with the disease. Because it is unethical to remove the appendix from patients when acute appendicitis and other serious causes for the pain are excluded, the final diagnosis plus follow-up of discharged, non-operated patients serve as the reference standard. There is a need to follow up those discharged without operation (often labelled as having NSAP) because of the possibility of false-negative diagnoses. Some discharged patients may have been suffering from acute appendicitis or other causes of AAP (e.g. perforated peptic ulcer or acute cholecystitis), which would require emergency surgery. However, in a 1-year follow-up of patients with NSAP, two-thirds never had a firm diagnosis.[1] The rest had minor conditions. The reference standards for other conditions that require surgery also involve histopathology, with final diagnosis and follow-up serving as the standard for those without the disease. Reference standards for these and non-surgical causes of AAP are discussed in various sources.[48–50]

Dilemmas in AAP decision-making

Making accurate decisions for patients with AAP is difficult in both primary and secondary care because many conditions cause it and no single clinical finding or laboratory test is both specific and sensitive.[5] Around 50% of hospital inpatients with AAP have a non-specific cause,[51] but many of the remainder have acute appendicitis or other conditions requiring emergency hospital admission and surgery. To avoid missing these seriously ill patients, large numbers are referred for unnecessary admission and surgery, with negative laparotomy rates of up to 25%.[51] However, patient, GP and surgical delays can lead to a perforated appendix in 20% of cases.[51] Even highly experienced doctors make these mistakes.[52] What makes clinical decision-making in this area particularly challenging is the trade-off between the perforated appendix rate and the negative appendicectomy rate.[53]

A recent population-based study found that the problem of misdiagnosis of acute appendicitis and other AAP conditions requiring urgent attention has not changed over time, despite the implementation of new diagnostic technologies in recent years.[54] The authors of the study offered possible explanations: "(1) computed tomography in the United States may not be performed frequently enough or in the appropriate subpopulations to affect the rate of misdiagnosis; (2) diagnostic tests may be less accurate in a typical clinical environment than in the research setting and (3) these tests may be accurate and performed routinely but may be overruled or not reported rapidly enough to influence decision making."

As a result of such difficulties, many tools that combine two or more clinical or laboratory findings to assist clinical decisions have been developed to aid the management of AAP and other conditions.[55,56] Such tools have a long history, with the first computer aid developed in 1959.[57] Some examples of decision tools include the Alvarado score,[58] the Leeds Acute Abdominal Pain system, which estimates patient-specific probabilities of different serious conditions causing the pain,[41] and Framingham-derived scores to assess individual cardiovascular risk.[59–62] While some AAP decision tools appear to be more accurate than junior doctors, no clear consensus exists on which, if any, is most appropriate for use by UK GPs or hospital doctors.[52]

Clinical decision tools

Definition and typology of decision tools

A decision tool is an active knowledge resource that uses patient data to generate case-specific advice, which supports clinical decision-making about individual patients by health professionals, the patients themselves or others concerned about them. A typology of decision tools is shown in *Figure 2*. The typology encompasses decision support systems, checklists, clinical algorithms, computer decision aids, slide rules, nomograms, preprogrammed calculators[63] and scoring systems.[55] For the purposes of this study, the focus is on decision tools that combine two or more clinical signs, symptoms or patient characteristics. A detailed discussion of decision tools can be found in Appendix 1.

The dual role of decision tools as prognostic and diagnostic models

The reason for assessing both the accuracy and impact of decision tools in Chapters 2–5 is that they have a dual and interrelated role as both diagnostic and prognostic tools; this is one of the reasons why they are called decision tools. Assessing the performance of a decision tool or another diagnostic technology provides an important first stage in the clinical management of patients. However, an arguably more important objective is the impact that the decision tool has on (the prognosis of) patients through a doctor's choice of treatment.[64]

A main difference between the diagnostic and prognostic applications of a tool is that of the time element. As Kraemer said, "if the diagnosis is obtained during the period of testing, the test is… called a 'diagnostic test', but if it is obtained during a follow-up period" [to assess the subsequent development of disease and the impact of the tool on patient outcomes], it then becomes a 'prognostic model'."[65] The methodological aspects of assessing the accuracy and impact of decision tools have been extensively discussed in the recent literature, particularly in regard to the balance between a simple tool that can be understood by healthcare practitioners and

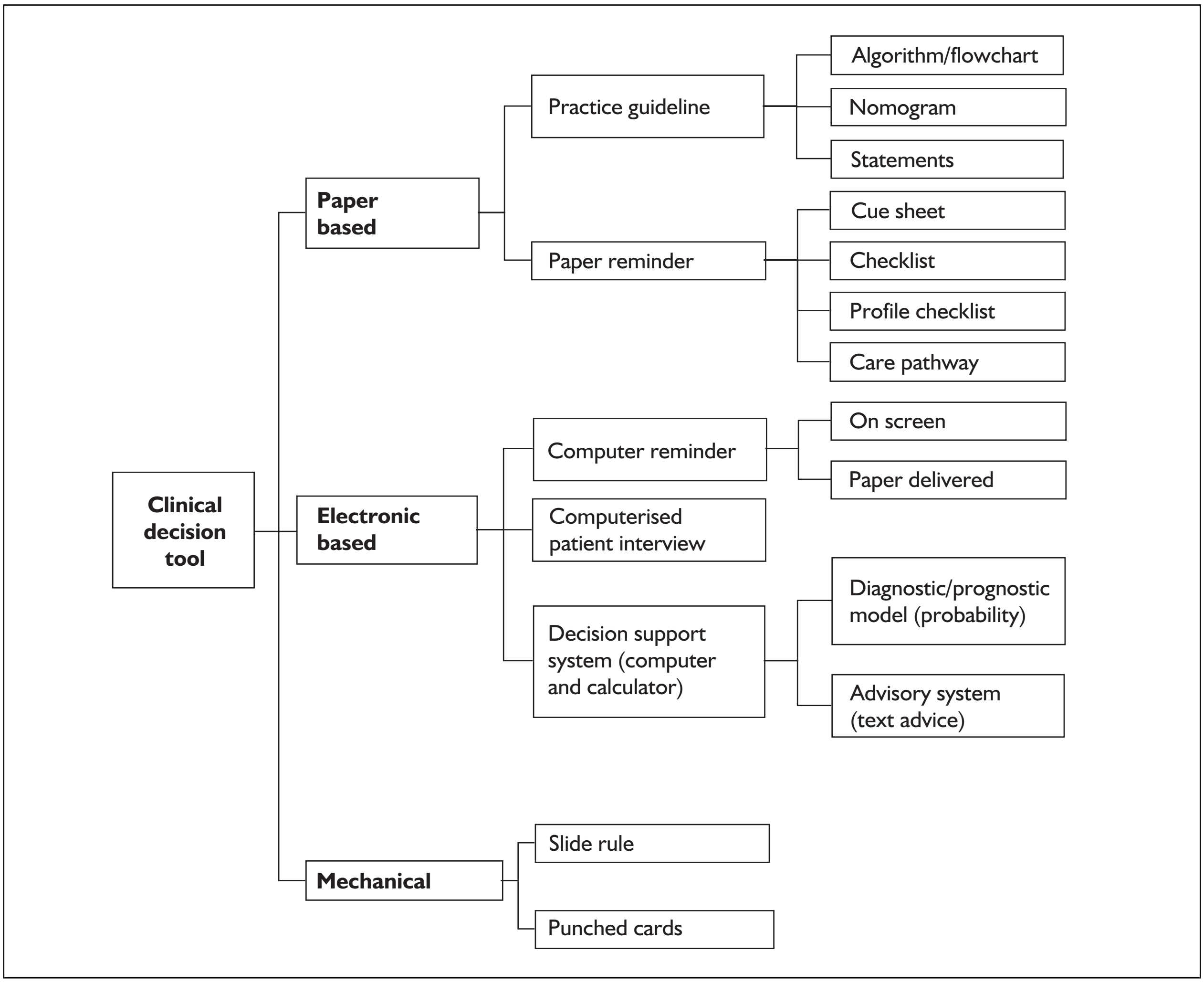

FIGURE 2 *Typology of clinical decision tools*

complex mathematical models that might be highly accurate but are a 'black box' to the user.[56,65–70] There are legal, ethical and safety questions concerning the role of the decision tool user as a learned intermediary: he or she needs to understand how the tool works to be able to make an informed decision.[71] The legal risk to doctors making uninformed decisions solely using the black box is that they breached their duty of care for patients who are harmed as a result.[72–74] A decision tool with high diagnostic accuracy may have little impact on patient outcomes for various reasons, including the black box phenomenon and poor usability as perceived by the user.[71]

A brief overview of common reasoning methods for decision tools

A brief overview of the reasoning methods underlying decision tools is given here. A more detailed exposition can be found in Appendix 2.

Bayesian methods

Bayes' theorem describes how the pretest or prior probability of disease changes as new information is taken into account; this revised probability is called the post-test probability.[75] Bayes' theorem can be extended in a simple way to combine multiple pieces of diagnostic information. The post-test probability obtained from the first test can serve as the prior probability for the next test. However, this approach (coined naïve or idiot's Bayes) has been noted to give overoptimistic predictions owing to double counting of diagnostic information when the individual test results that are being combined are not independent. In this report, methods that use simple naïve Bayesian probability updating will be referred to as Bayesian methods.

Logistic regression extensions of Bayes theorem

The problem of double counting diagnostic information has been tackled using logistic regression models, to take into account correlations between pieces of diagnostic information. The links between Bayes' theorem and logistic regression models can be fitted by re-expressing the theorem using 'weights of evidence' to account for the correlations.[76] The term logistic regression methods will be used to describe approaches that make the above adjustments.

Discrimination rules

Discrimination rules use statistical methods to produce a rule that can be used to allocate individuals to the group in which they are most likely to belong. For example, it is possible to produce predictions of group membership (diseased or non-diseased) from a logistic regression or discriminant function, categorising those with disease probability greater than some appropriate value (such as 0.5) into the diseased group and those with probabilities less than that value into the non-diseased group.

Clinical algorithms

An algorithm is a process for carrying out a complex task broken down into simple decision and action steps.[12] Clinical algorithms can be represented as paper-based flowcharts or computer programs.

Expert systems

An expert system is a computer program that simulates human thought processes "to provide the kind of problem analysis and advice that the expert might provide."[11]

Machine learning

Machine learning can be either supervised or unsupervised.[77] In the former, a system is provided with a sample of data and instructions on how to identify and classify patterns within the data by a trainer. In the latter, a system is also provided with data, but is left to identify patterns without external assistance. There are various types of machine learning methods, such as neural networks and genetic algorithms.

Study questions

The above discussion demonstrates the need to evaluate AAP decision tools. Thus, the overall study objective is to assess whether, and by how much, clinical decision tools improve the clinical management of AAP and whether any particular tool can be recommended. To achieve this objective, the following specific questions are addressed.

Study question 1: What are the diagnostic accuracies of decision tools and doctors aided by decision tools compared with unaided doctors in patients with AAP? (Chapters 2 and 3.)
Study question 2: What are the impacts of providing doctors with AAP decision tools on patient outcomes, clinical decisions and actions? (Chapters 4 and 5.)
Study question 3: What factors are likely to determine the usage rates and usability of each AAP decision tool? (Chapter 6.)
Study question 4: What are the associated costs and likely cost-effectiveness of these decision tools in routine use in the UK? (Chapter 7.)

To make this study manageable, the focus is on acute appendicitis, since the most important clinical decision when patients present with AAP is whether they might have appendicitis and whether surgery is required. The authors believe that the focus on acute appendicitis is justifiable. If one examines the ten most common causes of AAP, acute appendicitis is the one that inevitably requires a rapid diagnosis and rapid decision as to whether to operate immediately.[1,5] The other conditions requiring emergency surgery are perforated peptic ulcer and intestinal obstruction; however, these conditions are now rare compared with acute appendicitis, or are unlikely to be confused with appendicitis because of differing clinical features.[1,5,14,29] Emergency surgery is rare nowadays for acute cholecystitis, pancreatitis and diverticulitis. For patients with suspected cholecystitis, for example, the standard practice nowadays is to use "fasting, intravenous fluids and analgesia as initial treatments."[14] Similarly, emergency surgery is not the typical initial treatment for pancreatitis and diverticulitis.[1] Therefore, the most important decision facing the doctor is whether the AAP patient is suffering from acute appendicitis or other conditions, including NSAP. The restriction to acute appendicitis is an appropriate reflection of clinical problems in the NHS, based on the authors' communications with experienced surgical consultants at Whipps Cross Hospital, London, who participated in the original De Dombal studies. The overall scope of this Review of Abdominal Pain tolls (RAPT) is shown in *Figure 3*.

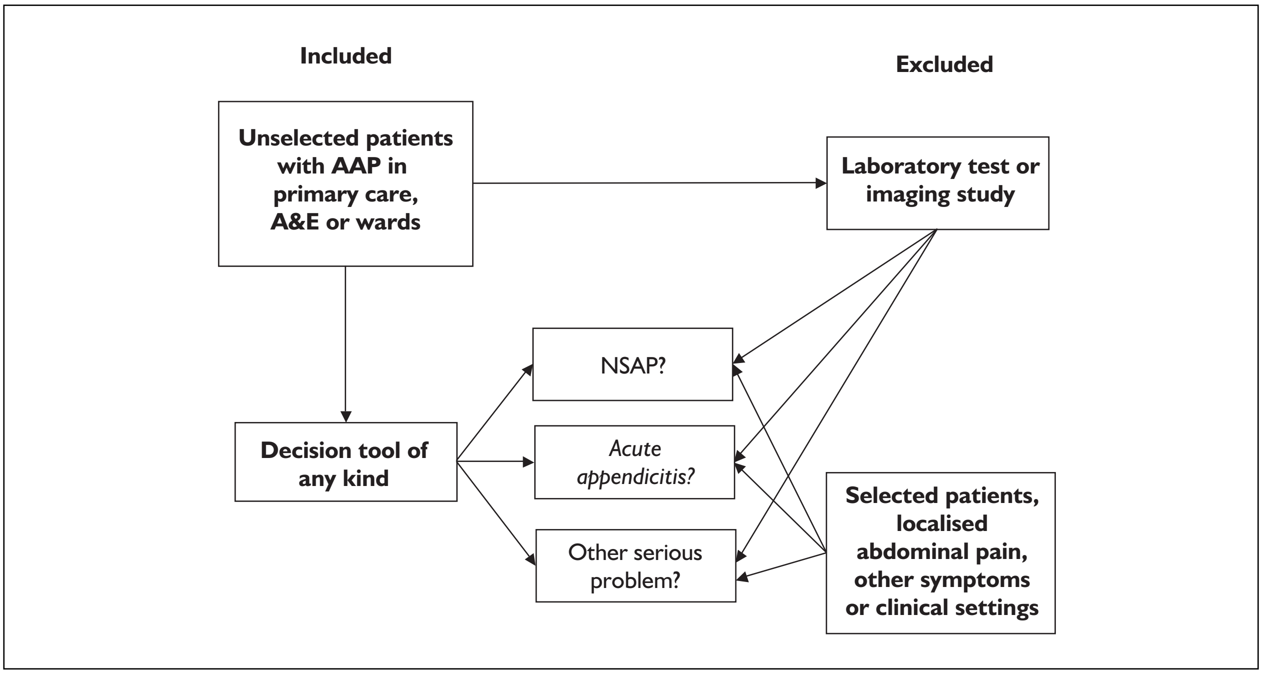
Included
Excluded
Unselected patients with AAP in primary care, A&E or wards
Laboratory test or imaging study
NSAP?
Decision tool of any kind
Acute appendicitis?
Other serious problem?
Selected patients, localised abdominal pain, other symptoms or clinical settings

FIGURE 3 *Scope of the RAPT review. To make the study manageable, the focus is on acute appendicitis.*

Chapter 2

Study question 1. Methods for the systematic review of accuracy studies

Search methods

Sources of studies

The following electronic databases were searched: MEDLINE (1966 to June 2003), EMBASE (1980 to June 2003), CENTRAL (The Cochrane Central Register of Controlled Trials (Issue 3 2003, Cochrane Library), CINAHL (Cumulative Index of Nursing and Allied Health) (1982 to June 2003), INSPEC (database provided by the Institute of Electrical Engineers, with literature on physics, electronics and computing) (1969 to June 2003), SIGLE (System for Information on Grey Literature in Europe) (1980 to June 2003) and HEALTH-CD (Database of full-text material from UK Department of Health and the Stationery Office containing British health management reports, grey literature). Searches were conducted from the earliest date of titles or abstracts available for each database to the latest title(s) or abstracts available as of 1 July 2003.

Further studies were located through citation searches of major papers introducing the tools and by checking the reference lists in primary and review articles retrieved from the database searches. Because the translation of foreign-language papers was not included in the original funding proposal, only English-language papers were included.

Search strategies

The sensitivity of searches was maximised by using both free text and controlled vocabulary terms, such as:

- specific target disorder or organ names (e.g. “appendi*” and “salpingitis”)
- the MeSH terms used to index the above papers and others that are found
- additional search terms to detect studies of decision tools (e.g. “decision support system”, “algorithm” and “scoring system”) and diagnostic accuracy (e.g. “specificit*”, “sensitivit*” and “likelihood ratio*”).[78–80]

A pilot MEDLINE search using this approach identified 3254 studies. Similar strategies (with additional or alternative search terms as appropriate) were adapted for the other databases. Appendix 3 contains the search terms used for each database.

Assessment of eligibility

Inclusion and exclusion criteria

Eligible studies

Studies of appropriate clinical cohorts that report the accuracies of AAP decision tools and/or the accuracies of doctors aided by decision tools were included. In regard to the former, reference is to the estimates of diagnostic accuracy provided by the decision tool itself. In regard to the accuracies of doctors aided by decision tools, reference is to the diagnostic accuracy of the doctor with access to the decision tool or its output. These are two distinct types of studies, and they will be referred to as the accuracy of the decision tool (or decision tool accuracy) and aided doctors’ diagnosis (or even aided doctors) for short. This review focuses on unselected patients with AAP recruited consecutively or randomly sampled from an appropriate cohort, that is, those typically seen by physicians in practice. For the purposes of this review, case–control studies or cohort investigations that studied patients who all had an operation for suspected appendicitis were excluded because of the likely referral bias in the evidence from such studies that prevents their application to clinical practice.[78,81]

Eligible patients

Patients with a main complaint of previously undiagnosed acute generalised, upper or lower abdominal pain lasting for not more than 7 days from onset were eligible for inclusion.[1]

Eligible investigations

Studies evaluating the accuracies of AAP decision tools and/or the accuracies of doctors aided by decision tools, compared with unaided doctors’ decisions, were included. Studies that reported only the accuracies of AAP decision tools and/or aided doctors were included, although such studies are less useful, because they cannot assess the relative accuracy of the decision tool or aided doctors and the unaided doctors. Studies of individual laboratory or imaging investigations, audit and feedback, continuing education activities and telemedicine were excluded. Studies in which the authors did not specify the reference standard were excluded. Without a reference standard, the accuracy of the decision tool cannot be assessed. The following reference standards were eligible:

- histopathology for those with the target disorder and final diagnosis for those without, with post-discharge follow-up of the latter
- histopathology for those with the target disorder and those without
- final diagnosis for both those with and without the target disorder (with standard criteria)
- final diagnosis for both those with and those without the target disorder (not standardised or unclear criteria).

Eligible study measures

Studies that reported sensitivity and specificity, likelihood ratios for positive and negative test results, area under the receiver operating characteristic (ROC) curve, or data that would enable these measures to be calculated, were included. Studies that reported crude accuracies only (i.e. the proportion of diagnoses which are correct) were excluded. Crude accuracy is not a useful measure because it entangles the accuracy of the decision tool among those with the target disorder and the accuracy among those without.

Procedures for assessing eligibility

The primary search aimed to identify all published studies that passed the inclusion/exclusion criteria outlined above, with an initial classification of all search results (through scrutiny of titles and abstracts) by one reviewer into:

- studies that were obviously irrelevant
- studies that were potentially or definitely relevant.

The hard copies of original articles for all studies in the latter category were obtained for more detailed review by a reviewer (JLYL). To assess the reliability of the above process, a second reviewer (JCW) independently categorised a sample of the full search results.

An eligibility criteria form (Appendix 4), which incorporated the criteria outlined in the section ‘Inclusion and exclusion criteria’ (p. 9), was developed, with reference to forms used in another HTA-funded systematic review.[82] One reviewer (JLYL) assessed the eligibility of all the retrieved studies for detailed data extraction. To assess the reliability of this process, a second reviewer (JCW) independently repeated the eligibility check for a sample of the retrieved studies. Disagreements were resolved through discussion between the two reviewers (JLYL and JCW). A third reviewer (JJD) was available to help to resolve the discrepancies in consultation with the other two reviewers. The extent of disagreements was quantified using the kappa statistic.[83]

Data extraction

A data collection form for extracting information from eligible studies was developed by adapting forms used in another review,[82] the methods recommended by the Cochrane Methods Working Group on Systematic Reviews of Screening and Diagnostic Tests, and the items in the Standards for Reporting of Diagnostic Accuracy (STARD) Initiative.[84–87] Data from each eligible study were extracted on to the form. The following data were recorded from each study:

- patients’ details, including demographic characteristics (age, gender, ethnicity), and medical condition(s) causing AAP
- the reasoning method used by each decision tool
- healthcare setting (ward admission, surgical department, A&E, other secondary care, primary care)
- estimates of the accuracy of decision tools and unaided doctors’ decisions
- indicators of methodological quality (see below).

A copy of the data collection form can be found in Appendix 4.

Methodological quality

Two reviewers (JLYL and JCW) from the project steering group extracted data and assessed the quality of all studies selected for inclusion in the review. Where disagreements continued after discussion between the two reviewers, a third reviewer (JJD) was available to help to resolve the discrepancies. The criteria below were used to assess study quality:

- Reference standard: what was the reference standard for those with the target disorder? Was it the same for everyone? What was the

reference standard for those without the target disorder? Was it the same for everyone? If the reference standard was not the same for all participating patients, then the accuracy of the decision tool could be distorted.
- Incorporation bias: did the reference standard exclude the output of decision tools or the unaided decisions of doctors and vice versa? Incorporation bias occurs when the decision tool uses signs, symptoms or tests that are part of the reference standard, leading to an overestimate of the test's performance.[88]
- Blinding: was the person allocating the reference standard blind to the decision tool results? Was the decision tool user blind to the reference standard?
- Verification/work-up bias: was there an attempt to compare all results of decision tools, aided doctors or unaided doctors to a reference standard and vice versa? Have all results been compared to the same standard? Verification bias occurs when the results provided by a decision tool affect the decision on whether to conduct the reference standard investigation.[78,89] This often happens when the decision tool results are negative, showing that a patient is unlikely to have the target disorder in question. There are two types of verification bias: 'partial verification', which occurs when the reference standard was not applied to all participating patients and 'differential verification', which occurs when results from a decision tool influence the choice of reference standards to apply. Estimates of sensitivity tend to be overestimated when partial verification bias is present, while estimates of specificity tend to be understated. Both sensitivity and specificity tend to be overstated when differential verification bias is present.
- Selection of the study sample: was a consecutive or random selection of cases sampled? Was a single relevant clinical population selected (as opposed to positive and negative groups)? The use of a single clinical population in a prospective or retrospective cohort design is more valid than the use of case–control designs, which tend markedly to overestimate accuracy. The results of a cohort study depend on the case-mix that it recruits, and may be quite different from the 'average' performance.
- Subgroups: were subgroups analysed separately? Were they prospectively defined?
- Completeness: how complete was the data set? Relevant information includes the number of patients originally considered for inclusion, the number of eligible patients, the number of patients included at the start, and the number of patients lost. A low completeness rate could result in bias.
- Indeterminate results: how were indeterminate scores and outputs of decision tools handled in the analysis? Excluding indeterminate results from the analysis could exaggerate the accuracy of the tool.
- treatment paradox: were patients treated for their AAP before the decision tools and the reference standard were employed? For example, if a decision tool indicates that a patient has a target disorder and he or she is successfully treated before being tested for a second time using the reference standard, misclassification bias would result. This is called the 'treatment paradox'.[78]
- Type of study: were the data collected retrospectively from case record reviews or prospectively?

Details on justifications for the above quality criteria can be found in other sources.[78,84,85]

Potential sources of heterogeneity

When sufficient data were available, the meta-analysis for each comparison was stratified by these factors:

- age groups and gender
- prevalence of acute appendicitis
- healthcare setting (ward, surgical department, A&E, other secondary care, primary care)
- type of decision tool studied (i.e. reasoning method of decision tool)

When significant heterogeneity was found to be present the following explanations were also considered:

- year of the study
- type of data set (e.g. prospective test set in a different centre from where the tool was developed, split-sample test set, jack-knife, training set data)
- number of data items used by the decision tool
- was the evaluator of the tool also its developer?
- follow-up of non-admitted, non-operated (defined as those who were admitted but not operated on) or postdischarge cases (defined as those who were admitted and operated on) (e.g. by telephone at 30 days)
- data collection method (prospectively or retrospectively)

- completeness of data as reflected by the sampling rate (number of subjects included/number of eligible subjects)
- training and seniority of the decision maker.

Statistical methods

Analytical techniques that are most relevant for the purposes of this systematic review of diagnostic accuracy are presented here.[78,90,91] The following software packages were used: STATA version 8.2 for constructing pooled error rate ratios, forest plots, scatterplots, metaregression models and basic univariate/bivariate statistics; S Plus 2000 Professional to construct summary ROC curves (explained in the next section) and forest plots to summarise graphically the sensitivities and specificities from individual studies; and CIA version 2.1 to compute confidence intervals.

Data synthesis

Pooling sensitivities and specificities

Pooled estimates of sensitivity and specificity can be obtained by considering the two measures as simple proportions $p_i = \frac{y_i}{n_i}$, and computing a weighted average of sensitivity or specificity of all studies.[78,92]

Pooling is reasonable as long as (1) sensitivity and specificity are not negatively correlated with each other, as the diagnostic threshold changes across studies, and (2) no heterogeneity exists between study-specific estimates of accuracy. In practice, conditions (1) and (2) often do not hold, so pooling is inappropriate, as explained elsewhere.[78,91]

Standard χ^2 tests were used to assess the extent to which accuracy data from the primary studies deviate from homogeneity. Correlation coefficients (e.g. Pearson's r for normally distributed data and Spearman's ρ for non-normally distributed data) were computed to assess the relation between sensitivities and specificities. A significant negative correlation signals a possible diagnostic threshold effect.

Among studies for which the diagnostic accuracies of decision tools/aided doctors and unaided doctors were compared, ratios of false-positive and false-negative error rates were used to provide estimates of the difference between the diagnostic accuracies of the decision tool/aided doctors and unaided doctors. A ratio for decision tools/aided doctors versus unaided doctors of less than 1 indicates that the tool or aided doctor has a lower false-negative rate or false-positive rate than the unaided doctor. A ratio of greater than one indicates that the decision tool/aided doctor has a higher false-negative rate or false positive rate than the unaided doctor. A ratio of unity indicates no difference. Pooled error rate ratios for the false-negative rates (1 – sensitivity) and false-positive rates (1 – specificity) were computed using the random effects model.[93] The presence of heterogeneity was tested using standard methods.[94] Sources of heterogeneity were investigated.[92,95–97]

Another way to illustrate graphically the presence (or absence) of heterogeneity of paired accuracies of decision tools (or aided doctors) and unaided doctors is to use an ROC plot, with study-specific data points for decision tools and doctors joined by a line. The magnitude of the difference between the data point for the decision tool (or the aided doctor) and data point for the unaided doctor is shown by the vertical and horizontal distances between the study-specific data points (corresponding, respectively, to differences in sensitivities and specificities between decision tools/aided doctors and unaided doctors).

The diagnostic odds ratio and other measures of diagnostic accuracy

The diagnostic odds ratio (DOR) is another summary measure of diagnostic accuracy; it is the ratio of the odds of a positive test result in a patient with the target disorder to the odds of a positive test result in a patient without the target disorder.[78] The DOR combines sensitivities and specificities (as well as likelihood ratios) into one measure of diagnostic accuracy.[78] It provides an assessment of how well a decision tool or doctor performs in distinguishing healthy from unhealthy patients. The bigger the DOR, the better the diagnostic accuracy. For many clinical applications, similar DORs may be observed in studies that use different cut-points, a useful property for the purposes of a systematic review (i.e. to combine studies). However, it can be difficult to interpret clinically. A decision tool with high specificity and low sensitivity could have the same DOR as one with low specificity and high sensitivity. *Table 2* provides examples of DORs for various sensitivities and specificities.

Summary ROC curves

An ROC curve (*Figure 4*) provides a graphical way to examine the relation between sensitivity and specificity for a decision tool.[98] It is a plot of the sensitivity (or the true-positive rate among those with the target disorder) against the complement

TABLE 2 *Examples of DORs and typical sensitivities and specificities*

DOR	Sensitivity	Specificity
9800	0.99	0.99
1880	0.99	0.95
1880	0.95	0.99
171	0.95	0.90
171	0.90	0.95
44	0.95	0.70
16	0.80	0.80
5.4	0.70	0.70
1.0	0.50	0.50

Source: adapted from Deeks.[78]

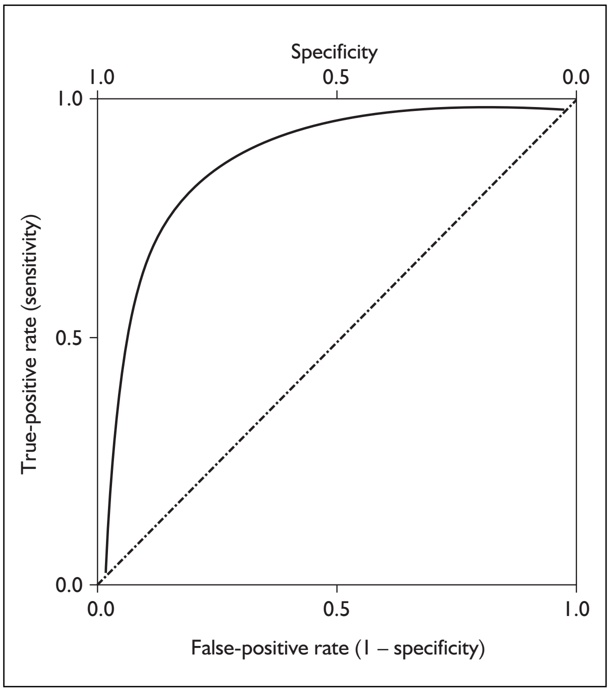

FIGURE 4 *ROC curve*

of the specificity (1 – specificity or the false-positive rate among those without target disorder) at different cut-points or diagnostic thresholds. ROC curves for diagnostic tools with excellent performance will have points that are close to the graph's top left corner, where the sensitivity and specificity are 100%. ROC curves for tools with poorer performance will lie more closely to the diagonal in *Figure 4*. The accuracy of a tool can also be summarised as the area under the receiver operating characteristic curve (AUROC curve), which ranges from less than 0.5 to 1.0.[99] An AUROC curve value of 1.0 means that the tool can perfectly discriminate between patients with and without the target disorder. An area of 0.50 implies that the tool is unable to discriminate patients with the target disorder from those without.[65]

The pooling of ROC curves from primary studies is usually impractical, because most studies report only a single point (i.e. sensitivity and specificity at a given threshold).[91,100] An alternative method of summarising diagnostic accuracy is therefore needed. One such method, proposed by Littenberg and Moses, fits a summary ROC (SROC) curve through sensitivity and specificity points plotted in ROC space, one point being plotted for each primary study in the meta-analysis.[91]

It is mathematically convenient to represent the SROC curve as a linear model (a statistical model in which the relation between the dependent variable y and the independent variable x can be fitted to the data using the equation, $y = \alpha + \beta x$, where α and β are constant terms). Littenberg and Moses[78,91] proposed fitting the following regression model:

$$D = a + bS \qquad (1)$$

where D is the natural logarithm of the DOR and S is the natural logarithm of the product of the odds of true-positive test results and the odds of false-positive test results. D is a summary measure of the diagnostic accuracy of the test. When a diagnostic threshold decreases (increases), the frequency of positive diagnoses (including both true and false positives) increases (decreases), which means that the product of the odds of positive test results and S also increases (decreases). S is thus a summary measure of diagnostic threshold. Equation (1) expresses how the diagnostic accuracy of a test changes with variations in the diagnostic threshold or other predictors of diagnostic accuracy. In this systematic review, weighted least squares regression was used, with the sample sizes of the primary studies used as weights.[91] To obtain an SROC plot, the regression model in equation (1) can be back-transformed.[78] The statistical significance of the regression coefficient b is tested to assess whether diagnostic accuracy varies significantly with changes in threshold. If such a relation exists, the SROC curve is asymmetrical in shape.[78]

If there is no evidence that diagnostic accuracy varies with changes in threshold, it means that the DOR stays constant, whatever the threshold. The SROC curve would then be symmetrical in shape.[70] The implication is that the usual methods for pooling odds ratios in a meta-analysis can be applied to obtain a pooled DOR.[78]

Assessment of heterogeneity using metaregression

The DOR may vary between primary studies for reasons other than the diagnostic threshold, such as heterogeneity in study design, patient subgroups and other effect modifiers, or differences in methodological quality between studies.[78,90,91,101] If the DOR varies with diagnostic threshold (i.e. the log DOR is associated with the measure of threshold S and the SROC curve is asymmetrical), potential effect modifiers or indicators of methodological quality (outlined in the section 'Potential sources of heterogeneity', p. 11) can be added as additional independent variables to the regression model in equation (1) (e.g. $D = a + bS + c \times \text{Prevalence}$).

If S is not statistically significant (i.e. the null hypothesis that the regression coefficient $b = 0$ fails to be rejected), it suggests the DOR does not change with diagnostic threshold (i.e. the SROC curve is symmetrical). Standard metaregression techniques for odds ratios can then be used to explore potential sources of heterogeneity, using the log DOR as the dependent variable and each potential effect modifier or indicator of methodological quality as independent variables (e.g. $D = a + c \times \text{Prevalence}$). There is no need to include S as a covariate since there is insufficient evidence to indicate that it is associated with D.

One of the limitations of metaregression is that a systematic review typically only includes a small number of eligible studies. Therefore, only one effect modifier was fitted at a time to avoid problems of overfitting. The metaregression models were also weighted by total study sample size.

The antilog of the regression coefficient for each effect modifier can be interpreted as a measure of the relative increase in the DOR that can be attributed to that effect modifier. This measure is also known as the relative diagnostic odds ratio (RDOR). The RDOR is a measure of differences in observed diagnostic ability between two tests, groups or types of study. It is calculated by dividing the DOR from one by the DOR from the other. The magnitude of bias or difference is reflected by the extent to which the ratio deviates from 1.

Effect modifiers included in metaregression

Effect modifiers for decision tools

The potential effect modifiers below were included in the metaregression models for decision tools. Some categories within each effect modifier were merged if the numbers of studies falling in them were too small.

- **Year of publication (as a surrogate for the year of the study):** when a decision tool first appears in the published literature, it may initially demonstrate very high accuracy. However, when additional studies are conducted over the course of time and the tool is scrutinised (e.g. in different settings and by independent evaluators), its accuracy may decline.
- **Prevalence of acute appendicitis:** the diagnostic accuracy (e.g. as measured by sensitivity and specificity) of a decision tool is known to vary with the prevalence of the target disorder.[89,98] The spectrum of patients may be different in populations with low and high prevalences of a target disorder.
- **Reasoning method of decision tool (bayesian methods, logistic regression methods, discriminant rules, clinical algorithms, expert systems, neural networks, other method):** this was included to compare the accuracies of different methods used to develop the decision tool.
- **Type of data set (prospective test-set data in different centre, prospective test-set data in same centre, non-random split sample used as test set, random split sample used as test set, resampled data used as test set, training-set data used as test set, and training-set data):** It has been argued that decision tools demonstrating high accuracy during the development phase (when 'training-set' data are used) tend to become less accurate when evaluated using postdevelopment 'test-set' data.[56,102,103] This could happen for a number of reasons (e.g. the often-small sample sizes used in training-set data to develop the tool, and model 'overfitting' during the development phase, when training-set data are used to fine-tune a tool to attain optimal performance).[102,103] The type of test-set data used to evaluate a decision tool after development can influence its estimated diagnostic/prognostic accuracy.[56]

The main types of test-set data,[104] in order of their rigour, are: (a) prospective data collected from one or more centres away from where the tool was developed (best); (b) prospective data collected in the centre where the tool was developed; (c) non-random split samples; (d) random split samples or samples generated using a resampling technique (such as bootstrapping or jack-knife); and (e) training-set data used also as test-set data (worst).

Option (a) is generally considered the best, particularly if the sample is selected from an environment that is typical of its intended use (e.g. in terms of the patient case-mix encountered and the users of the decision tool). Option (b) is inferior to option (a) because by collecting data from the centre where the tool was developed, it is more difficult to generalise findings. There is also more scope for bias. Option (c) refers to the non-random splitting of a sample into two subsamples, one for the training-set study and the other for the test-set study. It is prone to the same problems as option (b). However, a non-random split sample is better than a random split sample (option d), because the latter tends to produce samples with similar characteristics to the original sample. The random split sample goes against an important purpose of a test-set study, which is to assess performance in a sample with different characteristics. In a resampling procedure such as jack-knife (option d), in which a patient is taken from a sample, the model is constructed from the rest of the sample, the withdrawn patient is used as a case, and then another patient is withdrawn to repeat the procedure.[103,105]

- **Number of data items used by the decision tool:** if a particular decision tool requires doctors to record or input a very large number of data items, they may give up and not use the decision tool. Thus, doctors provided with this tool may perform no better than those who were not given the tool. However, a decision tool that uses a very small number of data items may not have made optimal use of the available information.
- **User of the decision tool (researcher, junior doctor, senior doctor):** the influence of the decision tool on diagnostic accuracy may vary according to the user. An experienced senior doctor may diagnose an AAP patient as or more accurately than a decision tool, but a junior doctor's performance may be worse.
- **Independence of evaluator:** investigators who evaluate tools that they developed may be favourably predisposed towards their performance. Independent evaluators are likely to be more objective.

Effect modifiers for unaided doctors' diagnosis

The following potential effect modifiers were included in the metaregression analyses for unaided doctors' diagnosis:

- **Year of publication:** with improvements in clinical practice and training of doctors, unaided doctors' diagnosis may be more accurate in later studies than in earlier ones. Changes in the case-mix of AAP conditions and the prevalence of acute appendicitis over time may also influence doctors' diagnostic accuracy.
- **Prevalence of acute appendicitis:** prevalence could also influence the sensitivity and/or specificity of a doctor's unaided diagnosis for the same reasons as those mentioned for decision tools.
- **Seniority of doctor making initial diagnosis:** consultant surgeons may have superior diagnostic accuracies to junior doctors.
- **Independence of evaluator:** an investigator who evaluates a tool that he or she had developed may be favourably predisposed towards its performance and may recruit a control group of doctors who are less experienced than in a typical clinical environment. This may exaggerate the performance of the decision tool in comparison to the unaided doctor. An independent evaluator is likely to be more objective.

Assessment of quality of study methods and reporting

Indicators of methodological quality were entered as covariates (one at a time) to Littenberg and Moses' regression model to assess the effects of study quality on the log DOR.[91] The following indicators were included in the analysis: (i) reporting of patient characteristics (age and gender distribution), whether non-operated cases were followed up, quality of the reference standard, potential for incorporation bias, whether the reference standard was allocated blind to the decision tool results, whether the decision tool user was blind to the reference standard, potential for differential verification bias, potential for partial verification bias, completeness of data, patient recruitment (consecutive recruitment versus random or representative sample), type of study (prospective versus retrospective), presence of treatment paradox, and methods of treating indeterminate outputs from the decision tool. Detailed definitions for the quality indicators used in the data analysis are given in Appendix 5.

This approach to assessing the effects of quality indicators is superior to the use of composite quality scores, the problems of which have been widely reported both empirically and methodologically.[106–109] The possibility of publication bias was assessed by a funnel plot.[110,111]

Chapter 3

Study question 1. Results of systematic reviews of accuracy studies

Studies included in the review

Thirty-four studies from 27 papers (out of more than 25,000 abstracts/titles screened) were found to be eligible for inclusion in the final review (Appendix 6 and *Table 3*). The flowchart in *Figure 5* depicts the number of papers that were screened from the various databases (more than 32,000 abstracts/titles before duplicates from different databases were identified and removed and more than 25,000 abstracts/titles screened for potential eligibility), the number possibly eligible and the number of papers that were read in full, and applies for study questions 1 and 2. The set of potentially eligible papers for the reviews (n = 489) was also used for study questions 3 and 4. Usage rates and data possibly useful for an economic evaluation were extracted from papers that reported them.

A total of 1462 abstracts was checked by both reviewers. The kappa statistic was 0.82 (95% CI 0.77 to 0.87), indicating good agreement. The rest of the titles and abstracts were checked by JLYL. Among the retrieved papers, 100 were checked by both JLYL and JCW for eligibility using an eligibility criteria form. The kappa statistic was 0.79. The rest of the retrieved papers were checked by JLYL.

Studies excluded from the review

Appendix 7 contains a summary table that lists the reasons for excluding retrieved studies for which data extraction was attempted. The most common reasons for exclusion were the inadequate reporting of results, such as crude accuracy rates only without providing data that would enable sensitivity and specificity to be computed.

Characteristics and development of decision tools for AAP

The AAP decision tools that were identified as eligible for this study used several different types of models: naïve Bayesian scores, logistic regression models, scores derived from discrimination rules and neural networks. The reasoning underlying these tools is discussed in Chapter 1 and Appendix 2. Some characteristics and developmental aspects of the 20 decision tools found to be eligible for the systematic review are first introduced.

Naïve Bayesian methods

The Leeds AAP system is a simple and widely cited Bayesian computer decision support system developed in the 1970s, but has not been used in clinical practice. The original 1972 study demonstrated very high crude accuracies, sensitivity and specificity.[113] The Leeds system has run on various platforms, from Apple computers in the 1970s to the latest personal digital assistants (PDAs). In its early years, turnaround time was a problem, since the system required the collection of 36 items on a data collection form (e.g. patient history items including gender, age, site at onset of pain, duration and severity, and physical examination items including temperature, blood pressure, tenderness of abdomen, rigidity and bowel sounds).[138] Relevant information frequently did not reach the decision-makers in time, so the actual impact of the system in practice was in question. Turnaround time ranged from a few to 20 minutes.[138,139] Compliance rates among doctors using the tool were low in some centres because of the system's slowness and complexity. This may change with the latest PDA prototype. The Leeds system uses naïve Bayesian methods to assess each new patient's information and provides a list of possible diagnoses with their estimated probabilities.[113,119,123–125,134,136] Other naïve Bayesian models were also found to be eligible for the review.[114,135]

Logistic regression extensions of Bayes' theorem

As mentioned in Chapter 1, logistic regression extensions of Bayes' theorem have been used to take into account correlations between pieces of diagnostic information. Bayes' theorem is re-expressed by logistic regression models that use weights of evidence to account for the correlations.[76] Several logistic regression models were found to be eligible for the review.[115,116,118,126,137]

TABLE 3 *Studies included in the accuracy review*

Study	Method used	Prevalence of appendicitis (%)	No. of items in tool	Evaluator evaluated own tool?	Type of data set	Type of accuracy study
Alvarado, 1986[58]	Alvarado score	82	8	Yes	Training-set data	DT
Bond, 1990[112]	Alvarado score (modified)	61	7	No	Prospective data set in different centre	DT
Bond, 1990[112]	Discrimination rule	61	7	Yes	Training-set data	DT
De Dombal, 1972[113]	Bayesian (Leeds)	28	36	Yes	Prospective data set in same centre	DT
Edwards, 1986[114]	Bayesian (Sheffield)	31	35	Yes	Prospective data set in different centre	DT
Eskelinen, 1992[115]	Logistic regression	20	22	Yes	Training-set data	DT
Fenyo, 1987[116]	Logistic regression	31	19	Yes	Prospective data set in same centre	DT
Hallan, 1997[117]	Logistic regression	38	6	Yes	Split sample	DT
Hallan, 1997[117]	Alvarado score (modified)	38	8	No	Prospective data set in different centre	DT
Hallan, 1997[118]	Logistic regression (AUROC only)	37	6	Yes	Split sample	DT
Horrocks, 1976[119]	Bayesian (Leeds)	33	36	Yes	Prospective data set in different centre	DT
Izbicki, 1992[120]	Discrimination rule	36	7	Yes	Training-set data	DT
Jahn, 1997[121]	Discrimination rule	42	11	Yes	Training-set data	DT
Jawaid, 1999[122]	Logistic regression	34	10	Yes	Prospective data set in same centre	DT
Kirkeby, 1987[123]	Bayesian (Leeds)	23	36	No	Prospective data set in different centre	DT
Kraemer, 1993[124]	Bayesian (Leeds)	17	36	No	Prospective data set in different centre	DT
Leaper, 1972[125]	Bayesian (Leeds)	26	36	Yes	Prospective data set in same centre	DT
Leaper, 1972[125]	Bayesian (Leeds)	26	35	Yes	Prospective data set in same centre	DT
Lindberg, 1988[126]	Logistic regression	28	10	Yes	Prospective data set in same centre	DT
Macklin, 1997[127]	Alvarado score (modified)	32	7	No	Prospective data set in different centre	DT
Malik, 1998[128]	Alvarado score (modified)	78	7	No	Prospective data set in different centre	DT
Ohmann, 1999[129]	Discrimination rule	21	8	Yes	Prospective data set in same centre	Doctor aided with DT
Owen, 1992[130]	Alvarado score (modified)	58	8	Yes	Prospective data set in same centre	DT
Pesonen, 1996[131]	Neural network: ART1	20	17	No	Split sample	DT
Pesonen, 1996[131]	Neural network: SOM	20	17	No	Split sample	DT
Pesonen, 1996[131]	Neural network: LVQ	20	17	No	Split sample	DT
Pesonen, 1996[131]	Neural network: BP	20	17	No	Split sample	DT
Pesonen, 1997[132]	Discrimination rule (AUROC only)	NR	9	Yes	Split sample	DT
Saidi, 2000[133]	Alvarado score	35	8	No	Prospective data set in different centre	DT
Staniland, 1980[134]	Bayesian (Leeds)	33	36	Yes	Prospective data set in different centre	DT
Sutton, 1989[135]	Bayesian (CADA-1)	15	NR	Yes	Prospective data set in different centre	DT
Wellwood, 1992[136]	Bayesian (Leeds)	6	36	No	Prospective data set in different centre	DT
Wellwood, 1992[136]	Bayesian (Leeds)	6	36	No	Prospective data set in different centre	Doctor aided with DT
Wong, 1994[137]	Logistic regression	18	8	Yes	Prospective data set in same centre	DT

ART, adaptive resonance theory; BP, back propagation; CADA, Computer Assisted Diagnostic and Audit; DT, decision tool; LVQ, learning vector quantisation; NR, not reported; SOM, self-organising map.

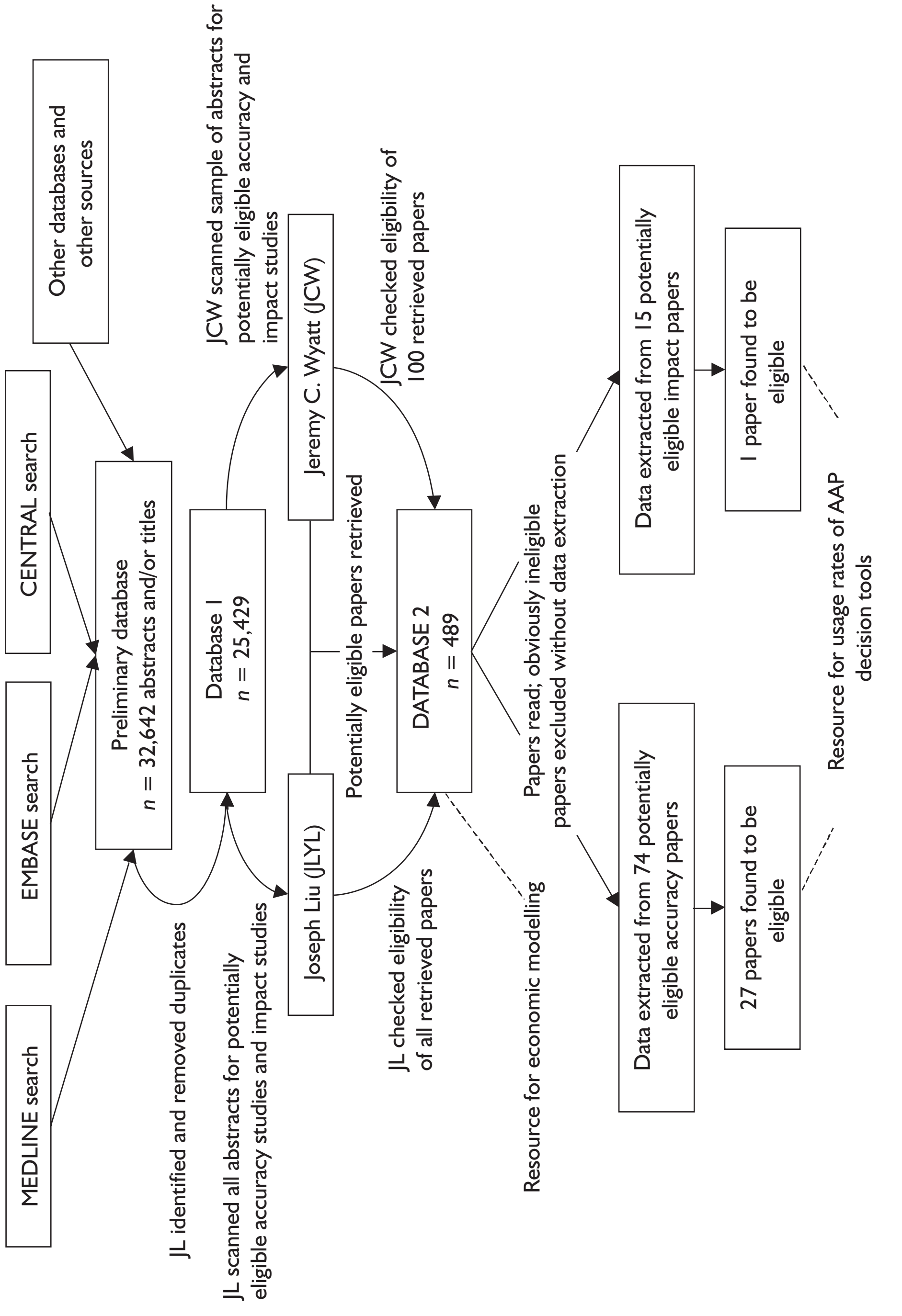
MEDLINE search
EMBASE search
CENTRAL search
Other databases and other sources
Preliminary database
n = 32,642 abstracts and/or titles
JL identified and removed duplicates
Database 1
n = 25,429
JL scanned all abstracts for potentially eligible accuracy studies and impact studies
JCW scanned sample of abstracts for potentially eligible accuracy and impact studies
Joseph Liu (JLYL)
Jeremy C. Wyatt (JCW)
Potentially eligible papers retrieved
JL checked eligibility of all retrieved papers
JCW checked eligibility of 100 retrieved papers
DATABASE 2
n = 489
Resource for economic modelling
Papers read; obviously ineligible papers excluded without data extraction
Data extracted from 74 potentially eligible accuracy papers
Data extracted from 15 potentially eligible impact papers
27 papers found to be eligible
1 paper found to be eligible
Resource for usage rates of AAP decision tools

***FIGURE 5** Search strategies and identification of eligible papers*

Checklists

The Alvarado score is a checklist of eight items (three symptoms, three signs and two laboratory tests), from which a score is computed.[58] The training-set sample consisted of 305 hospitalised patients in Philadelphia. A diagnostic weight was computed for each item, obtained by dividing the total number of patients by the number of true-positive or true-negative results. Eight items were found to be useful diagnostically, with the more important items given a value of 2 and the less important ones given a value of 1. In some evaluation studies, left shift of neutrophil maturation was not available and a nine-point modified Alvarado score was used.[127,128] Alvarado did not report the amount of time needed to use this method, but mentioned that it was much simpler than computer-based methods.[58]

Discrimination rules

In scoring systems using discrimination rules, data on clinical history, clinical signs and results of special investigations were collected from a sample of patients. An analysis would be conducted to identify factors that differ significantly between patients with and without acute appendicitis. The statistically significant factors were included in a multivariate model, which identified a final set of factors to construct a diagnostic score. This commonly used two-step approach is unnecessary.[83] Various other discrimination rules were used, including stepwise logistic regression[129] and discriminant analysis.[112] Some scoring systems were derived from bivariate analysis alone; that is, factors with statistical differences between patients with and without acute appendicitis were included in the score.[120,121] Items were included in the score if their prevalences were significantly different between patients with acute appendicitis and those without. Each of the items was assigned a subscore of 1 or 0 and the score was calculated by adding up the assigned value from each item.

Neural networks

Another group of decision tools is derived from learning systems, such as neural networks. In one study, the accuracies of four types of neural networks in diagnosing acute appendicitis in AAP patients were compared: Kohanen's self-organising map, binary adaptive resonance theory, learning vector quantisation and the backpropagation algorithm.[131] These four networks are described in Appendix 2 and in more detail elsewhere.[140] Four groups of symptoms were used to test the neural networks, each with a different number of clinical signs and clinical history items. The data were randomly split into a training-set sample (n = 454) and a test-set sample (n = 457).

Additional details

More details on the characteristics of these decision tools can be found in Appendices 2 and 8 and *Table 3*.

Results of the review

Summary of included studies

Table 3 provides a synopsis of some basic characteristics of the eligible decision tool studies. The breakdown of the reasoning methods used was as follows:

- bayesian (11, of which nine were studies of the Leeds AAP system)
- logistic regression (seven tools, of which one reported AUROC, but not sensitivity or specificity)
- Alvarado score (seven of which five used the modified version of the score)
- discrimination rules (five, of which one reported AUROC, but not sensitivity or specificity, two used stepwise logistic regression, two used bivariate analysis and one used stepwise discriminant analysis)
- artificial neural networks (four).

Among the 34 decision tool studies (from 27 articles), two studies[129,136] reported the sensitivities and specificities of doctors aided by decision tools (n = 2582), compared with unaided doctors. Thirty-two studies reported sensitivities and specificities of decision tools (including Wellwood[136]). Wellwood and colleagues' study reported the accuracy of both the decision tool and the aided doctor. Two studies reported AUROC curve only and could not be included in the forest plot or the meta-analysis (n = 1091).[118,132] The 30 decision tool accuracy studies with sensitivity and specificity estimates were used to construct SROC curves and metaregression models. Among these studies, 14 reported only crude accuracies, but contained data that enabled the calculation of sensitivities and specificities. Both studies of aided doctors reported sensitivities and specificities.

The prevalence of acute appendicitis in the studies ranged from 6 to 88%. Six to 36 items were included in a decision tool. Among the 34 decision tool studies (including two that reported AUROC curve only), 12 studies were independent evaluations of other researchers' decision tools, while the rest were evaluations of researchers' own tools. The breakdown of the type of data set for the decision tool studies was as follows:

- prospective test-set data collected in an independent centre, away from where the decision tool was developed (13 studies)
- prospective test-set data collected in same centre where tool was developed (nine studies)
- split samples (seven studies)
- training-set data (five studies).

Of the two studies of aided doctors' diagnoses, Wellwood and colleagues' study used prospective test-set data collected in an independent centre,[136] whereas Ohmann and colleagues used prospective test-set data collected in the same centre where the tool was developed.[129]

The forest plots in *Figure 6* provide a summary of the results for decision tool accuracy studies. The 25 papers containing 30 studies with information on sensitivity and specificity covered 15,040 patients. There is great variation in study size. Wellwood and Sutton were the largest studies (sample sizes of 5193 and 1985, respectively), with the narrowest 95% confidence intervals.[135,136] Kirkeby's study, a test-set validation of the Leeds AAP system, was the smallest study ($n = 77$) with the widest confidence interval.[123]

Among the 32 decision tool accuracy studies, ten studies also assessed unaided doctors' diagnoses ($n = 10{,}496$). Nine reported sensitivities and specificities or data that enabled their calculations. One study reported AUROC curve for doctors' unaided diagnosis only and could not be included in the forest plot or the meta-analysis ($n = 304$).[118] The nine studies in the forest plot for unaided doctors (*Figure 6*) therefore covered 10,496 patients, and were used to construct SROC curves and metaregression models.[123,135] Among accuracy studies of unaided doctors included in the meta-analysis, only two reported sensitivities and specificities together with 2 × 2 tables.[121,131] The rest reported crude accuracies, but contained data that enabled the calculation of these measures of diagnostic accuracy.

The studies in the review of decision tool accuracy were conducted from 1972 to 2000. For the review of aided doctors' accuracy, the two studies were conducted in 1992 and 1999. For the review of doctors' unaided diagnosis, the studies were conducted from 1972 to 1999.

The breakdown of countries for studies of decision tool accuracy was as follows: USA (three studies), UK (eight), Finland (six), Norway (five), Germany (two), Sweden (two) and Canada, Denmark, India, Pakistan, Iran and Hong Kong (one study each). These 32 studies were published in 19 journals and one book.

Accuracy of decision tools, unaided doctors' diagnosis and aided doctors' diagnosis

Reviews of decision tool studies and studies of unaided doctors' performance: overall results

The sensitivities of decision tools ranged from 53% (the Leeds AAP system in 1993[124]) to 99% (the Leeds AAP system in 1972[113]). The specificities ranged from 30% (Malik's study on the Alvarado score[128]) to 99% (the Leeds AAP system in 1976[119]). The sensitivities of unaided doctors' diagnosis ranged from 64 to 93%, while the specificities ranged from 39 to 91%.

The forest plots (*Figure 6*) for decision tools and unaided doctors' diagnosis demonstrate obvious heterogeneity in the estimated sensitivities and specificities. This was confirmed by tests of heterogeneity for both sensitivity ($\chi^2 = 65.6$, df = 8, $p < 0.001$ for unaided doctors, and $\chi^2 = 356$, df = 29, $p < 0.001$ for decision tools) and specificity ($\chi^2 = 368$, df = 8, $p < 0.001$ for unaided doctors, and $\chi^2 = 725$, df = 29, $p < 0.001$ for decision tools). Some AAP decision tools showed marked heterogeneity between studies assessing the same tool (Appendix 6). For example, the estimated sensitivities and specificities of the Leeds AAP system ranged from 52.9 to 99.2% and 52.5 to 98.6%, respectively. These results indicate that the estimation of pooled estimates of sensitivities and specificities is inappropriate.

Figures 7 and *8* show the SROC curves for decision tools and unaided doctors' diagnosis, respectively. The squares in the two plots correspond to the sensitivities and specificities from the individual studies. The area of each square depicts the size of a study. Using Littenberg and Moses' regression models,[91] *S*, the summary measure of diagnostic threshold, was significant neither for AAP decision tools nor for unaided doctors' diagnosis. This suggests that the DOR is constant regardless of changes in the diagnostic threshold and is reflected by a symmetrical SROC curve, as shown in (*Figures 7* and *8*).

Decision tool performance compared with unaided doctors' performance: overall results

Among the 14 studies for which data are available, how do decision tools fare in comparison with

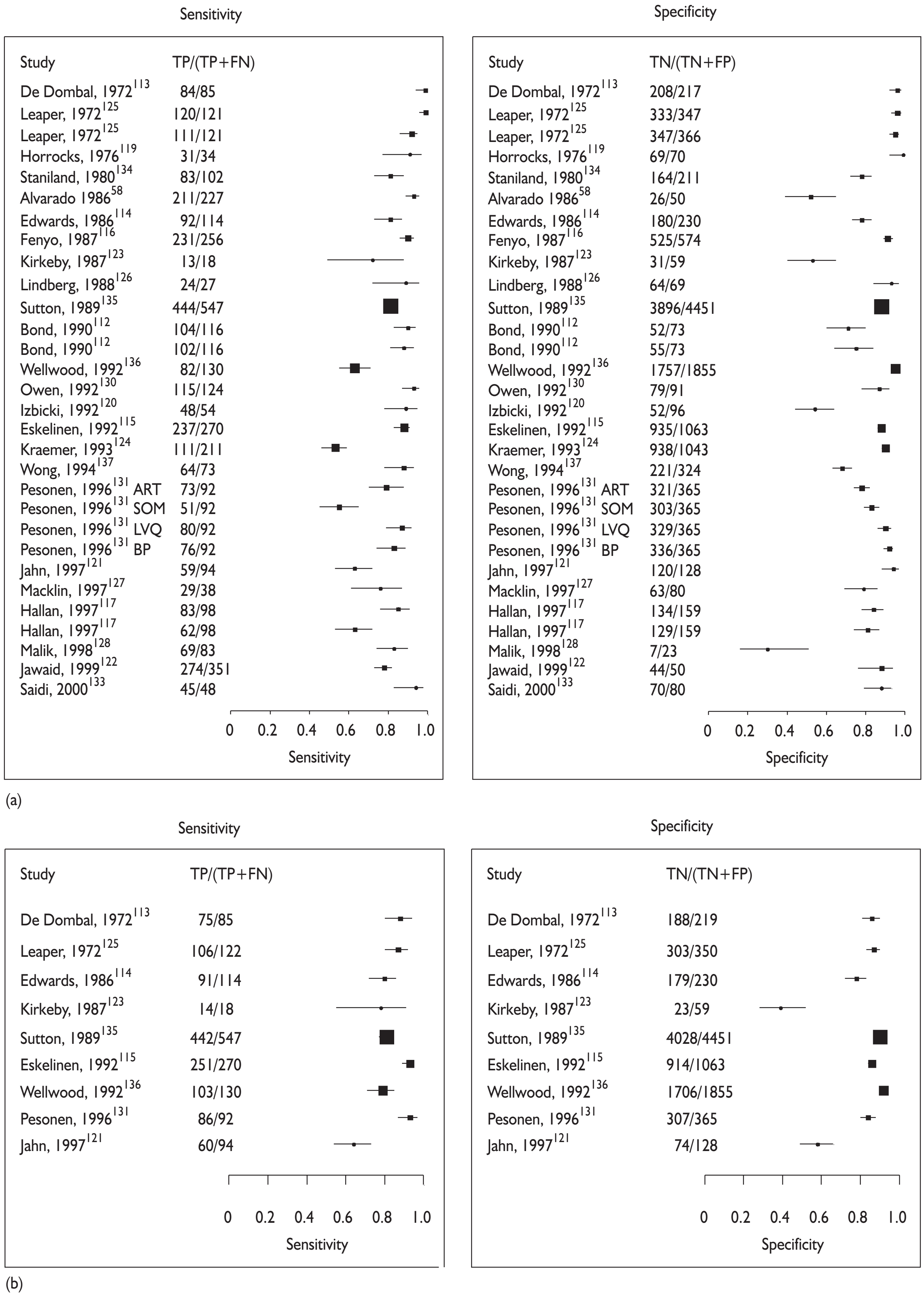

FIGURE 6 *Sensitivity and specificity estimates for (a) decision tools and (b) unaided doctors' diagnosis. FN, false negative; FP, false positive; TN, true negative; TP, true positive.*

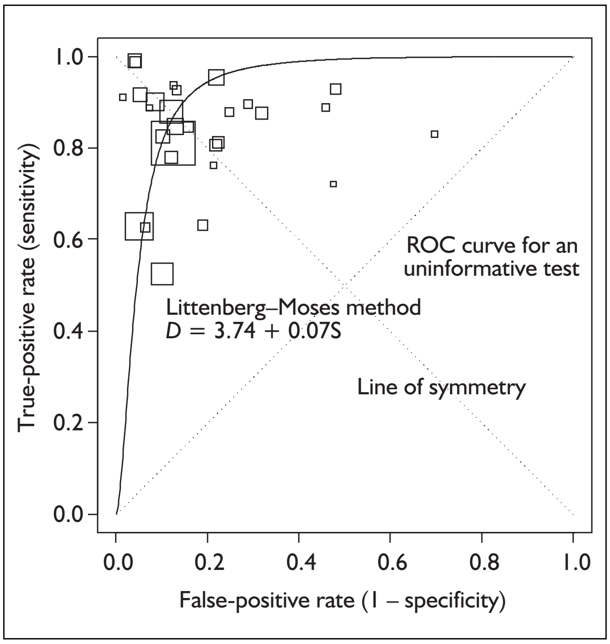

***FIGURE 7** SROC plot for decision tools*

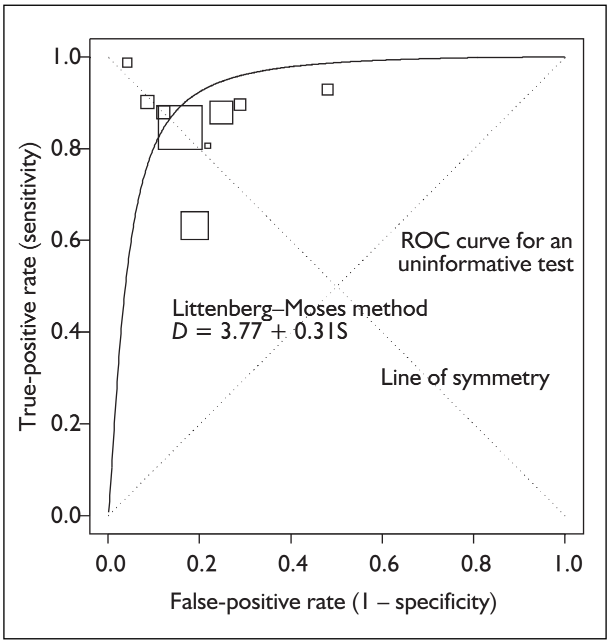

***FIGURE 8** SROC plot for unaided doctors' diagnosis*

doctors' unaided diagnosis? One of these comparisons reported AUROC curves only.[118] For Bayesian decision tools, seven out of 13 pairs of accuracy estimates (with the sensitivity and specificity of each tool forming a pair) were compared with unaided doctors' accuracy on the same cases. For other types of tools, seven out of 20 pairs of estimates were directly compared with unaided doctors.

In general, the performance of the Leeds AAP system as reported by its developers was superior to unaided doctors (*Table 4*). De Dombal and colleagues' landmark study in 1972 demonstrated this superiority.[113] Comparisons with other groups of doctors (e.g. registrars) by the Leeds group during the same period yielded similar results.[113,125] However, evaluations of the system by other investigators did not replicate the high performance achieved in the 1972 studies. In later independent evaluations by Kirkeby in 1987 and Wellwood in 1992, the sensitivities and specificities fell markedly.[123,136]

Looking at the performance of other Bayesian decision tools, the sensitivity of Sutton's system was very similar to that of unaided doctors (*Table 4*).[135] However, unaided doctors' diagnosis appeared to be more specific than Sutton's system.[135] There was little difference in the sensitivity and specificity of Edwards' Bayesian system and unaided doctors.[114]

Among tools that used other reasoning methods (e.g. discrimination rule-based scores and neural networks), six studies compared the performance of their tools with unaided doctors' diagnosis (*Table 4*). The directions are mixed. For example, Pesonen's neural network (Kohanen's self-organising map) was more specific but less sensitive than the unaided doctor.[131] Hallan and colleagues reported the AUROC curve with 95% confidence intervals for their logistic regression-derived decision tool.[118] The AUROC curve of the tool (0.809, 95% CI 0.797 to 0.824) was similar to that of the unaided doctor (0.813, 95% CI 0.797 to 0.829).

From the ROC plot in *Figure 9*, the paired differences in diagnostic accuracy (sensitivity and specificity) between decision tools and unaided doctors in the same comparison can be seen graphically. It appears that both the magnitude and direction of the paired differences are heterogeneous between studies.

This is confirmed in the meta-analyses of the error rate ratios presented as forest plots in *Figure 10*. The results are highly heterogeneous for both forest plots, as shown formally by the tests of homogeneity ($p < 0.001$ for ratios of false-negative rates and for ratios of false-positive rates) and visually by the obvious variations in the study-specific error rate ratios (*Figure 10*). On average, there was insufficient evidence to indicate a

***TABLE 4** Comparison of diagnostic accuracies of unaided doctors with decision support*

Study (decision tool)	Sensitivity			Specificity		
	Decision tools % (*n*) (95% CI)	Unaided doctors % (*n*) (95% CI)	Difference (95% CI of difference)	Decision tools % (*n*)	Unaided doctors % (*n*)	Difference (95% CI of difference)
De Dombal, 1972[113] (Leeds Bayesian)	98.8 (85) (94.9 to 99.8)	87.4 (87) (78.8 to 92.8)	+11.4 (+3.95 to +20.12)	95.9 (217) (92.3 to 97.8)	85.1 (235) (80.0 to 89.1)	+10.8 (+5.41 to +16.22)
Leaper, 1972[125] (Leeds Bayesian)	99.2 (121) (95.5 to 99.9)	86.9 (122) (79.8 to 91.8)	+12.3 (+6.16 to +19.45)	96.0 (347) (93.3 to 97.6)	86.6 (350) (82.6 to 89.7)	+9.4 (+5.27 to +13.68)
Leaper, 1972[125] (Leeds Bayesian)	91.7 (121) (85.5 to 95.4)	86.9 (122) (79.8 to 91.8)	+4.8 (+3.59 to +6.01)	94.8 (347) (92.0 to 96.7)	86.6 (350) (82.6 to 89.7)	+8.2 (+6.99 to +9.39)
Kirkeby, 1987[123] (Leeds Bayesian)	72.2 (18) (49.1 to 87.5)	77.8 (18) (54.8 to 90.1)	–5.6 (–22.05 to 32.17)	52.5 (59) (40.0 to 64.7)	39.0 (59) (27.6 to 51.7)	+13.5 (–4.30 to 30.25)
Wellwood, 1992[136] (Leeds Bayesian)	63.1 (130) (54.5 to 70.9)	79.2 (130) (71.5 to 85.2)	–16.2 (–26.7 to ––5.1)	94.7 (1855) (93.6 to 95.6)	92.0 (1855) (90.6 to 93.1)	+2.7 (+1.15 to +4.37)
Edwards, 1986[114] (Bayesian)	80.7 (114) (72.5 to 86.9)	79.8 (114) (71.5 to 86.2)	+0.90 (–9.48 to +11.22)	78.3 (230) (72.5 to 83.1)	77.8 (230) (72.0 to 82.7)	0.44 (–7.13 to +7.99)
Sutton, 1989[135] (Bayesian)	81.2 (547) (77.7 to 84.2)	80.8 (547) (78.3 to 83.8)	+0.4 (–4.29 to +5.02)	87.5 (4451) (86.5 to 88.5)	90.5 (4451) (89.6 to 91.3)	–3.0 (–4.29 to –1.69)
Eskelinen, 1992[115] (Logit)	88.0 (270) (83.7 to 91.5)	93.0 (270) (87.7 to 94.7)	–5.0 (–9.87 to +0.16)	88.0 (1063) (85.9 to 89.8)	86.0 (1063) (83.8 to 87.9)	+2.0 (–0.89 to +4.84)
Jahn, 1997[121] (Bivariate score)	62.8 (94) (52.7 to 71.9)	63.8 (94) (53.7 to 72.8)	–1.0 (–14.58 to +12.51)	93.7 (128) (88.2 to 96.8)	57.8 (128) (49.2 to 66.02)	+35.9 (+26.00 to +45.12)
Pesonen, 1996[131] Neural net (ART1)	79.3 (92) (70.0 to 86.4)	93.5 (92) (89.1 to 98.3)	–14.2 (–11.88 to –19.11)	78.1 (365) (73.6 to 82.0)	84.1 (365) (79.3 to 86.9)	–6.0 (–11.78 to –0.34)
Pesonen, 1996[131] Neural net (SOM)	55.4 (92) (45.3 to 65.2)	93.5 (92) (89.1 to 98.3)	–8.7 (–18.09 to +0.47)	83.0 (365) (78.8 to 86.5)	84.1 (365) (79.3 to 86.9)	–1.1 (–6.49 to +4.30)
Pesonen, 1996[131] Neural net (LVQ)	87.0 (92) (78.6 to 92.4)	93.5 (92) (89.1 to 98.3)	–6.5 (–5.86 to –10.48)	89.9 (366) (86.4 to 92.6)	84.1 (365) (79.3 to 86.9)	+5.8 (+5.65 to +7.07)
Pesonen, 1996[131] Neural net (BP)	82.6 (92) (73.6 to 89.0)	93.5 (92) (89.1 to 98.3)	–10.9 (–9.39 to –12.53)	92.1 (366) (88.8 to 94.4)	84.1 (365) (79.3 to 86.9)	+8.0 (+7.55 to +9.51)

CI, confidence interval.

difference in false-negative rates between decision tools and unaided doctors' diagnosis. Decision tools have lower false-positive rates than unaided doctors' diagnosis (error rate ratio 0.62, 95% CI 0.46 to 0.83). Unaided doctors' diagnosis has lower false-negative rates than decision tools (error rate ratio 1.34, 95% CI 0.93 to 1.93), although the lower end of the confidence interval falls just short of unity.

Meta-analysis of decision tools with unaided doctors using least biased test-set data

The analysis summarised in *Figure 10* was repeated for studies that were assessed by independent evaluators (i.e. those who had not developed the decision tool) in a different centre from where the decision tool was originally developed (*Figure 11*). These two criteria provide the least biased and independent type of test-set data, compared with alternative methods, such as split samples, jack-knife resampling and the use of training-set data as test data set, or developers evaluating tools that they had developed themselves. Four studies fulfilled the criteria.[114,123,135,136]

In the meta-analysis comparing the false-negative rates, there was insufficient evidence to indicate heterogeneity ($p = 0.073$) (*Figure 11a*). The pooled error rate ratio (or overall risk ratio as shown in *Figure 11*) indicates insufficient evidence of a difference between the sensitivities of the decision tools and unaided doctors' diagnosis. In the meta-analysis comparing the false-positive rates (*Figure 11b*), a high degree of heterogeneity was

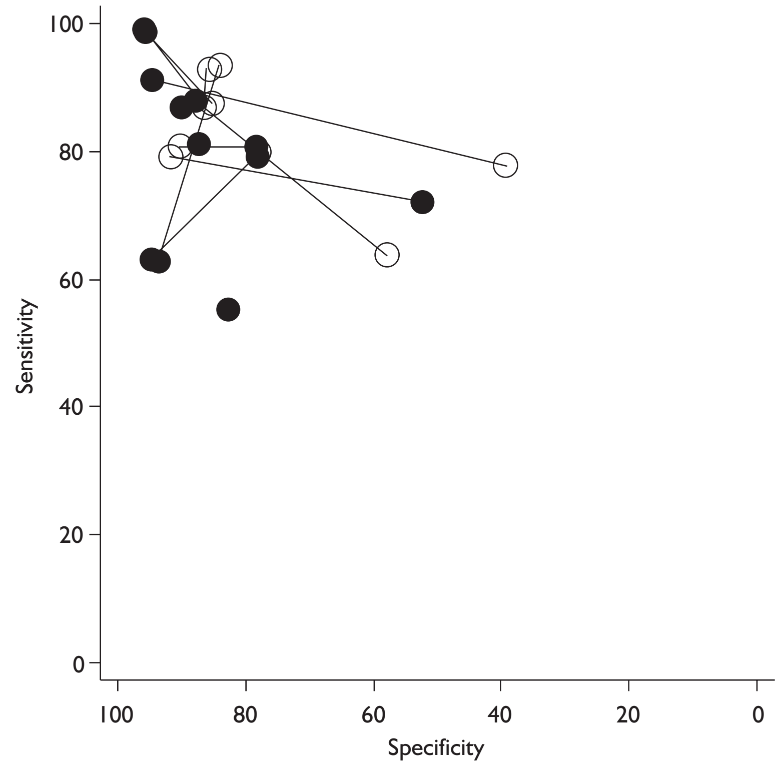

***FIGURE 9** ROC plot comparing paired accuracies of decision tools to unaided doctors' diagnosis. ●, performance of decision tool; ○, performance of unaided doctors' diagnosis. The lines connecting the black to the white spots indicate paired comparisons in the same study.*

detected (p < 0.001). The error rate ratio for Sutton was opposite that of Wellwood.[135,136] Wellwood's finding indicates that the specificity of the Leeds AAP system is higher than that of unaided doctors, while Sutton's finding indicates that specificity of the CADA (Computer Assisted Diagnostic and Audit database) system is lower than that of unaided doctors (*Figure 11b*). The pooled error rate ratio suggests that overall there was no difference between the specificities of the decision tools and unaided doctors' diagnosis.

Metaregression results

The analyses using Littenberg and Moses' regression[91] suggest a lack of evidence of a threshold effect with diagnostic accuracy. There is thus no need to adjust for threshold in the metaregression models and *S* was not included as a covariate.

Metaregression of AAP decision tools

Thirty studies with estimates of sensitivities and specificities of decision tools were included in the metaregression. The results of the metaregression for AAP decision tools are reported in *Table 5*.

There is a lack of evidence to suggest that diagnostic accuracy for AAP decision tools is associated with the prevalence of disease (*Table 5*), the number of items in the decision tool or the tool's reasoning method. In regard to prevalence, there was no evidence of a threshold shift. Sensitivity and specificity did not both vary with prevalence.

For the type of data set, 'testing in the same centre' suggests a significantly higher DOR than 'testing in a different centre', with an RDOR of 8.19 (95% CI 3.09 to 21.73) (*Table 5*). In other words, a decision tool that was tested in the same centre where it was developed was likely to demonstrate better diagnostic accuracy than a tool that was tested in a different centre from the original place of development. There is insufficient evidence to suggest that DORs for 'other types of data set' (jack-knife, random split sample or training-set data) are higher than 'testing in a different centre' (*Table 5*). The box-and-whisker plot in *Figure 12* provides a graphical representation of this relation.

The diagnostic accuracy of AAP decision tools is significantly associated with the evaluator of the tool (*Table 5*). Studies in which investigators evaluated a tool that they had developed

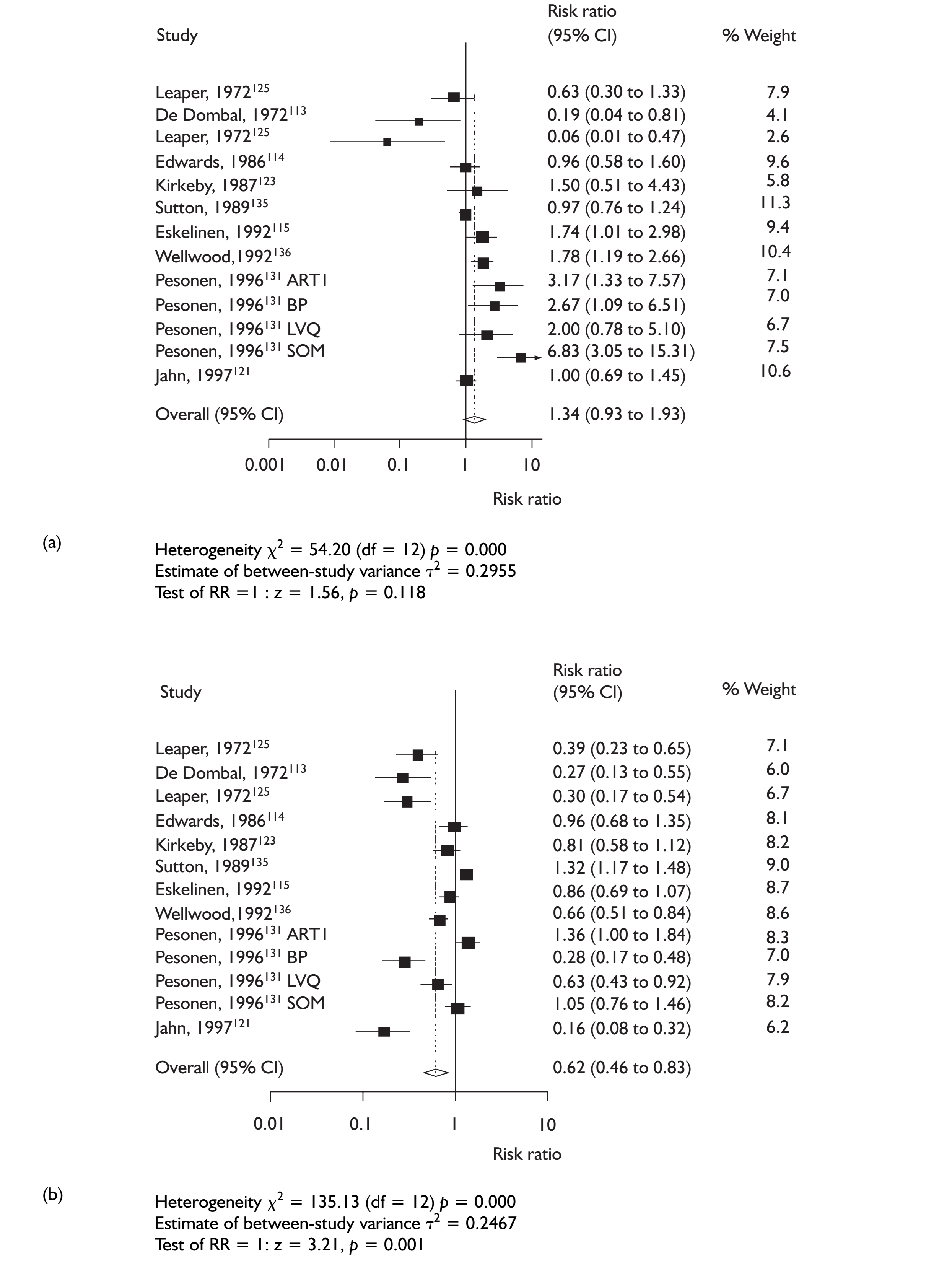

FIGURE 10 *Forest plots for error rate ratios of (a) false-negative rates and (b) false-positive rates for decision tools compared with unaided doctors' diagnosis. df, degrees of freedom; RR, relative risk.*

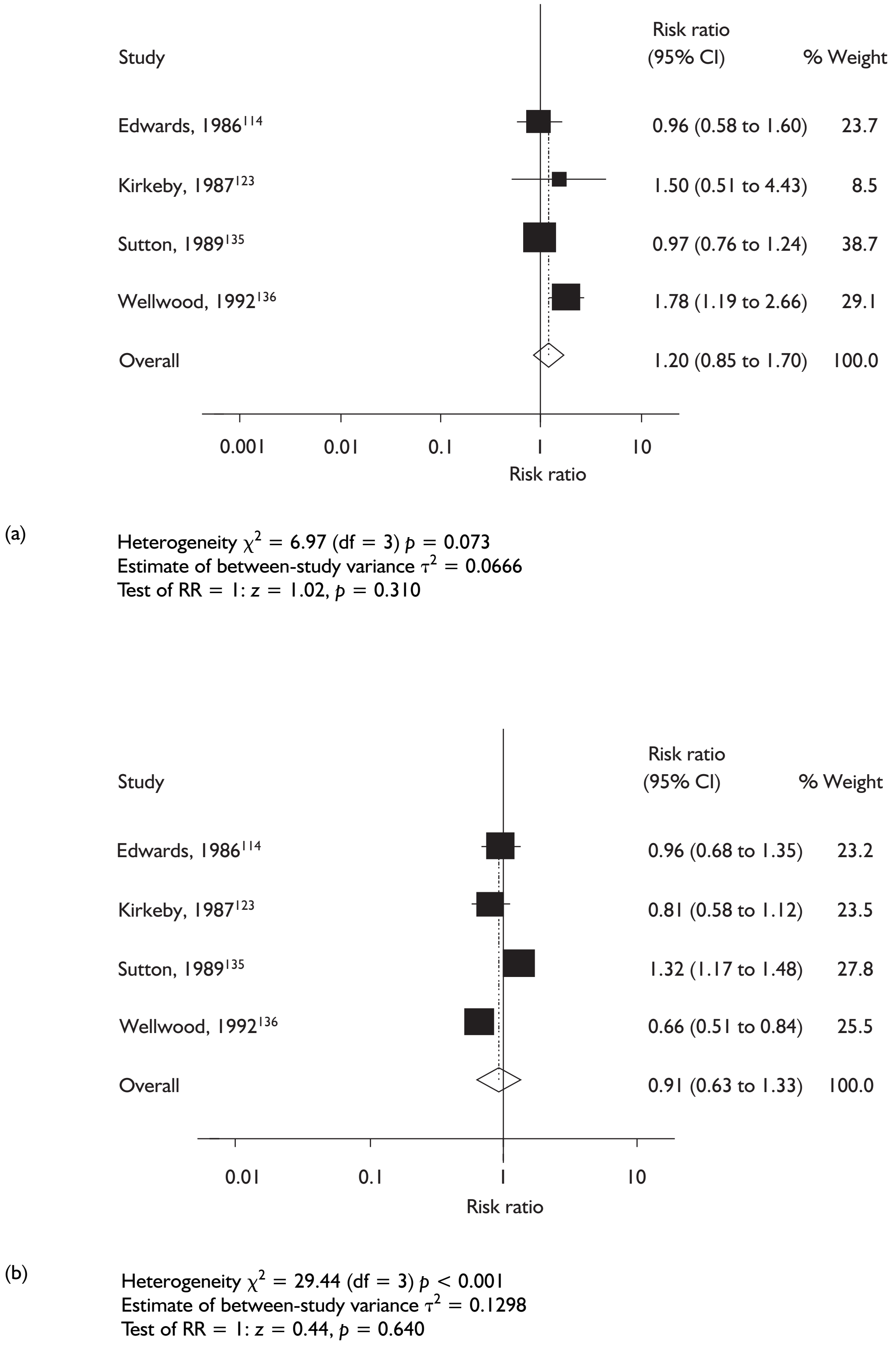

FIGURE 11 *Forest plots for error rate ratios of (a) false-negative rates and (b) false-positive rates for decision tools compared with unaided doctors' diagnosis in high-quality studies*

TABLE 5 *Summary table of metaregression of AAP decision tools*

Variable	RDOR (95% CI)	SE	*p*
S	1.38 (0.95 to 1.99)	0.189	0.12
Prevalence	1.00 (0.78 to 1.28)	0.127	0.96
Year	0.88 (0.83 to 0.92)	0.026	<0.001
Number of items	1.03 (0.98 to 1.08)	0.023	0.22
Reasoning method:			
Logit vs Bayes	1.40 (0.24 to 8.14)	0.899	0.71
Other vs Bayes	0.69 (0.26 to 1.84)	0.498	0.47
Type of test set:			
Same vs different centre	8.19 (3.09 to 21.73)	0.489	<0.001
Other vs different centre	1.19 (0.52 to 2.70)	0.418	0.680
Evaluated own tool	2.97 (1.31 to 6.77)	0.420	0.02

themselves showed a higher DOR than independent evaluations of decision tools (RDOR 2.97, 95% CI 1.31 to 6.77). The box-and-whisker plot in *Figure 12* provides a visual representation of this relation.

The diagnostic accuracy for AAP decision tools is strongly associated with the year of the study (i.e. change of one year). The RDOR of 0.88 (p < 0.001, 95% CI 0.83 to 0.92) (*Table 5* and *Figure 13*) means that the older the study, the higher the DOR, and vice versa. In 1972, the DORs for two test-set evaluations of the Leeds AAP system were 2854 and 1941.[113,125] In the 1980s, DORs for AAP decision tools ranged from 2.88 (Kirkeby's independent assessment of the Leeds system) to 99.0 (Fenyo's Bayesian score).[116,123] In the 1990s, DORs ranged from 2.16 (Malik's Alvarado score) to 76.5 (Pesonen's neural network).[128,131]

Metaregression of unaided doctors' diagnoses

The metaregression analyses indicated no evidence of associations between the DOR for unaided doctors and any of the covariates, including prevalence (*Table 6*). In regard to prevalence, there was a lack of evidence of a threshold shift. Sensitivity and specificity did not both vary with prevalence. The seniority of the doctor making the initial diagnosis of a patient was another possible effect modifier, but given the large number of missing data, it could not be included in the analysis.

Quality of study methods and reporting

All of the included studies recruited their patients consecutively or used a random sample of their patient population. A single relevant clinical population is typically studied, rather than separate positive and negative groups. The study design was typically reported as prospective cohort. These are particular strengths of the papers that have been included in this review. The studies tended to be quite poor in reporting the characteristics of the study population. Many did not give an age or gender breakdown. In one study children were included in a subgroup analysis of decision tool accuracy, but the age range used to define a child was not given.[128] Authors were often vague about blinding. Details of the quality assessment for each study included in the accuracy review can be found in the summary tables in Appendix 9.

Some items in the quality assessment form were omitted from the metaregression, mainly because there was not enough heterogeneity of responses for the analysis to be useful. The omitted items included potential for partial verification bias and the type of study (almost all studies were prospective), completeness of data (patients analysed divided by eligible patients), treatment paradox (no patients were treated for their AAP before decision tools were used) and indeterminate outputs (most studies were unclear on how indeterminate outputs from the included decision tools were handled).

The results of the metaregression of the quality indicators are shown in *Table 7*. None of the quality indicators was significantly associated with diagnostic accuracy, suggesting insufficient evidence that variations in the quality of the studies systematically influence diagnostic accuracy.

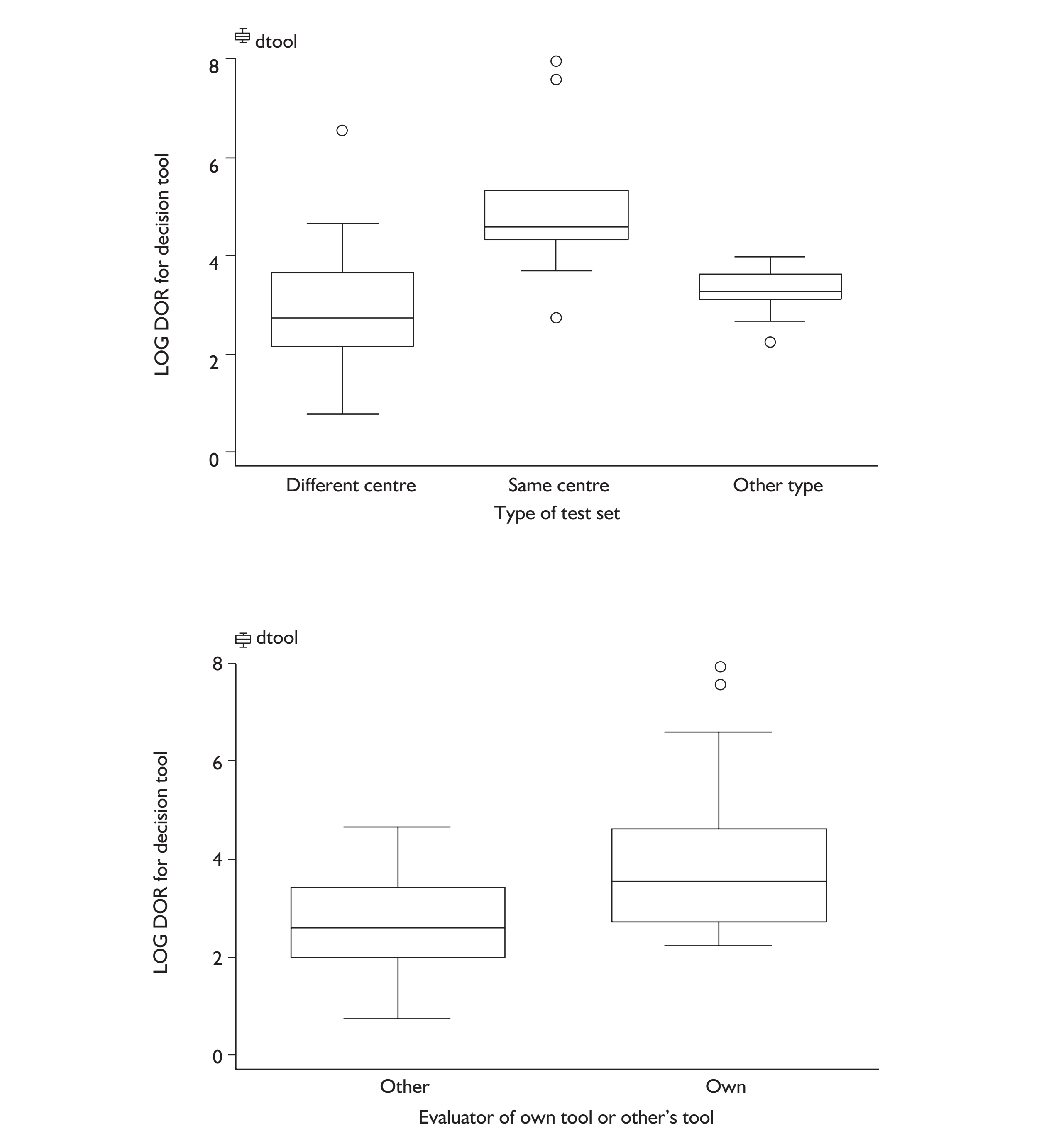

***FIGURE 12** Box-and-whisker plots of ln DOR of decision tool against type of test set and type of evaluator*

Aided doctors' diagnosis compared with unaided diagnosis

Two studies compared doctors aided by an AAP decision tool with unaided doctors.[136,141] In the Wellwood study[136] (described in more detail in Chapter 5, on the results of the impact study review), a multiarm cluster randomised trial randomised doctors to four groups: doctors who were allocated (1) no decision tool, (2) structured data collection forms only, (3) structured data collection forms and printed output from the Leeds AAP system, and (4) structured data collection forms, printed output from Leeds and monthly feedback meetings. As shown in *Table 8* (Chapter 5), there was insufficient evidence to conclude that the specificities and sensitivities were significantly different between the four groups.

The Ohmann study[141] compared the accuracy of doctors' diagnosis of appendicitis between those

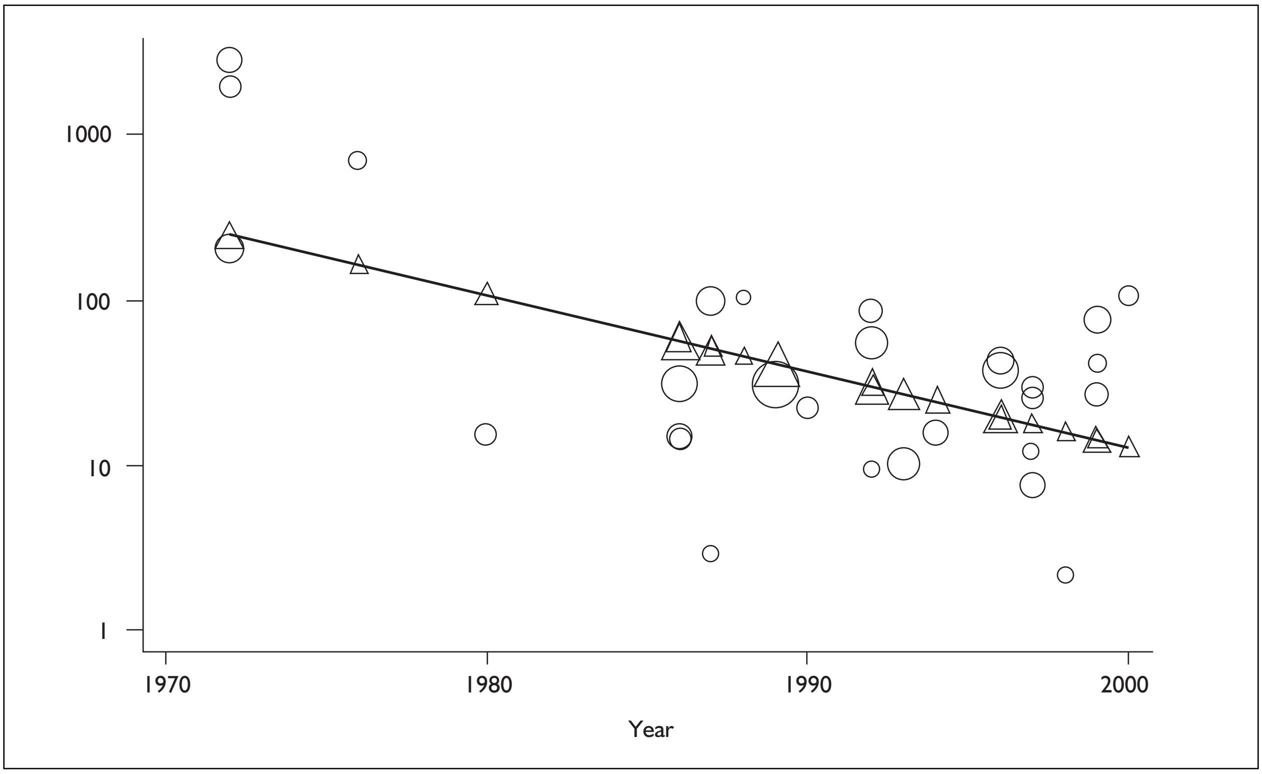

***FIGURE 13** Plot of DOR for decision tool against year of study. The area of each circle in the plot is the square root of the sample size of each study. The triangles represent fitted points on the regression line of best fit.*

***TABLE 6** Summary table of metaregression of unaided doctors' diagnosis*

Variable	RDOR (95% CI)	SE	*p*
S	1.07 (0.60 to 1.89)	0.29	0.82
Prevalence	0.27 (0.93 to 1.02)	0.024	0.27
Year	1.00 (0.93 to 1.07)	0.036	0.98
Type of test set			
Same vs different centre	1.13 (0.27 to 4.62)	0.720	0.87
Other vs different centre	1.44 (0.48 to 4.31)	0.559	0.53
Evaluated own tool	0.97 (0.39 to 2.40)	0.462	0.95

who received a diagnostic score developed using logistic regression (n = 829) and those who did not (n = 597). There was insufficient evidence to suggest a difference between the sensitivity of aided doctors' diagnosis (95.5%) and unaided doctors' diagnosis (91.5%) (p = 0.24). The specificity of aided doctors (78.1%) was significantly less than the specificity of unaided doctors (86.4%) (p < 0.001).[129]

The Ohmann study[141] is an uncontrolled before-and-after study, whereas the Wellwood study[136] is a cluster RCT. In the hierarchy of evidence,[142] the latter would be considered a higher quality study than the former, although both are informative.

Sensitivity analysis

Some of the studies included in the meta-analysis and metaregression (e.g. Pesonen[131]) used the same dataset to evaluate different decision tools. A sensitivity analysis was conducted in which the analysis was repeated by including one study from each article (even if an article contained more than one study). The direction of all results remained the same.

Publication bias

A visual examination of the funnel plot (*Figure 14*) suggests that it is symmetrical and that publication bias is not likely to be a serious problem for AAP decision tools. Given the many sources of

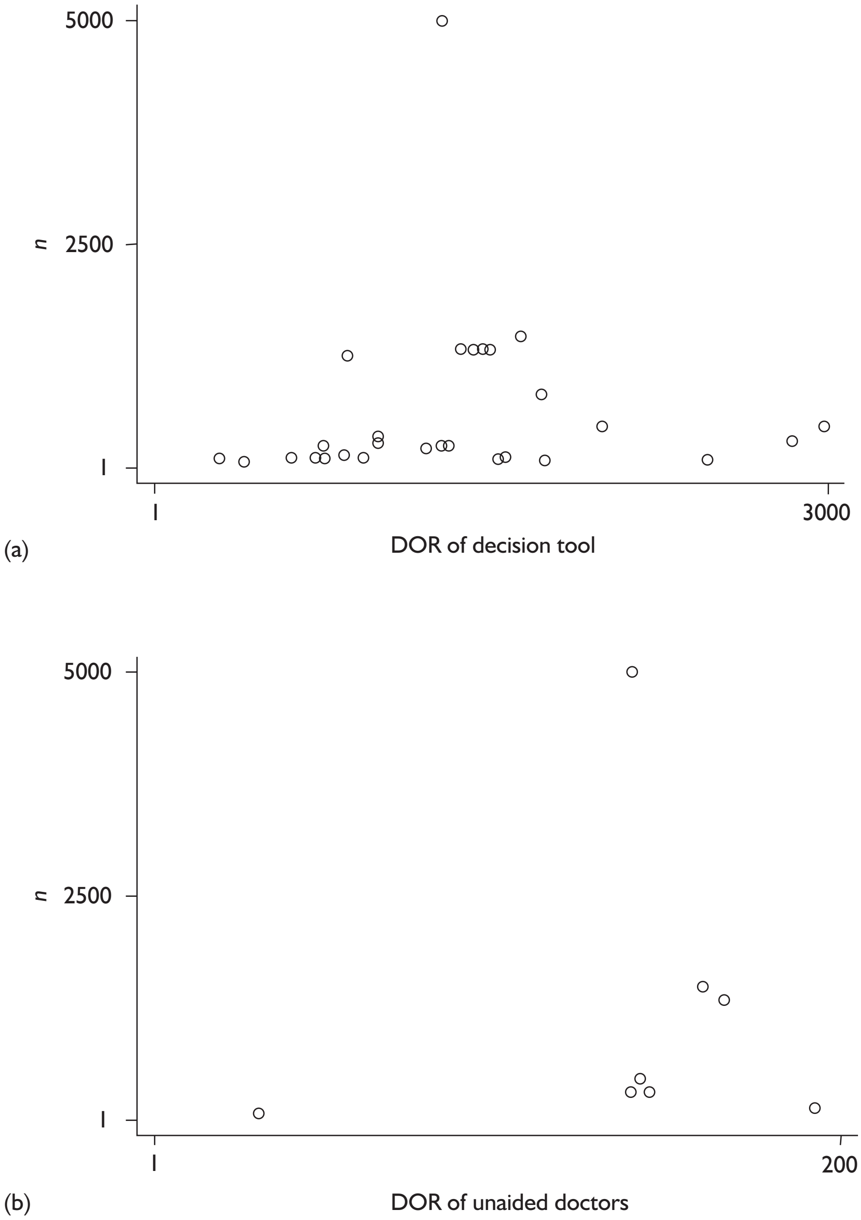

FIGURE 14 *Funnel plots of log DORs against sample size, for (a) decision tools and (b) unaided doctors' diagnosis*

heterogeneity, one must be cautious in visually interpreting a funnel plot, even when it looks symmetrical. There are probably not enough observations to judge whether the funnel plot for unaided doctors' diagnosis is symmetrical or not (*Figure 14*). Whether publication bias is a problem for studies of unaided doctors' diagnosis is uncertain.

Summary and discussion

Thirty decision tool studies (n = 15,040 patients) with sensitivity and specificity estimates of decision tools provided data for constructing SROC curves and metaregression models. Two studies compared the accuracies of doctors aided by decision tools with unaided doctors. Two studies reported AUROC only. The prevalence of acute appendicitis ranged from 6 to 88%.

Thirteen out of the 30 eligible studies of decision tool accuracy reported false-positive and false-negative rates for both decision tools and unaided doctors' diagnosis, enabling a direct comparison of their relative performance. In random effects meta-analysis, decision tools had significantly lower false-positive rates than unaided doctors'

TABLE 7 Summary table of metaregression of quality assessment indicators

Variable	RDOR (95% CI)	SE	*p*
S	1.38 (0.95 to 1.99)	0.189	0.17
Reported patient characteristics	0.54 (0.23 to 1.26)	0.436	0.10
Non-operated cases followed up	1.49 (0.35 to 6.29)	0.736	0.59
Quality of reference standard			
Fair vs good	0.46 (0.11 to 1.97)	0.741	0.31
Poor vs good	1.04 (0.24 to 4.53)	0.751	0.96
Potential for incorporation bias			
Moderate vs high	0.44 (0.18 to 1.07)	0.456	0.11
None vs high	0.87 (0.20 to 3.79)	0.748	0.86
Reference standard allocated blind to DT results	0.48 (0.20 to 1.12)	0.434	0.10
DT user blind to reference standard	0.63 (0.18 to 2.15)	0.629	0.47
Potential for differential verification bias	1.49 (0.35 to 6.29)	0.736	0.59
Patient recruitment			
Consecutive vs random/representative	1.92 (0.78 to 4.71)	0.458	0.17

diagnosis (error rate ratio 0.62, 95% CI 0.46 to 0.83). Unaided doctors' diagnosis had lower false-negative rates than decision tools (error rate ratio 1.34, 95% CI 0.93 to 1.93). These results suggest that decision tools may potentially be useful in confirming a diagnosis of acute appendicitis, but not useful in ruling it out.

Although subgroup analysis was prespecified, the results need to be treated cautiously, particularly given the low percentage of eligible studies that reported the performances of both decision tools and unaided doctors' diagnosis. This could potentially be a manifestation of outcome reporting bias (e.g. studies that analysed the performance of doctors did not report the results because it was better than the decision tool), a phenomenon that is empirically known to be serious for RCTs.[143]

Two studies compared doctors aided by an AAP decision tool with unaided doctors. In both studies, there was no indication to suggest that doctors aided by decision tools were more accurate in diagnosing AAP than doctors not aided by these tools.

In terms of the reported quality of the studies included in this review, the use of a single clinical population in a prospective or retrospective cohort design in the included studies is more valid than the use of case–control designs. The selection of study samples through consecutive recruitment or random sampling is another strength of the included studies, as the chances of selection bias are decreased.

In other ways, the reported quality of many studies was not of a high standard. For example, many studies did not report an age and gender breakdown of the patients. Yet, as discussed in Chapter 1, the epidemiology of AAP is related to age and gender. Most studies were unclear about how indeterminate outputs from AAP decision tools were handled. Authors were often vague about blinding, as well as other quality-related indicators. Sensitivity and specificity (or other suitable measures of performance such as likelihood ratios) were often not reported, and sometimes a 2 × 2 table was not even presented. The reviewers frequently had to scrutinise confusing text and unnecessarily complex tables to extract the data needed to calculate the appropriate measures of diagnostic accuracy.

None of the quality indicators was significantly associated with diagnostic accuracy, suggesting a lack of evidence that systematic variations in the quality of the studies influence diagnostic accuracy. However, if better quality studies had been available, a 'quality effect' on the diagnostic accuracy of the decision tools might have been detected. The results should therefore be treated with caution.

The practice of many authors in presenting crude accuracies as their main results is unhelpful. A highly sensitive tool would enable healthcare workers to rule out a disease, while a highly specific tool would enable them to confirm a diagnosis.[41] Knowledge of crude accuracies does not provide a basis for making such decisions about the clinical value of a decision tool.

The metaregression of AAP decision tools demonstrated statistically significant associations of diagnostic accuracy with the type of test-set data (RDOR of 8.19 for prospective testing in the centre where the tool was developed, compared with prospective testing in a different centre). The identity of the evaluator was also significantly associated with the DOR (RDOR of 2.97). Those who evaluated their own tools were more likely to report higher DORs than those who evaluated tools developed by others. It was previously suspected that validation of a decision tool by its developers at the centre where the tool was developed might provide exaggerated estimates of accuracy compared with evaluations conducted elsewhere by independent researchers.[56] One reason may be because independent evaluations are more objective than those conducted by developers in their own centres. Other reasons are also possible, such as different case-mix and over-optimism induced by data-dependent model selection (e.g. stepwise regression).[102] To the authors' knowledge, this analysis has provided the first empirical demonstration of such an effect.

A statistically significant association was also found between the DOR and the year of the study. The older the study the more accurate it was (RDOR of 0.88). This may be an indication that given time and scrutiny, the initially high accuracy of a decision tool would often be shown to be overly optimistic. Prevalence might be expected to have an effect on accuracy of diagnostic technologies such as decision tools,[41] but there was insufficient evidence to indicate this in the metaregression.

Metaregression has limitations. One problem is the risk of detecting spurious relations from overfitting because of the often small number of studies. This means that the effects of only one effect modifier at a time should be included as an independent variable. Since covariates and outcomes in a metaregression are defined at the group rather than at an individual level, ecological bias is possible and causality cannot be inferred from the observed associations.[144]

One might justifiably argue that appendicitis may spontaneously recover, and histological proof of appendicitis in operated patients may confirm the diagnosis, but not whether surgery would have been necessary. However, the danger with delaying surgery when appendicitis is suspected is that of perforation, peritonitis and other complications (the mortality risk ratio of perforation is 5.0[17]), which makes it risky and perhaps unethical to wait for the appendicitis to recover spontaneously. Waiting to see how the patient develops over a few hours may be acceptable in primary care, where AAP patients have a low probability of appendicitis and can initially be treated with antibiotics, but is arguably unacceptable after they have been referred to secondary care and surgical wards, where the prior probability of appendicitis is over 20%. There was, unfortunately, not a single primary care study that could be included in this review.

Statistical tests to assess funnel plot asymmetry were not used in this review,[111,145] because even for RCTs, there are various reasons why a funnel plot may be asymmetrical.[10] Publication bias, biased inclusion criteria, true heterogeneity and chance are some of the possible explanations for statistically tested asymmetry. The detection of publication bias in reviews of diagnostic technologies is more difficult than for RCTs, since the use of a conventional funnel plot is inappropriate, in contrast to other views.[78,145] Given a lack of consensus and ongoing methodological development in this area, the decision was made to include funnel plots without using statistical tests. Readers are advised to interpret the plots with caution and be aware that publication bias is a possibility.

Chapter 4

Study question 2. Methods for systematic review of impact studies

Introduction

As mentioned in Chapter 1, assessing the diagnostic accuracy of a decision tool provides an important first stage in the clinical management of patients.[64] However, an arguably more important objective is the impact that the decision tool has on patients, and their prognosis, through a doctor's choice of treatment based on the information provided by the tool. There is a need for decision tool evaluations to focus not only on accuracy, but also "on the likelihood that tests [and decision tools] detect clinical events of interest and the effect that tests can have on those events by the way in which the results affect subsequent management decisions."[146] In other words, the bottom line for a decision tool is whether its availability to doctors will have a discernible impact on health-related outcomes.[81,147] Lijmer and Bossuyt presented a rich typology of RCT designs that can answer a wide range of questions on the impact of a diagnostic technology on patient health outcomes.[146] In a systematic review, Hunt and colleagues identified 63 RCTs of computer decision support systems published between 1974 and March 1998, and observed that the quality of these trials is improving over time.[148] Since 1998, a range of good quality RCTs of decision support systems has been published in the top five general medical journals, including a large, rigorous RCT of a system that showed no evidence of a beneficial impact on clinical practice or health outcome.[149]

The systematic review in this chapter focuses on the impact of providing doctors with AAP decision tools on patient outcomes, appropriateness of clinical decisions and actions.

Search methods

See the section 'Search methods' (p. 9) for sources of studies and search strategies.

Assessment of eligibility

Inclusion and exclusion criteria

- Eligible studies: randomised trials and quasi-randomised trials for assessing the impacts of AAP decision tools and unaided doctors on patient outcomes, appropriateness of clinical decisions, and actions. Other studies using lower quality designs were excluded.
- Eligible patients: patients with a main complaint of previously undiagnosed acute upper or lower abdominal pain lasting for not more than 7 days from onset.[1]
- Eligible interventions: decision tools used to manage AAP compared with unaided doctors' diagnosis. Studies of individual laboratory and radiographic investigations, audit and feedback, continuing education activities and telemedicine were excluded.
- Eligible study measures: measures of the impact of the intervention on patient outcomes, appropriateness of clinical decisions and actions (e.g. perforation rates, negative appendicectomy rates, rates of admission to surgical ward, mortality rates or accuracy of doctors' diagnosis), expressed as relative risk reduction, absolute risk reduction, odds ratio, or data that allow these impact measures to be calculated.

Procedures for assessing eligibility

The primary search aimed to identify all published studies that passed the eligibility criteria outlined above, with an initial classification of all search results through scrutiny of titles and abstracts by one reviewer (JLYL) into:

- studies that were obviously irrelevant.
- studies that were potentially relevant.

The hard copies of original articles for all studies in the latter category were obtained.

To assess the reliability of the above processes, a second reviewer (JCW) independently categorised a sample of the full search results. Decisions on whether to include each study in the detailed data extraction were made on an eligibility criteria form. One reviewer (JLYL) assessed the eligibility of all the retrieved studies for detailed data extraction. To assess the reliability of this process, a second reviewer (JCW) independently repeated the eligibility check for a sample of the retrieved studies. Disagreements were resolved through discussion between the two reviewers (JLYL and JCW) and, if needed, a third reviewer (JD) was available to help to resolve the discrepancies in consultation with the other two reviewers. A copy of the eligibility form, which was also used for study question 1, can be found in Appendix 4.

Data extraction

A data collection form was developed incorporating eligibility and coding guidelines by adapting ones used in another HTA-funded systematic review[82] and items in the Consolidated Standards of Reporting Trials (CONSORT) statement.[150] Data from each eligible study were extracted on to the form. The following data were recorded from each study:

- details of patients, including demographic characteristics (age, gender, ethnicity) and medical condition(s) causing AAP
- the reasoning method used by each decision tool
- healthcare setting (ward admission, surgical department, A&E, other secondary care, primary care)
- method of presenting results from the decision tool (probability as a percentage, probability from 0 to 1, graph, prose report, raw score or other)
- measures of the impacts of decision tools and unaided doctors' decision-making on patient outcomes, appropriateness of clinical decisions and actions (e.g. reduction in negative appendicectomy rates, perforation rates, admission rates, mortality rates, and improvements in the accuracy of doctors' diagnostic accuracy
- indicators of methodological quality as outlined below.

Methodological quality

Two reviewers (JLYL and JCW) extracted data and assessed the quality of all studies selected for inclusion in the review. Where disagreements continued after discussion between the two reviewers, a third reviewer (JD) was available to help to resolve the discrepancies. The criteria below were used to assess study quality:

- Allocation process: alternation or randomisation (by patient, practitioner, team or other unit); method of randomisation (published random numbers, computer-generated random numbers, toss of coin, etc.); method of quasi-randomisation (birth dates, patient identification numbers, etc.).
- Allocation concealment: allocation concealment is only necessary if study allocation is by patient.[151,152] If allocation is by patient, then it is concealed if investigators were not aware of the allocation of each patient before trial entry, using schemes such as centralised telephone randomisation or opaque envelopes.[153]
- Contamination: where patients are randomised this is possible, so the allocation unit was recorded. For example, contamination could occur when doctors using a decision tool for patients in the experimental group also took care of patients in the control group.[147] The doctors might remember and use the advice generated by the decision tool in control group patients, perhaps unintentionally. When the controls are 'exposed' to the decision tool, the estimated impact of the decision tool may be understated. One way of avoiding this problem is to randomise the allocation of the tool to doctors (this is called the cluster randomisation method when the doctor, ward or hospital is the unit of randomisation rather than the individual patient), but contamination might still occur if doctors in the experimental group communicated the knowledge that they received from the decision tool to doctors in the control group. Allocation by ward or hospital can help to reduce the likelihood of serious contamination.[154]
- Unit of analysis error: in cluster-randomised trials, the analysis should take clustering effects into account,[155] for example by inflating the standard error or confidence interval by a 'design effect'.[156,157] If data from such trials were analysed on the assumption that individuals were randomised, a unit of analysis error would result and the precision of estimated impact would tend to be overstated.[158–160]
- Analysis by 'intention to provide or communicate information': if some doctors did not receive decision support even though they were assigned to use a decision tool, they should still be included in the data analysis and remain in the group to which they were assigned.[147] In clinical epidemiology, this is known as analysis by 'intention to treat'.[153] In the case of decision tools, a more accurate expression would be 'intention to provide or communicate information'.[81] To check whether this criterion is fulfilled, the percentage of patients excluded from the data analysis or with

missing outcome data were reported for each trial arm. It was noted whether the analysis presented included all data for which final outcomes were known. The completeness of data presentation for known final outcomes was noted.

- Blinding: the findings of a study would have more credibility if patients and study personnel were blind to the decision tool.[147] Although there may be logistical difficulties in achieving this, with some improvisation it can be done. For example, to blind patients from the decision tool, the doctor could use it away from the patient and only bring the printed output into the consultation room.[103] Blinding of the intervention and blinding of outcome assessment were therefore recorded.
- Checklist effect: where a paper checklist was used in addition to another decision tool (e.g. a computer score) this was noted, as a paper checklist alone can improve diagnostic decisions by 10%,[161] and may exert an independent impact on patient outcomes, appropriateness of clinical decisions and actions.
- Feedback effect: the feedback effect occurs when the decision tool improves the performance of users by providing them with feedback and audit on their decisions.[162]
- Other co-interventions: checklists, and audit and feedback are examples of co-interventions that can confound the estimated impact of the decision tool.[162] Other co-interventions include an algorithm or a special investigation in addition to the decision tool.
- Prior sample size calculation: did the authors report how they arrived at their sample size?

Data synthesis

Methods for pooling results and investigating between-study heterogeneity were abandoned when only one study was found to be eligible.

Chapter 5

Study question 2. Results of systematic review of impact studies

Studies included in the review

Only one study out of the 15 that were retrieved for detailed review was eligible: a four-arm cluster RCT conducted by Wellwood and colleagues to assess the impact of the Leeds AAP system on patient health outcomes.[136] The characteristics and development of the Leeds system were discussed in Chapter 3.

Results of the systematic review

Summary of included study

Wellwood and colleagues[136] randomised 40 doctors to four groups and analysed the results using nested analysis of variance (ANOVA). The four groups consisted of (1) doctors who did not use any decision tool; (2) doctors who used structured data collection forms to fill in clinical data; (3) doctors who used structured forms to fill in clinical data and also received output from the Leeds AAP system; and (4) doctors who, in addition to using forms and the Leeds output, attended monthly audit and feedback meetings on AAP. Doctors in the four groups examined a total of 5193 patients suspected of having acute appendicitis. The objective of the trial was to estimate the impact of advice from the Leeds AAP system after taking into account the effects of doctors using structured data collection and receiving regular feedback and audit. Decision support systems often derive part of their impact from the checklist and feedback effects discussed in Chapter 3. This four-arm trial design allows the estimation of unconfounded impacts of the advice from the Leeds AAP system.

The results of Wellwood and colleagues' RCT are summarised in *Table 8*. The overall incidence of acute appendicitis was 6% (5.6% to 6.9%), with 42% males and 58% females. The mean age was 30.3 (SD 20.96), with a range from 0 to 97 years old; 13.1% of patients were aged 60 years or older, while 37.2% were aged 18 years or younger.

Impact of the decision tool

Table 8 shows that admission rates for patients from A&E were higher ($p < 0.001$) for doctors who did not use any decision tool (43%, 95% CI 40.5 to 45.2%) than for doctors who used structured data collection forms (38%, 95% CI 35.8 to 40.6%), doctors who used data collection forms and output from the Leeds AAP system (39%, 95% CI 35.6 to 41.6%) and doctors who used data collection forms, Leeds output, and monthly audit and feedback meetings (34%, 95% CI 31.4 to 37.2%). It appears that the use of structured forms by doctors and feedback meetings accounted for much of the decrease in admission rate for patients. The risk ratios for admissions demonstrate this.

Perforation rates did not significantly differ between the comparison groups: doctors who did not use any decision tool (1.3%, 95% CI 0.9 to 2.0%), doctors who used structured data collection forms (0.8%, 95% CI 0.4 to 1.3%), doctors who used data collection forms and output from the Leeds AAP system (0.6%, 95% CI 0.3 to 1.3%) and doctors who used data collection forms, Leeds output, and monthly audit and feedback meetings (0.7%, 95% CI 0.3 to 1.4%). The risk ratios for the three groups that used decision tools were about half that of the controls, indicating a lower risk of perforation for those who used a combination of forms, computer support and/or feedback. However, the differences were not statistically significant ($p = 0.19$). Negative laparotomy rates did not appear to differ between the comparison groups ($p = 0.48$): doctors who did not use any decision tool (1.9%, 95% CI 1.3 to 2.6%), doctors who used structured data collection forms (1.5% 95% CI 1.0 to 2.2%), doctors who used data collection forms and output from the Leeds AAP system (1.2%, 95% CI 0.7 to 2.1%) and doctors who used data collection forms, Leeds output, and monthly audit and feedback meetings (1.2%, 95% CI 0.7 to 2.1%). The point estimates for the groups that used decision tools were slightly lower than the controls. However, the 95% confidence intervals of the risk ratios for negative laparatomies overlap and all include unity, indicating insufficient evidence of differences between the groups.

In terms of the impact of the decision tools on doctors' diagnostic accuracy, the sensitivities of the

TABLE 8 *Results of the review of impact study*

Study, country Study design	Patient characteristics	Decision tool	Diagnosis			Admissions		Perforations		Negative laparotomy	
			Sensitivity (95% CI)	Specificity (95% CI)	DOR (95% CI)	Admission rate (95% CI)	Risk ratio for DT and outcomes (95% CI)	Perforation rate (95% CI)	Risk ratio for DT and outcomes (95% CI)	Negative laparotomy rate (95% CI)	Risk ratio for DT and outcomes (95% CI)
Wellwood 1992,[136] UK Randomised 40 doctors to four groups Cluster randomised trial, with doctors as units of allocation instead of patients Simultaneous/overlap design Used nested ANOVA to take into account unit of analysis error	Overall incidence of appendicitis very low: 324 (6%) of population 2165 males (41.7%), 3027 females (58.3%) Mean age 30.32 (SD 20.962) years Age range 0–97 years Age band (years) *n* (%) <13 941 (18.1) 13–18 995 (19.2) 19–59 2577(13.1) ≥60 680 (13.1) Total 5193 (100) 5193 patients in four groups	No DT allocated (12 doctors, 1610 patients)	28.4% 29/102 (20.6 to 37.8)	96.0% 1448/1508 (94.9 to 96.9)	9.5 (5.81 to 15.84)	42.8% 689/1610 (40.5 to 45.2)	1.00	1.3% 21/1610 (0.9 to 2.0)	1.00	1.9% 30/1610 (1.3 to 2.6)	1.00
		Forms only (8 doctors, 1598 patients)	42.4% 39/92 (32.8 to 52.6)	95.6% 1439/1506 (94.4 to 96.5)	15.8 (9.77 to 25.56)	38.2% 610/1598 (35.8 to 40.6)	0.89 (0.82 to 0.97)	0.8% 12/1598 (0.4 to 1.3)	0.58 (0.28 to 1.17)	1.5% 24/1598 (1.0 to 2.2)	0.81 (0.47 to 1.37)
		Forms and Leeds DSS output (12 doctors, 986 patients)	47.9% 35/73 (36.9 to 59.2)	97.6% 891/913 (95.1 to 97.5)	25.4 (14.21 to 45.25)	38.5% 380/986 (35.6 to 41.6)	0.90 (0.82 to 1.15)	0.6% 6/986 (0.3 to 1.3)	0.47 (0.19 to 1.15)	1.2% 12/986 (0.7 to 2.1)	0.65 (0.34 to 1.27)
		Forms, Leeds DSS output and monthly feedback meetings (8 doctors, 999 patients)	45.6% 26/57 (33.4 to 58.4)	95.5% 900/942 (94.0 to 96.7)	18.0 (9.80 to 32.95)	34.2% 342/999 (31.4 to 37.2)	0.80 (0.72 to 0.89)	0.7% 7/999 (0.3 to 1.4)	0.54 (0.34 to 1.27)	1.2% 12/999 (0.7 to 2.1)	
			$p = 0.07$ (correcting for different doctors)	$p = 0.09$ (correcting for different doctors)		$p < 0.001$ (correcting for different doctors)		$p = 0.19$ (correcting for different doctors)		$p = 0.46$ (correcting for different doctors)	0.65 (0.33 to 1.25)

The study shows that the major benefit is forms requiring a diagnosis to be recorded. Among senior house officers (SHOs) who were in the 'forms only' group, SHOs used forms in 65% of eligible cases (95% CI 63.5 to 66.6%). Among SHOs who used forms and the Leeds AAP system, SHOs used computers in 50% of eligible cases (95% CI 47.8 to 52.2%) and output was available before their diagnosis in 39% of cases (95% CI 37.0 to 41.3%).
DSS, decision support system.

groups that used decision tools (ranging from 42 to 48% for the three groups) were significantly higher than the 28% sensitivity of SHOs' initial diagnosis of acute appendicitis ($p = 0.035$). Although statistically significant, the difference is arguably not clinically significant, since the sensitivities of all four groups were too low to be used to rule out acute appendicitis as a cause of a patient's AAP (Sackett and colleagues' 'SnNout' rule: if the sensitivity of a diagnostic test for a target disorder is very high, say more than 95%, then a negative test result rules out the target disorder as the cause of a patient's symptoms[41]). The specificities of the control group and the three groups of decision tools were all high (all over 95%) and not significantly different from each other. The specificity of the unaided SHOs' diagnosis was similar to those who used decision tools. Thus, with a specificity of 96%, SHOs can confirm a diagnosis of acute appendicitis as confidently as SHOs in the three 'decision tool' groups (Sackett and colleagues' 'SpPin' rule: if the specificity of a diagnostic test for a target disorder is very high, say more than 95%, then a positive test result 'rules in' or confirms that a patient has the disorder[41]). The DORs for all four groups were low (see *Table 2* for examples of typical sensitivities and specificities associated with DORs of different magnitudes), largely because the sensitivities were so low for all of the compared groups.

Quality of study methods and reporting

A particular strength of the study is that it is an RCT. It was unclear to what extent contamination might have weakened the estimated impact of the Leeds AAP system on clinical practice or patient health outcomes. The use of doctors instead of the individual patient as the unit of allocation should reduce the problem. Still, communication between doctors allocated to different groups about their respective experiences in managing AAP patients might have led to some contamination.[163] The level of contamination between the clusters (doctors) would need to be substantial to alter the findings.[163]

Although not explicitly stated, it is clear from the results that the analysis for admissions rate and surgical operations rate was appropriately carried out according to the 'intention to provide or communicate information' principle. Regardless of the doctors' compliance with their allocated regimen, patients remained in the group to which they were allocated in the data analysis.

However, the negative laparotomy rates were analysed using the number of negative laparotomies as the numerator and the number of laparotomies as the denominator. Although this provides a strictly correct definition of the negative laparotomy rate (see Glossary), the randomisation was compromised, as only a subsample from each of four groups was included in the comparison, thus violating the principle of 'intention to provide or communicate information' (the rationale of which was discussed in the section 'Methodological quality', p. 36). The results shown in *Table 8* corrected for this by using all patients recruited to each group as the denominator. Perforation rates, an important outcome, were not reported in the published paper. The perforation rates for the four groups were therefore calculated based on the available data, with the appropriate denominator to preserve the randomisation.

In the paper, the diagnostic accuracy of the doctors was only reported as crude accuracy rates. This is an inappropriate measure of performance. The data were therefore recalculated, with sensitivities, specificities and DORs as reported earlier.

It was unclear from the paper whether the final or discharge diagnosis was assessed by someone who was blind to the allocation of the diagnostic aids. For example, the authors did not mention whether the pathologists who examined the excised appendices of patients who underwent laparotomy were blind to the group to which the patients' doctors were allocated. By randomising the doctors to four groups (no decision support versus data collection forms only versus data collection forms plus output from Leeds AAP versus forms plus output and feedback), it was possible to assess the impact of each component of the intervention (i.e. providing doctors with the AAP decision tool). The authors did not specify how they determined the sample size of the trial ($n = 5193$). Unit of analysis errors could arise if the patient was used as the sampling unit in the data analysis, when it was in fact the doctor.[158,159] However, in this study the authors used nested ANOVA to take the unit of analysis into account.

Studies excluded from the review

Of the 15 papers that were retrieved for detailed review, 14 were excluded. The summary table in Appendix 10 lists the reasons for exclusion. Thirteen studies were excluded because they were neither randomised nor quasi-randomised controlled trials. The breakdown of the designs used was as follows: four uncontrolled before-and-after studies,[164–167] one externally controlled before-and-after study,[116] three interrupted time-series studies with fewer than six data points,[168–170] two interrupted time-series studies with six or more data points[171,172] and three prospective cohort studies.[51,173,174] One study was excluded because it did not ask an appropriate study question (upon scrutiny it was found to be an accuracy study only, not an impact study).[135]

Fenyo's externally controlled before-and-after study (which reported the negative appendicectomy rates and the perforation rates of suspected appendicitis patients) assessed the impacts using a Bayesian score on patient outcomes in a hospital in Stockholm.[116] Patient data from eight separate external studies were used as controls for Fenyo's score. The external controls (covering 8393 patients) reported a mean negative laparotomy rate of 29.2% (95% CI 28.2 to 30.2%), significantly higher than the rate of 15.5% (95% CI 13.2 to 18.2%) for Fenyo's score (p < 0.001). However, the external controls all reported results for only patients who have already undergone appendicectomy, whereas the group using Fenyo's score consisted of unselected patients with suspected acute appendicitis. The comparison does not rule out changes within the hospital over the study period, which might explain the impact of introducing the scoring system.

Wilson and colleagues published two switchback interrupted time-series studies (both with more than six data points) in 1975 and 1977.[171,172] The 1977 study appeared to include data from the earlier one.

Among the four prospective cohort studies, Adams and co-workers[51] reported results from a multicentre investigation involving eight participating centres, each of which used a different protocol and a different study design. In effect, they were eight separate studies. The reporting of the results as if they were all comparable and part of the same study was inappropriate. The study by Clifford and colleagues which was one of the four uncontrolled before-and-after studies excluded from this review, reported findings from another centre that had participated in Adams' investigation.[165]

In Gough's prospective study, half of the hospital registrars were non-randomly allocated to use the Leeds AAP system, while the other half were not given decision support.[174] The method of allocation of registrars was unclear. Although the number of patients included in the study was reported, the number of registrars was not. A unit of analysis error occurred by using the patient as the sampling unit, when it should have been the registrar. Fenyo and colleagues' study of a simplified scoring system was primarily a study on diagnostic accuracy. Although measures of impact were also reported, no control group was included.[173]

Summary and discussion

Summary of results

Only one out of 15 potentially relevant papers was eligible, showing that there is a clear need to improve the design and implementation of studies in evaluating the impact of AAP decision tools. The evidence base of good quality impact studies in this area is virtually non-existent.

The case for using RCTs in assessing the impact of decision tools

It is obvious from the results of this review that developers and evaluators of AAP decision tools, as well as other medical informatics applications, chose not to use RCTs to assess their impact. Heathfield and colleagues' view that the randomised trial has a limited role in evaluations of information technology is currently widespread among medical informatics professionals.[175–177] The most popular designs in the 15 studies considered were the before-and-after study (five studies, only one of which was externally controlled) and the interrupted time series (five studies). The uncontrolled before-and-after study is probably the weakest of all study designs in an evaluator's toolkit.[178,179]

Externally controlled before-and-after studies and the interrupted time series can both take limited account of known confounding factors but, like all other observational designs, cannot make adjustments for unknown confounding.[81,179,180] In clinical medicine and medical informatics, the RCT is almost universally considered to be the gold standard for assessing cause and effect or therapeutic impact.[81,148,181,182] It is the only design that deals effectively with the problem of unknown confounding. Decision tools are designed to benefit patients, so ethical issues should not provide a barrier to the use of the RCT to assess their impact, as would be the case with studying risk factors.

There is a strong need to defend the use of RCTs of decision tools, given the current level of opposition to their use in the medical informatics community. In the World Medical Informatics Conference in 2001, a session was held on the NHS Information Authority's Electronic Record Development and Implementation programme, concluding that RCTs have a limited or no role to play in the evaluation process.[183]

Some have called for a fundamental or even 'paradigm' change in the way in which medical informatics applications are evaluated.[177,184,185] There is a need to define clearly the type of problems that need to be solved and then design an appropriate study to solve them. The design can be chosen or adapted from a diverse set of existing objectivistic and subjectivistic methods of evaluation.[81] Many concerns expressed by sceptics of the RCTs in the medical informatics profession can be countered by Macintyre and Petticrew's rigorous rebuttal.[186] A large number of RCTs in non-drug technologies and decision support systems has been conducted,[148,187] although very few RCTs (only five) have been conducted for diagnostic decision tools.

Some medical informatics researchers believe that trials answer questions that do not interest them. RCTs can answer many questions of interest in medical informatics; for example, does doing *A* cause the same benefit or harm as doing *B*? Furthermore, the RCT has sound underlying logical principles.[178,188]

Conclusions

Only one cluster RCT was identified from the impact study review. This RCT, of the Leeds AAP system, showed that the use of structured data collection forms alone can significantly lower admission rates to the surgical ward. However, this is based on one primary study, so caution should be exercised in the interpretation of the review's results. A recent systematic review demonstrated that paper checklists have an impact on a wide range of patient outcomes and clinical applications.[189] With a specificity of 96%, an unaided SHO can confirm a diagnosis of acute appendicitis as confidently as SHOs who were assigned to the three 'decision tool' groups. The sensitivities of unaided SHOs and those assigned to the three 'decision tool' groups were all low, which means that acute appendicitis cannot be confidently ruled out, with or without a decision tool. There is a need for guidelines concerning the reporting and conduct of impact studies in decision tools for AAP. This will be discussed further in Chapter 8.

Chapter 6

Study question 3. What factors are likely to determine the usage rates and usability of each AAP decision tool?

Background

Chapter overview

If a decision tool is to have an impact on clinical decision-making and patient health outcomes, then usage rates for the tool must be high. Usability is thus an important attribute of a decision tool. A low usage rate would limit the impact of the decision tool and its influence on decision-making. Where usage rate data are not available (e.g. for a new tool, or lack of data for prototype tools), the usability of a tool needs to be assessed since it is likely to be an important determinant of usage rates.

User satisfaction with decision tools

User satisfaction with decision support systems is important, since no system can succeed without the support of users.[190] Although beliefs, behaviours, perceptions and attitudes affect the uptake of computer and decision tool support by healthcare professionals, a vital indicator of the tool's usefulness is the users' reactions to the system's characteristics.[190] If users are satisfied, they will modify their behaviour to use the decision tool to their advantage; if they are not, then they will not use the tool or use it in a suboptimal manner.[190]

A literature review presented a typology of factors that may influence user satisfaction with computer systems (*Figure 15*).[141] Although the authors focused on user satisfaction with computer-based systems, their typology can be applied to decision tools in general. Overall, user satisfaction is a combination of several components: system-dependent and system-independent factors. The latter include personal factors (e.g. 'computer anxiety' and 'attitudes towards computers') and organisational factors (e.g. the environment in which the system is used). System-dependent factors include 'satisfaction with the content of the decision support system' and 'satisfaction with the interface of the system'. A series of constructs relating to the factors was also identified (*Figure 15*).[141] For example, usability was found to be a marker of user satisfaction with a system's content and interface.

An evaluation or a systematic review of all aspects of user satisfaction with AAP decision tools would be a complex task. A pragmatic approach was

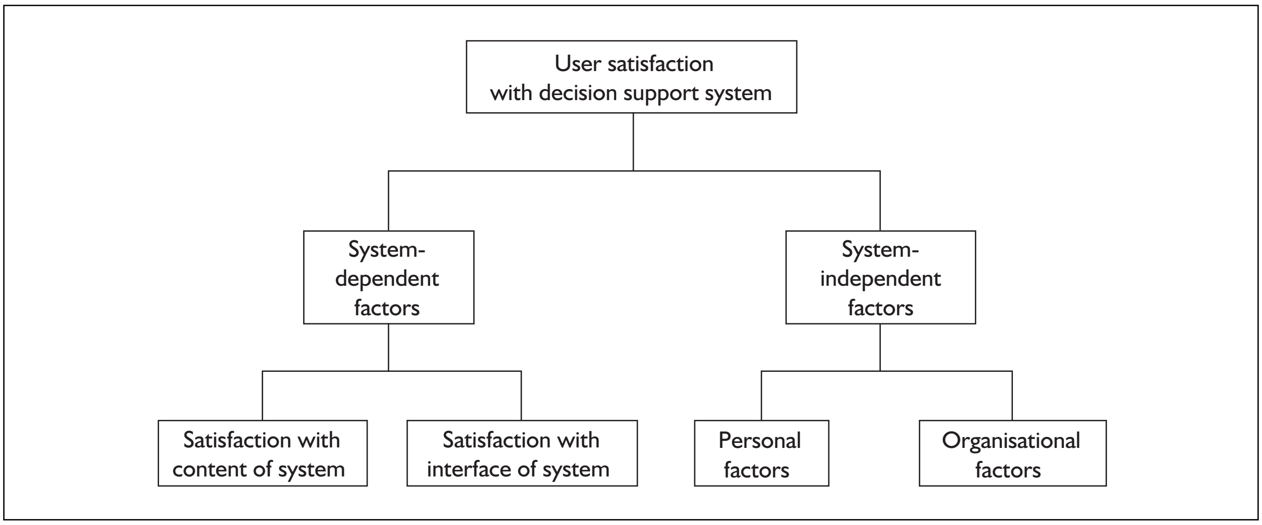
User satisfaction with decision support system
System-dependent factors
System-independent factors
Satisfaction with content of system
Satisfaction with interface of system
Personal factors
Organisational factors

FIGURE 15 *A typology of factors that influence user satisfaction with a decision support system. Adapted from Ohmann and colleagues, 1997.*[141]

therefore taken by focusing on usability. The usage rates were also extracted from the reports of the AAP decision tools identified in the systematic reviews of accuracy and impact studies reported in Chapters 3 and 5. Usage rates are assumed to be an indirect indicator of user satisfaction.

Usability and usage of AAP decision tools

The usability and usage rates of a decision tool depend on various factors:

- the time taken to obtain the decision tool
- the time taken to use the decision tool
- its conceptual complexity (which affects ease of understanding and usage)
- the number of data items to collect (if the data are not already available in electronic patient records)
- the ease of data entry
- the ease of computing the results (e.g. using a calculator to compute cardiovascular risk directly with the Framingham risk equation[191] is usually considered too complicated for clinical use, hence the development of decision tools[59,62,192])
- the ease of interpreting the results (numbers, probabilities, graphs, advice, etc.)
- the high cost of complex equipment or procedures needed to operate the decision tool (e.g. if a hospital cannot afford the required hardware, the tool cannot be used)
- the perceived applicability of the decision tool to the doctor's own patients, perhaps based on new evidence.[60,62]

Given the paucity of impact study data, the usability of decision tools in this project could be ignored, but there may be some promising and accurate tools that have never been tested in real-world clinical settings. Before recommending them, it seems sensible to assess their potential to be used in clinical practice.

One possible approach is to carry out a systematic review of past studies of usability and usage rates of AAP tools. However, there are problems with using systematic review methods to address study question 3. Few studies will have included more than one tool, so the context, including selected samples, methods to assess knowledge, attitude and perceptions, and the definitions of measures such as usability and satisfaction, will vary too much between studies for useful comparisons. Some studies may have used an obsolete method to deliver the decision tool, such as mainframe computers or slide rules, but the underlying algorithms of such tools may be sound and physicians may use them if appropriate modern techniques were employed. The impacts of clinical information on medical decisions and actions can change, depending on the format and design used.[193,194] In other words, the methods of assessment are confounded with the tools themselves. Therefore, a primary study of the usability and/or usage rates of a range of promising AAP decision tools would be likely to be more appropriate than a systematic review. A shortlist for such a primary study was therefore assembled.

Aims of Chapter 6

The aims of this chapter are:

- to report on the usage rates of AAP decision tools from retrieved studies that reported them, based on data extracted from studies in the reviews of accuracy and impact studies
- to discuss factors that may affect the usability and/or usage rates of AAP decision tools, based on data extracted from studies in the reviews of accuracy and impact studies
- to outline possible primary studies on the usability and/or usage rates of AAP decision tools.

Usage rates of AAP decision tools

Articles retrieved for data extraction as accuracy and/or impact studies were checked for reporting of usage rates. Six papers reported the rates of usage of two different decision tools.[51,136,139,165,195,196]

Harvey and colleagues[195] developed a non-sequential Bayesian program for AAP diagnosis. It was originally installed on a mainframe computer. Emergency physicians were offered the use of the decision tool in 1982 and had access to it from a remote terminal connected to the mainframe with a telephone modem. The rate of usage was 15% (95% CI 12.3 to 18.3%) of AAP cases encountered.[195] The system was transferred in 1991 to a personal computer with improved user-friendliness. A second study was carried out with a different group of doctors from the earlier study.[196] The usage rate tripled to 44% (95% CI 40.1 to 47.9%). The time taken to use the 1982 version of the tool was 6 minutes compared with 2 minutes for the 1991 version. In comparing attitudes towards the 1982 and 1991 versions of the tool, more users found the 1991 version informative, reliable, easy to use and acceptable, while fewer thought that it took too long to use. It is uncertain from the two studies to what extent the two groups of doctors are comparable to each other.

Usage rates for the Leeds AAP system were reported in two studies.[136,165] Clifford and colleagues observed that in the early phase of their study, staff interest in the computer system was high and the initial usage was 64% (95% CI 60.7 to 67.3%).[165] Interest gradually waned and the usage rate fell to 10% (95% CI 8.1 to 12.2%). The mean usage rates of structured forms and computers in the RCT by Wellwood and colleagues were 65.2% (95% CI 63.6 to 66.7%) and 50% (95% CI 47.8 to 52.2%) of cases, respectively.[136] When the decision tool was not used it was because the doctor was overworked or the case was too easy to diagnose. The average amount of time spent by doctors was 5 minutes for the computer and 3 minutes for the structured forms.

In the multicentre investigation of the Leeds AAP system (also reported in Chapter 5), the overall usage rate from the eight centres participating in the study was 77.3% (95% CI 76.6% to 78.8%) for those who used structured forms and 74.1% (95% CI 73.1 to 75.1%) for those who used the system.[51]

From the six papers above, the reported usage rates of AAP decision tools ranged from 10% to 77%.

Potential determinants of usability and usage rates of AAP decision tools

Some of the items included for data extraction in the reviews for accuracy and impact studies are likely to contribute to the usability and usage rates of AAP decision tools. These include the number and type of items used by the decision tool, the reasoning method used and the output format. A tool requiring doctors to input data for an overly large number of items is likely to be considered too cumbersome for routine clinical use. Thus, Edwards' Bayesian system,[114] which uses 35 signs and symptoms, and the Leeds AAP system,[138] which uses 36 items, could be considered less usable than tools such as the Alvarado score and Ohmann's diagnostic score, which require fewer than ten items of information.[58,129] Waiting for the results of imaging investigations or laboratory tests can be time consuming. Quick decisions are often needed in managing AAP patients. It seems reasonable to surmise that a tool that requires only the input of clinical signs and symptoms is preferable to one that requires test results. Thus, Hallan's score, which uses six signs and symptoms, fares well compared with tools such as Jawaid's score, which requires leucocyte and neutrophil counts in addition to signs and symptoms.[117,118,122] Doctors probably feel more comfortable with a decision tool that uses a simple and transparent reasoning method than one that uses a black box method.[73,197] For example, the Alvarado score,[58] obtained by adding up values assigned to eight signs, symptoms and biochemical test results, is readily understood by health professionals, whereas the underlying reasoning method of neural networks (e.g. Pesonen[198]) is opaque. A decision tool's output format could also affect its usability. Users' preferences for tools may differ between those that generate raw scores with interpretation (e.g. the Alvarado score[58]), printed lists of possible diagnoses (e.g. the Leeds AAP system[51]) or the area under an ROC curve (e.g. Pesonen's neural networks[198]).

Methods for a primary study on usability and usage rates

Alternative methods for a primary study on usability and usage rates

As mentioned in the section 'Usability and usage of AAP decision tools' (p. 46), the review approach has limitations in assessing usability and usage rates, hence the need for appropriately designed primary studies. The following alternatives can be used for AAP decision tools, or for decision tools in general:

- Nominal group or Delphi study: a panel of experts and/or potential users of decision tools convenes to decide what makes a decision tool usable using formal consensus development tools, such as a Delphi or nominal group study.[199]
- Focus group study: a focus group of potential users is asked questions about what makes a decision tool usable. The results from this focus group study can then be applied to a sample of AAP decision tools to assess their usability.[200]
- Opinion survey: a survey of a large representative sample of typical users of decision tools can be carried out to elicit their opinions on the usability of these tools.
- Conjoint analysis: analytical techniques can be used to elicit individuals' preferences for specific tools. Conjoint analysis can be carried

out in potential user groups to assess explicitly the key factors that determine usage and perceived usability.[201–203]
- Observational study to assess usability based on simulated patients: a sample of potentially useful tools can be obtained. Two of each are given to potential users with patient scenarios. They are then asked to use each tool to classify the simulated patients and, based on this experience, rate the usability of the tools.
- Observational study to assess the impact of decision tools on process measures or outcomes: observational data can occasionally provide good-quality evidence on the impact of decision tools or technology on process and/or outcome measures.
- Randomised controlled trial to assess the impact of decision tools on process measures related to usability and usage rates: a few tools for a given condition are randomly allocated to a large sample of doctors, who are asked to use them to manage their next set of patients (the precise number to be determined after sample size calculations). Outcome measures will include the time it takes doctors to use the tools and their satisfaction.
- RCT to assess the impact of decision tools on outcomes: here, usability is ignored. Instead, a pragmatic approach is taken. The impacts of the different tools are on clinical practice. Patient outcomes are assessed, assuming that the most usable tool will be used more and have more impact (e.g. reductions in rates of mortality rates, infection and perforation).

The evidence to date suggests that decision tools have not yet changed actual clinical practice, require periodic updating and may understate or overstate the actual risk faced by different populations (an example is the treatment of dyslipidaemia in primary care,[204] despite the many available decision tools for coronary heart disease[59,192,205,206]). Primary studies and systematic reviews in the past decade have shown that randomised trials of decision tools are feasible and can produce useful results.[148,207] Friedman and Wyatt discuss the principles of designing, implementing and analysing such studies.[81] The systematic review on the impact of AAP decision tools (Chapter 5) shows that this is an under-researched area.

Each of the above methods has advantages and disadvantages. Nominal group, Delphi and focus group studies are useful for assessing the degree of consensus in structured discussions, but the generalisability of these studies is debated.[208,209] However, they are useful in identifying possible determinants of usability for further research. Opinion surveys provide a good snapshot of the views of users at a particular point in time. A challenge is whether a representative sample of typical users can be found. An observational study of simulated patients may suffer from learning effects, so that rankings may be different in routine use. Observational studies using real data can be compelling, for example, if there were one million UK users of broadband last year and five million this year, that is strong evidence that users value the Internet. They are weaker at detecting moderate or small effects. RCTs have been covered in previous chapters. Two of the other methods are examined in further depth by short protocols of possible primary studies drafted below.

Outline protocol of a nominal group study

The nominal group technique

The nominal group technique is a structured method for collecting information from a panel of experts about a particular topic; in this case, factors important to usability. Members of the panel are given two rounds to rank, discuss and, based on the discussion, rerank a series of scenarios or questions related to the topic. The meeting is chaired by a facilitator and consists of the following steps:[13,210]

- Reviews of the relevant literature are provided to participants before the meeting.
- Participants spend several minutes writing down their views about each topic in question.
- Each participant, in turn, contributes one idea to the facilitator, who records it on a flipchart.
- Similar suggestions are grouped together, where appropriate. There is a group discussion to clarify and evaluate each idea.
- Each participant privately ranks each idea (round 1).
- The ranking is tabulated and presented.
- The overall ranking is discussed and reranked (round 2).
- The final rankings are tabulated and the results fed back to the participants.

The advantages and in-depth methodology of the nominal group technique over unstructured group discussions are covered elsewhere.[199,209,210]

Outline of a nominal group study to assess usability of AAP decision tools

A nominal group study to assess the usability of AAP decision tools is outlined in *Figure 16*.

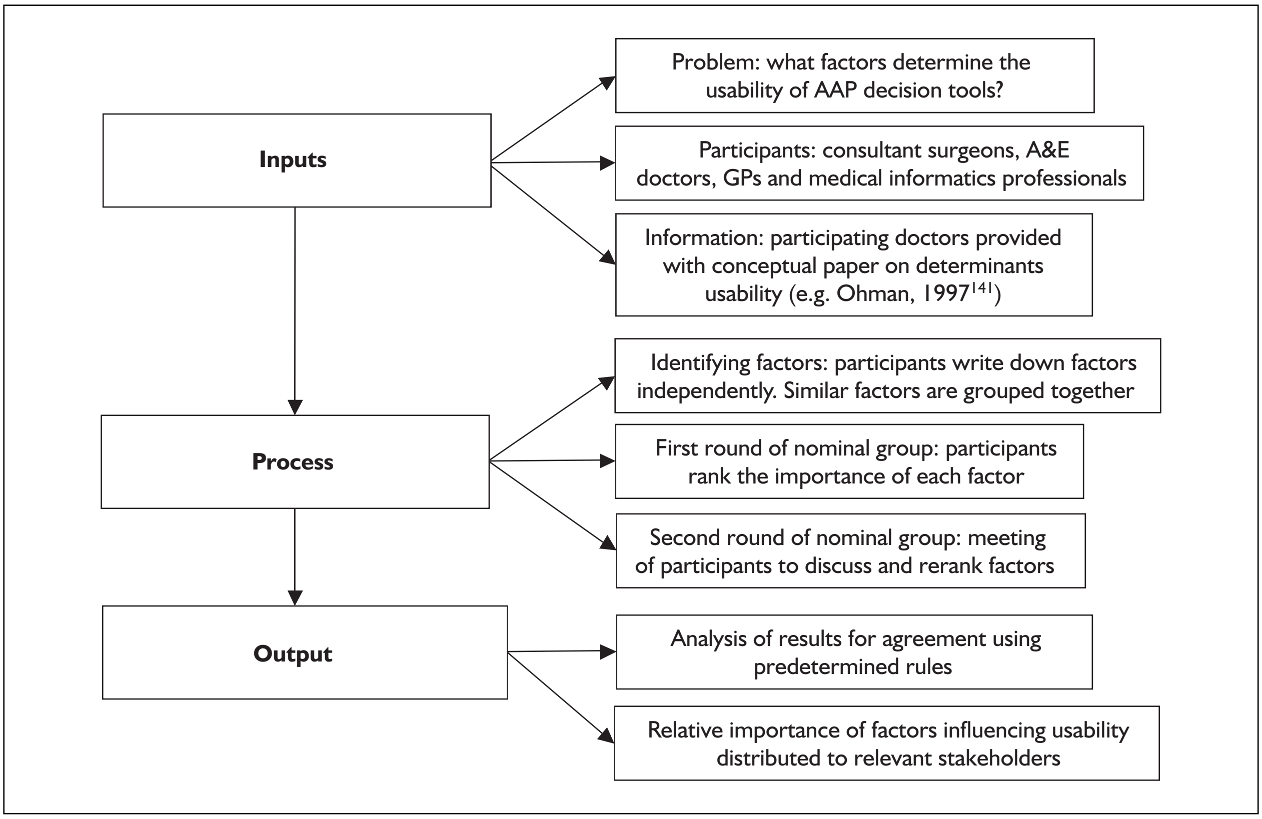
Inputs
Process
Output
Problem: what factors determine the usability of AAP decision tools?
Participants: consultant surgeons, A&E doctors, GPs and medical informatics professionals
Information: participating doctors provided with conceptual paper on determinants usability (e.g. Ohman, 1997[141])
Identifying factors: participants write down factors independently. Similar factors are grouped together
First round of nominal group: participants rank the importance of each factor
Second round of nominal group: meeting of participants to discuss and rerank factors
Analysis of results for agreement using predetermined rules
Relative importance of factors influencing usability distributed to relevant stakeholders

FIGURE 16 *Nominal group study for usability of AAP decision tools. (Adapted from: Jones and Hunter, 2000.[208])*

The basic components of the proposed study are as follows (the components were based on Jones and Hunter[13,210]):

- Definition of the question: a clear definition of the study question needs to be given. In this case, the question is: 'What factors determine the usability and usage rates of AAP decision tools?' The objective of the study is to derive a set of factors that participants agree are likely to influence usability and/or usage rates of AAP decision tools, or nobody feels enough dissent to challenge the overall agreement.[199]
- Study participants: four samples of participants will be selected: consultant surgeons, A&E doctors, GPs and medical informatics professionals. Each sample will undergo a separate nominal group exercise. Because of the different perspectives of the groups and the different case-mix encountered, their views on what constitutes a usable decision tool may differ.
- Information: the following documents will be distributed to nominal group members before the meeting: a published review paper on conceptual issues underlying the usability of decision tools or a literature review conducted by the nominal group facilitator, and an up-to-date literature review on the epidemiology of AAP and the decision-making process involved in treating AAP. While there are generic issues on the usability of decision tools, group members should also be verbally briefed on AAP epidemiology and decision-making, so that they can place the use of decision tools in the specific context of AAP. All group members will have read the papers carefully before the meeting.
- Identifying factors: at the nominal group meeting, each member will write down, in private, factors that they consider important in determining the usability of AAP decision tools. The facilitator will then ask each member to suggest one factor and classify similar factors together. A discussion then takes place between group members to assess each factor and to help to improve understanding of the issues involved.
- First round of nominal group study: group members independently use a nine-point ranking scale to rate each factor in terms of its contribution to the usability of a decision tool. The facilitator then collates and presents the results as summary tables.
- Second round of nominal group study: group members discuss the rankings and based on insights from the discussion, are given a chance

to revise their ranking in the second round after taking into consideration fellow members' views. The facilitator will emphasise to group members that they must not feel pressured to conform to the others' views and should ignore the seniority of fellow group members. However, during the discussion before the second round, members with divergent opinions from the group can be asked to explain their views. Further rounds of rating and discussion may take place if considered useful.

- Analysis, feedback and interpretation of results: Jones and Hunter define agreement as follows: "first, the extent to which each respondent agrees with the issue or statement under consideration (typically rated on a numerical or categorical scale) and second, the extent to which respondents agree with each other – the consensus element of these studies (typically assessed by statistical measures of dispersion)."[13] The methods for analysing the results are determined before the nominal group exercise begins. The interquartile range can be used to assess the degree of consensus, with the lower and upper quartiles used in a sensitivity analysis to measure the uncertainty associated with the estimates. The reporting of kappa statistics can also be considered if appropriate. The results of the analysis are fed back to group members and findings are distributed to stakeholders interested in AAP management.[13]

Outline protocol of a conjoint analysis study

The method

Conjoint analysis is a technique that can be used to establish the views of stakeholders on healthcare-related issues.[201,202] Its basic rationale is that "any good or service can be described by its characteristics (or attributes) and that the extent to which an individual values a good or service depends on the levels of these characteristics".[201] According to Phillips and colleagues, it "is an approach to measuring preferences (utilities) that estimates both overall preferences for a good or service as well as preferences for its specific attributes."[7] The technique was initially used in market research, transport economics, environmental economics and recently in health economics, but can be applied to a wide range of healthcare-related problems.[211,212] One application is in the measurement of the 'relative importance' of healthcare attributes, which lets decision-makers observe the independent effect of each attribute on 'overall benefit' (such as a usability or user satisfaction score for a decision tool).

The following are typical steps in designing and implementing a conjoint analysis exercise:[7,201]

- Stage 1. Eliciting the attributes: attributes can be elicited in several ways. If attributes are not known in advance, systematic reviews, literature surveys, interviews and group discussions can be used to identify them.
- Stage 2. Assigning a measurement scale to the attribute: the scale can be ordinal, cardinal or categorical, and should have face validity and be "plausible and actionable'[201] to convince participants (who are often stakeholders in a healthcare system) that the exercise is credible and worthwhile.
- Stage 3. Choice of scenarios: scenarios are then created by combining the varying levels of the different attributes. The number of possible scenarios rises exponentially with the number of attributes and levels.
- Stage 4. Establishing preferences: ranking, rating and discrete choices are methods used to tease out participants' preferences for the scenarios presented in the questionnaire. With the ranking method, participants are requested to rank-order their preferences for each scenario. With the rating method, participants are asked to give a preference score to each scenario. With the discrete choice method, participants are asked to make their preferred choice when presented between a set of choices (say A and B). Five-point scales can be used, in which 1 means 'definitely prefer A' and 5 means 'definitely prefer B'. Alternatively, participants can just state whether they prefer A or B.
- Stage 5. Data analysis: responses are analysed using regression models. The specific modelling method depends on the nature of the collected data.

Outline of possible conjoint analysis study on usability of AAP decision tools

A search of the MEDLINE literature using the search strategy below (1965 to 30 June 2003) did not identify a conjoint analysis study of decision support systems or in the field of medical informatics:

Search	Search terms	Hits
#1.	informatics or decision support system* or computer or aided diagnos*	27,436
#2.	conjoint analy* or discrete choice experiment*	355
#3	#1 and #2	2

The papers identified in the two hits from search #3 were irrelevant.

A possible conjoint analysis study to assess the usability of AAP decision tools is outlined below.

Study question
The objective of the study will be used to assess the relative importance of attributes that influence doctors' perceived usability of several published or simulated AAP decision tools.

Participants and setting
Samples of consultant surgeons, A&E doctors and GPs selected from a group of hospitals and GP surgeries in a geographical region in the UK. Given the different case-mix of AAP faced by the three groups of doctors and their different specialties, their perceived usability of decision tools could well be different. The three samples will therefore be analysed separately.

Study design
A structured questionnaire survey will be carried out to elicit preferences between different attributes of an AAP decision tool using conjoint analysis. A discrete choice approach is used, as it is generally considered to be superior to the rating and ranking approaches in reflecting actual choices faced by consumers in the marketplace. The determination of sample size for a conjoint analysis is complex and beyond the scope of this outline proposal. This subject is covered elsewhere.[213]

Main outcome measure
Doctors' perceived usability score for AAP decision tools.

Defining and assigning levels to attributes
Attributes of usability will be identified using, for example, a systematic review, interviews or the nominal group study outlined earlier in this chapter. Levels are then assigned to each of these attributes. Examples of possible attributes and assigned levels for AAP tools include time taken to use the tool (e.g. 2 minutes or less versus more than 2 minutes), number of items used (e.g. ten items or fewer versus more than ten items), format of output (e.g. probability, raw score, list of possible diagnoses or other), any laboratory tests or special investigations performed (yes or no), users' understanding of the decision tool's reasoning method (e.g. good understanding, vague or poor understanding), accuracy of tool (e.g. almost always accurate versus usually accurate versus less accurate). The attributes and their levels must be explicitly defined and explained to study participants in a standardised manner before the start of the study.

Scenarios
Scenarios are created by combining varying levels of attributes. Study participants are asked to choose between different scenarios (e.g. a tool that uses fewer than ten items but employs special investigations versus one that uses more than ten items but employs no special investigations). It will be assumed for now that the numbers of attributes and levels are small, which means that all possible combinations of scenarios can be presented to participants. A full factorial design can then be used. However, if there are too many attributes and assigned levels, then a fractional factorial design is used to reduce the number of scenarios.[7,202,214,215]

Establishing preferences
A discrete choice approach is used in which participants are asked to make choices for each scenario. An example of a conjoint analysis scenario is given in *Table 9*.

Data analysis and interpretation
A logistic regression model can be constructed to estimate the usability of AAP decision tools, with the general functional form as follows:[201]

$$\Delta B = \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \ldots + \beta_n X_n$$

where the dependent variable ΔB is the change in perceived usability in changing from tool A to tool B, X_1 to X_n represent the differences in levels of attributes between the tools, and β_1 to β_n represent the estimated regression coefficients.

The independent variables represent the differences in attributes between the decision tools and the regression coefficients "the relative importance of the different attributes", and demonstrate the change in usability given an incremental change in an attribute.[201] The ratio of any two of the regression coefficients demonstrates the willingness of doctors to trade off between attributes of the decision tool. Usability scores can be computed from the fitted model.

Summary

This chapter discussed the usage rates and usability of AAP decision tools and their possible determinants.

User satisfaction is a key determinant of doctors' uptake of a decision tool. A typology of user satisfaction was discussed.[141]

TABLE 9 *Example of conjoint analysis question for AAP decision tool*

	Decision tool A				Alternative decision tool B			Which tool would you find more usable? (please tick one for each choice)				
	Time to complete	Sensitivity in diagnosing appendicitis	Reasoning method	Or	Time to complete	Sensitivity in diagnosing appendicitis	Reasoning method	Definitely A	Probably A	No Preference	Probably B	Definitely B
Choice 1	15 minutes	95%	Bayesian	or	5 minutes	90%	Paper checklist					
Choice 2	15 minutes	95%	Bayesian	or	15 minutes	98%	Neural network					
Choice 3	15 minutes	95%	Bayesian	or	5 minutes	95%	Stepwise regression					

Table format adapted from Ryan and Farrar.[201]

Usability is one of the constructs of the typology. Because of budgetary constraints, this chapter focused on usability and usage rates as surrogates for user satisfaction.

The limitations of a review approach in assessing usability and usage rates of AAP decision tools were discussed. A primary study is needed to compare directly the usability and usage rates of a range of tools against each other.

With these limitations in mind, the usage rates of AAP decision tools were extracted from studies retrieved for the accuracy and impact study reviews. Factors that may be associated with usability and usage rates were discussed.

A set of alternative methods for a primary study was outlined.

Draft protocols for possible primary studies on assessing usability were outlined.

Chapter 7

Study question 4. What are the associated costs and likely cost-effectiveness of these decision tools in routine use in the UK?

Background

The systematic review of impact studies indicates that insufficient data exist for a cost-effectiveness analysis of the identified decision tools. Little good-quality information exists on the impact of AAP decision tools on patient health outcomes. However, various issues related to economics in this area can still be explored.

- The few studies that were identified on the economics of managing AAP will be discussed, although the papers that reported outcome measures were not of a sufficiently good quality to be used in a cost-effectiveness analysis using advanced methods (such as decision-analytic, stochastic or partially stochastic models). However, useful insights can still be gained by analysing the available information using the basic principles of economics.
- The barriers to cost-effectiveness evaluations of AAP decision tools or in medical informatics in general will be discussed.
- The policy implications of the state of research and development (R&D) in AAP decision tools will be discussed.

Economic aspects of AAP decision tools in the literature

Of the 489 papers retrieved for the impact and accuracy reviews, only eight were found that possibly contained information on the cost-effectiveness of AAP decision tools.[51,171,172,216–220] Of these eight papers, five attempted to examine the costs and benefits of AAP decision tools. One other paper was identified through other means.[221]

De Dombal's economic analysis of the Leeds AAP system

The Leeds AAP system will be looked at first, since it provided the most information on the system's potential cost savings for the NHS. An appraisal will be conducted on a cost–benefit analysis,[219] which elaborated or updated on the economic data presented in other Leeds studies.[51,171,172,217] De Dombal[219] reported that the study by Adams and colleagues[51] demonstrated conclusively that the Leeds AAP system reduced length of hospital stay, the number of AAP patients admitted to hospital, perforated appendix rates and negative laparotomy rates. Based on these figures, De Dombal estimated that at a nationwide level, annual savings to the NHS would amount to £28 million, and reported the cost of each computer system at £3000 and each part-time computer assistant at £3000 per year (1986 prices).[219] De Dombal did not provide overall nationwide estimates of these costs, but argues that the savings from implementing the Leeds AAP system across the country would outweigh the costs.

De Dombal's analysis had several weaknesses, including; first, a population-based cost estimate of implementing the Leeds system was not provided, only the cost of each individual computing system. Secondly, a decision index was not presented in the analysis, for example, a net present value (sum of benefits minus the sum of costs discounted to convert projected future costs/benefits into the present value[2]) or a discounted benefit–cost ratio to give the reader some idea of the magnitude of potential improvement in economic efficiency using the tool (see the Glossary for definition of discounting, under 'Discount rate'). Thirdly, a health service perspective was used. It would have been helpful to have also included a societal perspective, which is the preferred practice today.[222–224] Fourthly, an inappropriate comparator was used. The Leeds system was compared with the status quo. One of Drummond's recommendations in his health economics texts on conducting an economic evaluation is that a comprehensive set of feasible alternative interventions should be considered, described and compared in an incremental analysis.[222,223] As De Dombal admitted, one of the advantages of the computer system was that "the doctor is encouraged to follow 'good' clinical practice" and "the analysis may draw the doctor's attention to some point forgotten due to distraction or tiredness". He was referring to the checklist effect of the computer system, described by Friedman and Wyatt.[81]

Although cost-benefit analysis has been extensively applied in non-health areas such as environmental economics,[2,225] and exhaustively theorised,[226–229] cost-effectiveness analysis and cost–utility analysis are much more commonly used in healthcare, mainly because of difficulties in assigning monetary value to health benefits.[2,223,230] Cost-effectiveness analysis and cost–utility analysis identify the programme option with the least cost per unit of effect (e.g. life-years gained or quality-adjusted life-years gained). They are more limited in their aims than cost–benefit analysis, in that they cannot directly identify the most economically efficient policy option, since the benefits have not been monetised, but they also avoid some of the problems of benefit valuation.

Gorannsson and Lasson's cost estimates of various diagnostic methods for AAP

In a non-systematic literature review of the accuracy and costs of diagnostic methods for AAP in Sweden, the costs, sensitivity and specificity of 18 methods were assessed.[218] The accuracy measures were given a score on an ordinal scale by the authors for each method. It appears the authors did this because they did not attempt to synthesise the data from the studies they identified. The costs of the diagnostic methods ranged from $9 (£5.75) for C-reactive protein, through $450 (£288) for computed tomography to $650 (£416) for angiography. The authors made subjective assessments of sensitivities and specificities for clinical examination, scoring system and computer aid, and showed they were similar, but less than computed tomography and angiography. It is unclear why a subjective assessment was made when quantitative estimates of the performance of these tests are available from the literature.

Gill and Jenkins' cost-effectiveness of management options for AAP

Gill and Jenkins reported they had conducted a cost-effectiveness analysis of management options of the acute abdomen.[220] The study was excluded because it did not study a decision tool, and was therefore considered ineligible for the systematic reviews. Nevertheless, its subject matter is of interest to this chapter. The authors presented a series of tables with number of patients, average length of stay and average charges stratified by various factors: expected payment source, patient's residence and patient's age group. As the authors correctly noted, it is preferable to use cost data than charge data in an economic evaluation as the former better reflect resource use.[223] The average cost data were presented for two medical centres where they were available, stratified by various diagnostic related groups related to appendicectomy (complicated with co-morbidities, non-complicated without co-morbidities, non-complicated with co-morbidities and complicated without co-morbidities). Then, without providing any data or analysis to back up their claim, the authors concluded that "the single most cost-effective way to evaluate and manage a patient with an acute abdomen is to consult the responsible general surgeon immediately upon hearing the patient's history so that the surgeon can direct the diagnostic work-up". It appears that the authors had not conducted a cost-effectiveness analysis.

This study's cost-effectiveness comparison of the Leeds AAP system with structured data collection form

The RCT of the Leeds AAP system[136] reported in Chapter 5 found that the benefit (in terms of admissions rate) could be attributed to the use of structured data collection forms, implying that the impact of the computer system can be explained by the checklist effect. The absolute reduction in admissions rate was virtually the same as using the computer alone. Friedman and Wyatt reanalysed the data on crude diagnostic accuracy in Adams[51] and found that, of the 27% improvement observed 6 months after implementing the system, 1% was attributable to the computer advice, 14% to the checklist effect and 13% to the monthly feedback given to doctors.[161] The reader should be reminded that Friedman and Wyatt's reanalysis was based not on an RCT, but on a study that was a composite of studies at eight centres with different design protocols.

Wellwood and colleagues therefore recommended that "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",[136] an option that appears to be much more cost-effective than the Leeds AAP system. It may be argued that at today's prices, a PDA version of the Leeds AAP system would be much cheaper. However, the cost of a PDA alone (without estimating development and maintenance costs at today's prices) is still in £70–450 range (the prices were checked in December 2004 on the website of PC World, http://www.pcworld.co.uk). A structured paper data collection sheet costs much less, at less than 50 pence per sheet of A4 paper (the prices were checked in December 2004 on the website of Office World, http://www.office-world.co.uk).

The appropriate comparisons in an economic evaluation are not between computer-based AAP decision tools only. The relevant comparison is between incremental cost-effectiveness ratios of AAP decision tools and paper-based decision tools such as checklists, the doctor's unaided diagnosis and special investigations such as computed tomography. In health economics, if intervention *A* is less costly than but as effective as or more effective than intervention *B*, intervention *A* is said to dominate intervention *B*. In the analysis above, the structured paper checklist appears to dominate the Leeds computer system.

Given the prices of the two decision tools and the similar impact that they have on patient outcome, a paper checklist is likely to be 100–900 times more cost-effective than a computer-based decision tool. It must be emphasised this is based on one RCT, so this result should be interpreted with caution.

It can be argued that there is no need here for an advanced cost-effectiveness evaluation (e.g. using the state-of-the-art 'net benefit' approach[231] or the traditional decision-analytic approach to cost-effectiveness analysis), since the results of such a study would be so obvious. However, counter-arguments can also be made. There are likely to be additional advantages to using a PDA version of the Leeds AAP system. Changes in the format of presenting information are known to have an effect on the impact of a decision tool.[151,194] Thus, there is the possibility that this 'format' effect might improve its performance, although the improvement would have to be huge to overcome the lower cost advantage of the structured data collection sheet.

A second, more important, counter-argument is that, if the PDA is closely integrated with the environment in which it is used (e.g. the hospital or the GP clinic), it could in some instances be an efficient way to store most patient information electronically, not only about AAP. This is one of perhaps several benefits of using this technology, and an economic evaluation would need to take these additional external benefits into account. Whether a PDA could then be more cost-effective than a paper data collection sheet could depend partly on the existence of economies of scale,[232] (e.g. a decrease in operating costs of PDAs as a result of an increase in the scale of operations of a hospital or clinic) and economies of scope,[232] (e.g. a hospital that has just installed an electronic patient record system may find the incremental cost of installing a PDA version of the Leeds system to be minimal).

For the compared technologies, the existence of economies of scale and scope could vary according to the clinical environment in which a decision tool is used. There may be decreasing returns to scale (i.e. output increases less than in proportion to inputs as the firm's production expands, in which case overall costs would increase[232]) if the Leeds PDA were installed in a small local GP surgery, but increasing returns to scale (i.e. output increases more than in proportion to inputs as the firm's production expands, in which case overall costs would decrease[232]) if it were used in a large hospital with an integrated electronic patient record system.

There are mixed opinions on the importance of the issue of returns to scale in an economic evaluation.[223,224,233] The conventional practice is to assume constant returns to scale.[223]

To conclude, the structured data collection sheet is likely to be much more cost-effective than the Leeds PDA under assumptions of constant returns to scale (the conventional assumption) or decreasing returns to scale, but less certain under the assumption of increasing returns to scale. More research is needed to assess the generalisability of this finding.

Barriers to cost-effectiveness evaluations

The lack of economic evaluations of AAP decision tools may be due to a number of reasons. There is a general lack of good-quality evidence on patient health outcomes in the literature, one of the key inputs for a cost-effectiveness analysis. Ideally, these should be in the form of life-years gained or quality-adjusted life-years gained, but no AAP study in the systematic reviews reported these measures. The outcomes reported included negative laparotomy rates, positive appendicectomy rates, rates of postoperative complications and admission rates. It is possible to conduct economic evaluations using these outcomes if a robust evaluation method is used (i.e. randomised or quasi-randomised trial, or at the very least a cohort study that adjusted for known confounders).

Life-years gained and quality-adjusted life-years gained are the preferred outcomes because conceptually they enable incremental cost-effectiveness ratios of different interventions from different disease areas to be compared directly with each other, possibly in a comprehensive standardised cost-effectiveness league table.[234,235] Together with healthcare utilisation figures, population figures and estimates of a health authority's maximum willingness to pay for a healthcare intervention, league tables can theoretically be used to aid decision-makers with making resource allocation decisions that are indicative of economic efficiency.[236]

There is a lack of currently relevant evidence on key resource use variables in the AAP decision tool literature in the UK. There also appears to be a paucity of good-quality economic evaluations in other areas of medical informatics. Most evaluations on the impact of decision tools are carried out with no or little mention of cost-effectiveness (e.g. Dexter[237]). A recent systematic review of cost-effectiveness studies of telemedicine interventions indicated that of 612 articles identified, four used cost-effectiveness analysis and none used cost–utility analysis.[238] The few eligible studies were of poor quality and the authors of the review concluded, "there is presently no persuasive evidence about whether telemedicine represents a cost-effective means of delivering health care".

Medical informatics researchers may not be as aware as biomedical researchers of the standards set by the US Public Health Panel on Cost Effectiveness Analysis in Health and Medicine.[224,239] Developers of AAP decision tools may be concerned that their computer-based systems are not cost-effective compared with other tools; or there may be the perception that economic evaluations will not do their applications justice, since computer technology may initially be very expensive but the price often falls rapidly as the application matures. More work needs to be done to elicit medical informatics researchers' attitudes towards economic evaluations and the barriers to the implementation of these studies. It may be useful for clinicians and developers of decision tools to collaborate more closely with health economists at the beginning of their studies.

Policy implications of the research base in AAP decision tools

A lot of time, effort and money has been used to develop and evaluate AAP decision tools (particularly computer-based tools) since the early 1970s, with relatively little progress. Research councils need to be more careful when considering future funding applications for research in this area.

Summary and discussion

There are insufficient data on the impact of decision tools to conduct a cost-effectiveness analysis, based on stochastic or decision-analytic modelling.

The few papers that were identified on the economics of decision tools in AAP were discussed in depth. In particular, it was noted that the benefits of the Leeds AAP system (in terms of both improvements in crude accuracy and impact measures) mainly appear to be attributable to the checklist effect. Further research is needed to assess whether this finding from one study is generalisable.

A basic cost-effectiveness comparison was made and concluded that a paper checklist is likely to be far more cost-effective than a computer system, even at today's prices, assuming constant or decreasing returns to scale. Although the analysis was basic, it adhered to the fundamental principles of economics and provided valuable insights.

Economics publications on AAP are rare. None used life-years gained and quality-adjusted life-years gained as the denominators to incremental cost-effectiveness ratios. Medical informatics researchers should be made more aware of the standards set by the US Public Health Panel on Cost Effectiveness in Health and Medicine.[224,239]

The policy implications were discussed.

Chapter 8

Reporting guidelines for decision tool performance evaluations and impact studies

Standardised reporting of studies and a common understanding of basic concepts between researchers from different disciplines would make it easier for clinicians and researchers to understand published articles and systematic reviewers to synthesise the evidence. Reporting guidelines exist for various study types, some of which are relevant to studies of decision tools.

Recommendations for good practice in the reporting of decision tool evaluations

The recommendations in the following two sections 8.2 and 8.3 below were adapted from STARD[86,87] and CONSORT[150] for the purposes of evaluations of decision tools and decision support systems. The discussion of the items in each checklist was tailored to the context of decision tools. There is recognition of the need to tailor CONSORT (and arguably STARD) to different study designs and areas of implementation (e.g. herbal medicine and decision support systems).[240] CONSORT guidelines have recently been tailored to cluster-randomised trials and the reporting of harms.[241,242] Comments that are particularly relevant to decision tools were added, based on the results of the systematic review, on reading the medical informatics literature and discussions with doctors who are potential users and medical informatics researchers. The adapted STARD and CONSORT recommendations were the result of an informal process between the report authors over a period of months. These included a consultant surgeon, a GP, statisticians, a health economist, two medical informatics practitioners and two medical informatics practitioners, one of whom is a lapsed physician. These modified guidelines should be treated as a first attempt for subsequent polishing by a group of experts.

Recommendations for the reporting of studies measuring accuracy of decision tools

Table 10 shows the original checklist for STARD. The examples and notes on the checklist items below relate to the implementation of STARD for decision tools. They are meant to supplement the examples and comments in the original STARD explanatory document and pertain specifically to decision tools. In the modified guidelines, the terms ‘diagnostic decision tool’ or ‘decision tool’ should be used to replace ‘test’ and ‘index test’ in the original STARD checklist. Examples are hypothetical except for those that are referenced.

Title, abstracts and keywords

Item 1. Identity the article as a study of diagnostic accuracy (recommend MeSH healing ‘sensitivity and specificity)

Example(s) Title: Diagnostic accuracy of a scoring system for acute abdominal pain: a prospective cohort study.

Abstract: Objective – To estimate the sensitivity and specificity of the scoring system for diagnosing acute abdominal pain by using a two-year follow-up of all patients as the reference standard.

Notes Mention in the abstract that the study is on decision tools and report the comparisons or estimates made. Also propose the term ‘decision tool’ as a keyword.

Introduction and objectives

Item 2. State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups

Example(s) This study compares the diagnostic accuracy of decision tool X with the performance of unaided senior house officers.

The study’s primary objective is to compare the diagnostic accuracy of decision tool Y in different patient subgroups (by age groups and gender).

Methods

Participants

Item 3. Describe the study population: the inclusion and exclusion criteria, setting and locations where the data were collected

TABLE 10 Checklist of items to include when reporting the diagnostic accuracy of a diagnostic decision tool (the STARD statement)[87]

	Item	Descriptor
Title/abstract/ keywords	1	Identify the article as a study of diagnostic accuracy (recommended MeSH heading 'sensitivity' and 'specificity')
Introduction	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between diagnostic accuracy between tests or across participant groups
Methods		**Describe**
Participants	3	The study population: the inclusion and exclusion criteria, setting and locations where the data were collected
	4	Participant recruitment: was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	Participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected
	6	Data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?
	7	The reference standard and its rationale.
Test methods	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for the index tests and reference standard
	9	Definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence interval)
	13	Methods for calculating test reproducibility, if done
Results		**Report**
Participants	14	When study was done, including beginning and ending dates of recruitment
	15	Clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centres)
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended)
Test results	17	Time interval from using the index tests to the reference standard, and any treatment administered between
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition
	19	A cross-tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard
	20	Any adverse events from performing the index tests or the reference standard
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)
	22	How indeterminate results, missing responses and outliers of the index tests were handled
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centres, if done
	24	Estimates of test reproducibility, if done
Discussion	25	Discuss the clinical applicability of the study findings

Example(s) Patient population: Patients with suspected acute appendicitis admitted to surgical wards in hospitals X, Y and Z in London during 2003.

Inclusion and exclusion criteria: Eligible patients include those admitted to surgical wards with previously undiagnosed acute abdominal pain for 7 days or less. Patients with recurrent abdominal pain were excluded as the aetiologies of the two groups of conditions are different.

Notes If special groups or subgroups are studied, mention whether they were prespecified or post hoc. If adults only are studied, then investigators should define the age threshold for an adult. The type of demographic information to be collected needs to be specified in the methods section of a research protocol and summarised in the corresponding section of the article. Evaluations of decision tools often fail to report these demographic characteristics.

Was the study conducted in a surgical department, hospital ward, A&E or GP surgeries? It may be insufficient to report 'secondary care', as the case-mix of patients could vary between different secondary care departments. This may have implications for the estimated diagnostic accuracy of the decision tools.

Item 4. Describe participant recruitment: was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?

Example(s)

i. Patients presenting with acute abdominal pain lasting for 7 days or less were recruited
ii. Patients given a diagnosis with the aid of a decision tool were recruited.
iii. Patients who had undergone a negative laparotomy (as indicated by histopathology) were recruited.

Item 5. Describe participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected

Example(s)

i. Consecutive patients presenting with acute abdominal pain lasting for 7 days or less admitted to A&E using a prospective cohort design were selected.
ii. All patients who had acute abdominal pain and their appendix removed in the surgical ward were recruited to participate in the study.

Notes Good reporting of the sampling scheme (such as examples i and ii above) is important, because it will help readers to assess the internal and external validity of findings about the decision tool.

Item 6. Describe data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?

Test methods

Item 7. Describe the reference standard and its rationale

Example(s) The reference standard consisted of histopathology of the excised appendix and the follow-up of all discharged patients, including those who were not operated on.

Notes The purpose of the follow-up is to ascertain the subsequent health status of all patients and whether any were falsely classified as not having the condition or falsely classified as having the condition. The follow-up helps one to assess how often such errors are made. Many decision tool studies fail to do this.

Item 8. Describe technical specifications of materials and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard

Example(s) The decision tool was developed using naïve Bayesian analysis with training-set data from 200 consecutively recruited patients with acute abdominal pain in the A&E department of hospital W in city X. It was tested on a test set of 1250 consecutively recruited patients from a surgical ward in hospital Y in city Y by independent evaluators. The decision tool was used before the choice was made on whether to operate on the patients. After the operation, the reference standard was carried out. It consisted of histopathology of the excised appendix and a 1-year follow-up of patients who were discharged without an operation.

Notes The following points are of particular relevance to decision tools:

i. Describe the development of the decision tool or cite references where this information can be found.
ii. Describe or cite the reasoning method of the decision tool(s) (e.g. independent naïve Bayes, logistic regression model, statistical model based on discrimination rules, machine learning).
iii. If available or relevant, describe the method of diagnosis used by the unaided doctor. What signs, symptoms and other indicants had he or she obtained from the patient?
iv. Identify the article as describing either a training-set study, test-set validation study, or both.
v. If the article reports a training-set study, describe the training-set sample used.

vi. If the article contains a test-set study, describe the test-set sample used. The main options are: (a) prospective test-set data collected from one or more centres away from where the tool was developed (best type of test-set data); (b) prospective data collected in the centre where the tool was developed (c) non-random split samples; (d) random split samples or samples generated using a resampling technique (such as bootstrapping or jack-knife); and (e) use of training-set data as test-set (worst type of test-set data).

vii. If the article contains a test-set study, clarify whether the developers of the decision tool or independent evaluators carried out the evaluation.

Item 9. Describe definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard

Example

i. The cut-point of the Ottawa ankle rules was chosen to achieve high sensitivity, since the rules were designed to rule out fractures of the ankle and mid-foot.[243]

ii. The reference standard chosen to assess the accuracy of the decision tool for acute appendicitis is histopathological examination of the excised appendix and final diagnosis at discharge for patients who had not undergone surgery.

iii. Patients are followed up for 1 year upon discharge, regardless of whether they had undergone an appendicectomy. This allows the investigators to determine the health status of patients who were discharged without an operation and any subsequent complications of patients who were discharged after an operation.

Notes No reference standard is perfect, whether in health informatics or in other areas of healthcare research. Friedman and Wyatt mention that in assessing the diagnostic accuracy of a decision tool, a patient's diagnosis at discharge is often used as the reference standard.[81] The discharge diagnosis may later turn out to be incorrect for some patients. Ideally, the patients should be followed up for a period of time to confirm the diagnosis was indeed correct, but for pragmatic and logistical reasons, this is often not possible or is difficult. The reference standard is accepted, as it is the best available approximation of the unachievable perfect standard. Because reference standards are often 'fuzzy' or imperfect, some health informatics researchers have argued against the use of reference standards in assessing the accuracy of a decision tool. Friedman and Wyatt argue that imperfect standards are better than no standard.[81]

Item 10. Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard

Example Two groups of users of a diagnostic decision tool were assessed: consultants and house officers.

Notes In the example above, the seniority of the doctor may have an effect on the usefulness of the decision tool. Senior doctors may tend to be more accurate than junior doctors in diagnosing a particular condition and stand to benefit less from decision support.

Item 11. Describe whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers

Statistical methods

Item 12. Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty

Example(s) The relative diagnostic odds ratio was used to assess the differences in observed diagnostic ability between the decision tool and the unaided doctor.

Notes Appropriate measures of diagnostic accuracy should be used and reported. Possible measures include sensitivity, specificity, positive predictive value, negative predictive value, likelihood ratios for positive and negative test results, the ROC curve (with a table of sensitivity and specificity at different cut-points), the area under the ROC curve (with the raw data used to calculate the area), relative true-positive rates and relative false-negative rates, the diagnostic odds ratio and the relative diagnostic odds ratio. Crude accuracies are of limited use and should not be reported because they entangle true-positive rates among those with the target disorder and true-negative rates among those without the target disorder. Crude accuracies are commonly reported in decision tool studies.

Item 13. Describe methods for calculating test reproducibility, if done

Example(s) The level of inter-observer agreement in interpreting the output of a decision tool was assessed by using the kappa statistic with a 95% confidence interval.

Notes The reproducibility of the results obtained from a diagnostic decision tool is often not assessed or not reported, yet poor reproducibility can have an adverse effect on a decision tool's accuracy.

Results

Participants

Item 14. Report when study was done, including the beginning and ending dates of recruitment

Item 15. Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centres)

Item 16. Report the number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended)

Test results

Item 17. Report the time interval from the index tests to the reference standard, and any treatment administered between

Item 18. Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition

Item 19. Report a cross-tabulation of the results of the reference standard (including indeterminate and missing results); for continuous results, the distribution of the test results by the results of the reference standard

Notes: If results of unaided doctors' diagnosis are available then report a cross-tabulation for them as well.

Item 20. Report any adverse events from performing the index tests or the reference standard

Example(s)

- Errors in reference database calculation of Down's syndrome screening, giving false-negatives, Sheffield.[74]
- Errors in updated embedded clinical coding software giving false plain-language representation of diagnoses, UK.[74]

Notes Unlike drugs, decision tools, including computer-based tools, are currently exempt from regulation. However, this may change.[244] It is well documented that errors sometimes occur when decision tools are used.[74] These adverse events should be reported.

Estimates

Item 21. Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)

Example(s) The measures of diagnostic accuracy used include the diagnostic odds ratio, relative true positive rates and relative true-negative rates.

Notes If relevant, report estimates of unaided doctors' diagnostic accuracies and measures of statistical uncertainty. If relevant, formally compare the diagnostic accuracy of decision tools with unaided doctors' diagnosis using an appropriate statistical test and a suitable summary measure of diagnostic accuracy, for example, the diagnostic odds ratio or the relative error rate.

Item 22. Report how indeterminate results, missing responses and outliers of the index tests were handled

Item 23. Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centres, if done

Item 24. Report estimates of test reproducibility, if done

Discussion

Item 25. Discuss the clinical applicability of the study findings

Notes Given the diverse ways in which discussion sections are written in decision tool studies (e.g. AAP decision tool studies as well as many other studies of medical informatics systems), the tendency sometimes to overinterpret results, the habit of some researchers to report methods and new results in the discussion section instead of the relevant sections, and the perhaps understandable lack of maturity in a young science such as health informatics,[245] there may be a case to be made to go beyond STARD's original recommendations for discussion sections.

TABLE 11 Checklist of items to include when reporting an RCT of a decision tool (the CONSORT statement)[150]

	Item	Descriptor
Title and abstract	1	How participants were allocated to interventions (e.g. 'random allocation'; or ' 'randomised' or 'randomly assigned')
Introduction		
Background	2	Scientific background and explanation of rationale
Methods		
Participants	3	Eligibility criteria for participants and the settings and locations where the data were collected
Interventions	4	Precise details of the interventions intended for each group and how and when they were actually administered
Objectives	5	Specific objectives and hypotheses
Outcomes	6	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (e.g. multiple observations, training of assessors)
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules
Randomisation		
Sequence generation	8	Method used to generate the random allocation sequence, including details of any restriction (e.g. blocking, stratification)
Allocation concealment	9	Method used to implement the random allocation sequence (e.g. numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned
Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups
Blinding (masking)	11	Whether or not participants, those administering the interventions and those assessing the outcomes, were blinded to group assignment. If done, how the success of blinding was evaluated
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analyses and adjusted analysis
Results		
Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analysed for the primary outcome. Describe protocol deviations from study as planned, together with reasons
Recruitment	14	Dates defining the periods of recruitment and follow-up
Baseline data	15	Baseline demographic and clinical characteristics of each group
Numbers analysed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was by 'intention to treat'. State the results in absolute numbers when feasible (e.g. 10/20, not 50%)
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (e.g. 95% confidence interval)
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory
Adverse events	19	All important adverse events or side-effects in each intervention group
Discussion		
Interpretation	20	Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of outcomes
Generalisability	21	Generalisability (external validity) of the trial findings
Overall evidence	22	General interpretation of the results in the context of current evidence

Discussion sections structured with the following headings are suggested:[246]

- Statements of principal findings
- Strengths and weaknesses of the study
- Strengths and weaknesses in relation to other studies, discussing particularly any differences in results
- Meaning of the study: possible mechanisms and implications for clinicians or policy makers
- Unanswered questions and future research.

Recommendations for the reporting of studies measuring impact of decision tools

Table 11 shows the updated checklist for CONSORT, published in 2001. Examples and/or notes on the checklist items are given where appropriate. They are meant to supplement the examples and comments in the original CONSORT explanatory document and pertain specifically to decision tools.

Title and abstract

Item 1. How participants were allocated to interventions (e.g. 'random allocation', 'randomised', or 'randomly assigned')

Example(s) *Title:* Computer-aided diagnosis compared with unaided doctors' diagnosis: a cluster randomised trial of impact on management of acute abdominal pain.

Abstract (Design): Cluster randomised controlled trial with ward as the unit of randomisation.

Notes: The objective of the impact study should be clearly indicated in the title and/or abstract. Specify the sampling unit and method of allocation of sampling units to decision tools (e.g. randomisation, cluster randomisation). Examples of sampling units include the patient, doctor, ward and hospital.

Introduction

Item 2. Background: scientific background and explanation of rationale

Example(s) Making accurate decisions with patients with acute abdominal pain (AAP) is difficult in secondary care because many conditions cause it and no single clinical finding or laboratory test is both specific and sensitive. About half of hospital inpatients with AAP have a non-specific cause, but many of the remainder have acute appendicitis or other conditions requiring emergency surgery. To avoid missing these seriously ill patients, large numbers are referred for unnecessary admission and surgery, with negative laparotomy rates of up to 25%. However, delays can lead to a perforated appendix in 20% of cases. What makes clinical decision-making in this area particularly challenging is the trade-off between the perforated appendix rate and the negative appendicectomy rate. As a result of these difficulties, many computer-based diagnostic decision tools have been developed to aid the management of AAP. It is unclear which, if any, of these tools are effective in improving patient management. No RCTs on this area were identified in a literature search.

A cluster randomised trial investigated whether doctors' use of a new computer-based decision tool improved the diagnostic accuracy and management of patients, compared with doctors who are not using any tool.

Notes The scientific basis for the study needs to be explained. Having outlined the problem, the authors should examine the evidence on the impact of different types of decision tools and special investigations, and the impacts of unaided doctors' diagnosis. Peer-reviewed papers or, if available, systematic reviews should be cited and discussed. The rationale for a study on a new decision tool would be stronger if the evidence indicated that other tools have little impact and are unlikely to be cost-effective. Need should be based on the existence of a clinical problem, not the appearance of a new technology.[247] Cluster designs are commonly used in decision tool studies. If a cluster design was used, this should be mentioned.

Methods

Item 3. Participants: eligibility criteria for participants and the settings and locations where the data were collected

Example(s) *Eligibility:* All patients presenting with previously undiagnosed acute abdominal pain for 7 days or less were eligible for inclusion in the study.

Setting: A&E.

Location: Hospitals A, B and C in city X.

Notes In an RCT of diagnostic decision tools, there is often a risk of the carry-over effect, which is "a contamination of the management of the 'control condition' by care providers who also have, or have previously had access to the [decision tool]."[81] To reduce the risk of contamination, a cluster design is often used, where the doctor (instead of patient), ward or even hospital is used as a sampling unit. If cluster sampling was used, the eligibility criteria for patients and clusters should be specified.

Item 4. Interventions: precise details of the interventions intended for each group and how and when they were actually administered

Example(s) Senior house officers were allocated to one of four groups: (a) no decision tool allocated; (b) data collection forms only; (c) data collection forms and output from Leeds Acute Abdominal Pain system; and (d) forms, output from the Leeds system and monthly feedback of doctors' performance.[136]

Notes In the example, the four groups indicate that a multiarm design was used. The checklist effect is "the improvement observed in decision-making due to more complete and better-structured data collection when paper- or computer-based forms are used to collect patient data".[81] This checklist effect may explain part of any observed effect of the decision tool, hence the need to quantify the effect.

Many decision tools "provide the doctors with the opportunity to capture their diagnoses on a form", which may encourage doctors to audit their own performance.[81] This feedback effect may explain part of any observed effect of the decision tool, hence the need to quantify the effect.

The following details of the decision tool(s) should be included:

i. Describe the development of the decision tool(s) or cite references where this information can be found.
ii. Describe the reasoning method of the decision tool(s) (e.g. independent naïve Bayes, logistic regression models, statistical models based on discrimination rules, machine learning) or cite references where this information can be found.
iii. If available, describe the method of diagnosis used by the unaided doctor. What signs, symptoms and other indicants had he or she obtained from the patient? The characteristics of clusters should be described, e.g. doctors (seniority, speciality) and the healthcare centres involved.
iv. Describe how each decision tool was actually used for each assigned patient and/or cluster of patients.

Item 5. Objectives: specific objectives and hypotheses

Example(s)

i. The objective of the study was to assess the impact of doctors using the decision tool(s) on patient outcomes, appropriateness of clinical decisions, actions and/or use of healthcare resources.
ii. The object of the study was "to evaluate the use of a computerised support system for decision making for implementing evidence based clinical guidelines for the management of asthma and angina in adults in primary care".[149]

Item 6. Outcomes: clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (e.g. multiple observations, training of assessors)

Example(s) Mortality rates, rates of infection, perforation rates, adherence to guidelines, usage rates.

Notes The primary outcome(s) should be prespecified. Details should be provided on how primary and secondary outcomes were measured. As shown in the examples above, outcome measures from trials of decision tool studies can be very diverse.

Item 7. Sample size: how sample size was determined and, when applicable, explanation of any interim analyses and stopping rules

Example The design was regarded as two embedded trials, and the sample size was determined for each separately. Each trial required 80% power to detect a 10% difference in adherence to guideline recommendations (e.g. between 45 and 55%) with a significance level of 5%. Adherence to the guidelines was defined by measures of process as recorded in the patients' records. Because the intraclass correlation coefficients for measures of process were estimated to be around 0.05, data were collected from 40 patients with each condition in each of 60 practices. Changes in patient outcome were assessed with summated Likert scales that could be considered as continuous variables with a normal distribution. Again, an intraclass correlation coefficient of 0.05 was assumed. Application of standard methods indicated that if data were collected from 40 patients from each of 60 practices the study would have 80% power to detect an effect size of 0.2 standard deviations with a significance of 5%.[149]

Notes Sample size calculation should be based on the most important primary outcome and must take the unit of allocation into account. It should be clear whether individual or cluster randomisation was used. The method of calculating the sample size should be reported. If cluster randomisation was used, the method for adjusting the sample size to take into account the intra-cluster correlation should be reported.

Item 8. Randomisation – sequence generation: method used to generate the random allocation sequence, including details of any restriction (e.g. blocking, stratification)

Example(s) Senior house officers were randomly allocated to the use of different diagnostic aids using the 'biased coin' technique to balance gender and career intention as far as possible in groups.[136]

Notes RCTs of decision tools often used a cluster randomised design. While it is reasonable to assume that characteristics of randomly allocated clusters are balanced between groups, the same assumption cannot be made about the characteristics of individuals within clusters.[241] Therefore, matching, stratification and other forms of constraints such as the biased coin technique (see example) are often used to reduce imbalance across intervention groups.[241] These methods should be noted if used. Sample size calculations and analysis should take into account these constraints and this should be reported.[241]

Item 9. Randomisation – allocation concealment: method used to implement the random allocation sequence, including details of any restriction (e.g. numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned

Notes The method of implementing the random allocation sequence should be reported (e.g. central telephone registration or labelled containers). State whether the sequence was concealed until the intervention assignment. State whether allocation was based on clusters or individuals.

Item 10. Randomisation – implementation: who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups

Notes The individuals responsible for enrolling participants, assigning participants to the use of decision tools or unaided doctors' diagnosis and generating the allocation sequence need to be stated.

Item 11. Blinding (masking): whether or not participants, those administering the interventions and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated

Example(s) Potentially, if one group of patients enrolled in a trial of a decision aid noticed that their doctors were consulting an impressive computer, while the other group had no such experience, this could unbalance the groups … [A remedy is] to arrange that all doctors left the patient briefly to visit another room where some would use the decision aid, or to arrange that all doctors consulted a computer in front of the patient, but that this delivered specific advice only in certain cases, and neutral information in the remainder.[103]

Notes Participating patients, health professionals, data collectors and data analysts should ideally be blind to the intervention assigned, the rationale being that bias is prevented by their ignorance of the assignment. However, blinding of the decision tool to the user and patient can be difficult, and authors should report whether it was done. They should also report whether assessment of outcome from a decision tool was blinded.

Item 12. Statistical methods: statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analyses and adjusted analysis

Notes Analysis should be conducted in accordance to the principle of 'intention to provide advice' or "intention to provide decision support', which is equivalent to the 'intention to treat' principle in therapeutic trials.

Results

Item 13. Participant flow: flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analysed for the primary outcome. Describe protocol deviations from the study as planned, together with reasons

Notes The numbers (of individual patients and clusters if relevant) randomly assigned to use the decision tool or not to use it, receiving the intended decision support, adhering to study protocol and included in the data analysis should all be reported with the aid of a flowchart. Deviations from the study protocol should be reported and explained.

Item 14. Recruitment: dates defining the periods of recruitment and follow-up

Notes The dates covering the recruitment and follow-up periods should be given; these are often not clearly stated.

Item 15. Baseline data: baseline demographic and clinical characteristics of each group

Item 16. Numbers analysed: number of participants (denominator) in each group included in each analysis and whether the analysis was by 'intention to treat.' State the results in absolute numbers when feasible (e.g. 10/20, not 50%)

Item 17. Outcomes and estimation: for each primary and secondary outcome, a summary of results for each group and the estimated effect size and its precision (e.g. 95% confidence interval)

Item 18. Ancillary analysis: address multiplicity by reporting any analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory

Notes Results of subgroup analyses and adjusted analyses should be reported, with statements on whether they were defined prospectively or post hoc.

Item 19. Adverse events: all important adverse events or side-effects in each intervention group

Example(s)

- Errors in reference database calculation of Down's syndrome screening, giving false-negatives, Sheffield.[74]
- Errors in updated embedded clinical coding software giving false plain-language representation of diagnoses, UK.[74]

Notes Unlike drugs, decision tools, including computer-based tools, are currently exempt from regulation. However, this may change.[244] It is well documented that errors sometimes occur when decision tools are used.[74] These 'adverse events' should be reported.

Discussion

Item 20. Interpretation: interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision, and the dangers associated with the multiplicity of outcomes

Notes Given the diverse ways in which discussion sections are written in decision tool studies (e.g. AAP decision tool studies as well as many other studies of medical informatics systems), the tendency sometimes to overinterpret results, the habit of some researchers to report methods and new results in the discussion section instead of the relevant sections, and the perhaps understandable lack of maturity in a young science such as health informatics,[245] there may be a case to be made to go beyond CONSORT's original recommendations for discussion sections.

Discussion sections structured with the following headings are suggested:[246]

- Statements of principal findings
- Strengths and weaknesses of the study
- Strengths and weaknesses in relation to other studies, discussing particularly any differences in results
- Meaning of the study: possible mechanisms and implications for clinicians or policy makers
- Unanswered questions and future research.

Item 21. Generalisability: generalisability (external validity) of the trial findings

Notes The generalisability or external validity of study findings should be discussed, at the individual patient level and/or cluster level as appropriate.

Item 22. Overall evidence: general interpretation of the results in the context of current evidence

General comments on the reporting of decision tool studies

There should be a better understanding of terms from different specialities in a multidisciplinary field such as medical informatics. If authors report that they have conducted a controlled trial they must be sure they have done one. In the impact study review of AAP decision tools, some authors described their studies as clinical trials or experiments in the article abstracts, when uncontrolled before-and-after studies or cross-sectional opinion surveys were conducted. These studies are often mistakenly indexed in MEDLINE or CENTRAL as 'controlled clinical trials'. As another of many examples, a study with a primary objective of measuring decision tool accuracy should not be called an experiment or a 'trial', because it is not. An experiment would be the wrong study design to answer the study question.

There is a case for a common multidisciplinary glossary of concepts to be published simultaneously in journals in medical informatics, general medicine, public health and other specialities that report decision tool studies to improve the understanding and usage of others' specialities. Two of the authors of this monograph have made an initial effort in compiling a glossary,[12] but there is a strong argument for wider dissemination of these documents and better communication between fields connected to healthcare, medical informatics and evaluation methodology.

Readers should not get the impression that this monograph is claiming that controlled trials and properly designed accuracy studies are always the most important types of evaluations for decision tools and in the field of medical informatics in general. Rather, for the types of question that were investigated for studies 1 and 2, they were the best study designs to provide the answers. Evaluation methodology in medical informatics is a rich field, with a diverse range of primary studies to evaluate the usability and cost-effectiveness of decision tools. Certain types of study question require the subjectivist or qualitative evaluation methods that are covered in depth elsewhere.[81,248,249]

Towards good practice in the conduct of decision tool evaluations

Given the diverse toolkit of evaluation methodologies in medical informatics, there is also a need to develop guidelines for the conduct (as opposed to the reporting) of accuracy studies, impact studies, usability studies, economic evaluations and subjectivist studies for decision tools in medical informatics. However, this task is beyond the scope of this monograph. One way to begin is to start with a set of methodological obstacles (e.g. in conducting RCTs in medical informatics) and come up with methods to solve these problems. These proposed solutions can then serve as a first step towards the development of conduct guidelines.

One should keep in mind that the conduct of evaluation studies, even when restricted to a specific method (e.g. accuracy studies), depends largely on the question asked.

Chapter 9
Discussion and conclusions

Summary and discussion of results

Accuracy

Thirty-two studies (from 27 papers), all based in secondary care, were eligible for the review of decision tool accuracies, while two were eligible for the review of the accuracy of hospital doctors aided by decision tools. Sensitivities and specificities for decision tools ranged from 53 to 99% and from 30 to 99%, respectively. Those for unaided doctors ranged from 64 to 93% and from 39 to 91%, respectively. Thirteen studies reported false-positive and false-negative rates for both decision tools and unaided doctors' diagnosis, enabling a direct comparison of their relative performance. In random effects meta-analyses, decision tools had higher false-negative rates than unaided doctors (error rate ratio 1.34, 95% CI 0.93 to 1.93). These results suggest that, overall, decision tools may potentially be useful in confirming a diagnosis of acute appendicitis, but not useful in ruling it out.

Two studies compared the diagnostic accuracies of doctors aided by decision tools with unaided doctors. In a multiarm cluster randomised trial ($n = 5193$), there was insufficient evidence to suggest a difference between doctors not given access to decision tools (sensitivity 28.4% and specificity 96.0%) and the three groups of aided doctors (sensitivities of 42.4–47.9% and specificities of 95.5–96.5%). In an uncontrolled before-and-after study ($n = 1484$), the sensitivities and specificities of aided and unaided doctors were 95.5% and 91.5% ($p = 0.24$) and 78.1% and 86.4% ($p < 0.001$), respectively. In both studies, there was no indication that doctors aided by decision tools were more accurate in diagnosing AAP than doctors not aided by these tools.

Thirty-two studies were found to be eligible for data extraction in the decision tool accuracy review (of which 30 presented data for sensitivities and specificities of decision tools). The metaregression of AAP decision tools showed that: (1) prospective test-set validation at the site of the tool's development showed considerably higher diagnostic accuracy than prospective test-set validation at an independent centre (RDOR 8.19, 95% CI 3.09 to 14.73); (2) the earlier the year that the study was performed the higher the performance (RDOR 0.88, 95% CI 0.83 to 0.92); and (3) developers evaluating their own tool showed better performance than independent evaluators (RDOR 2.97, 95% CI 1.31 to 6.77). Metaregression showed no evidence of an association between other quality indicators and DT accuracy.

These results should be treated with caution because the quality of studies was generally not of a high standard. The problems included:

- The gender/age breakdown of patients was often not given.
- It was often unclear whether doctors and pathologists making the final diagnosis were blinded from the results of the decision tool.
- Incorporation and verification biases were potential problems in most studies.

Impact studies

In the one eligible study of the impact study review, a four-armed cluster randomised trial ($n = 5193$) showed that unnecessary hospital admission rates of patients by doctors not allocated to a decision tool (42.8%) were significantly higher than those by doctors allocated to three combinations of structured forms, the Leeds AAP computer system, and audit and feedback (34.2–38.5%) ($p < 0.001$).[136] There was insufficient evidence of a difference between perforation rates ($p = 0.19$) and negative laparotomy rates in the four trial arms ($p = 0.46$). The limited evidence suggested that the impact of a structured paper checklist on patient outcomes is comparable to that of a computer-based decision tool, but likely to be at a much lower cost. The results indicated that much of the benefit of the Leeds AAP system (in terms of reductions in admissions rates) came from a structured data collection sheet and not from advice from the computer-based decision tool. This result needs to be treated with caution, since only one paper was eligible in the impact study review. The other abstracts and retrieved papers were rejected mostly because of inappropriate study designs (e.g. uncontrolled before-and-after studies).

Usage rates and cost-effectiveness

The usage rates of AAP decision tools were extracted from six papers that presented results and ranged from 10 to 77%. Potential determinants of usability and usage rates include the number of types of items used by the decision tool, the reasoning method used and the output format. Possible primary study designs have been outlined, because of the limitations of a review approach towards the assessment of usability. From the systematic review of impact studies, it is evident that insufficient data exist to conduct an advanced cost-effectiveness evaluation. A deterministic cost-effectiveness comparison was made, and the structured data collection sheet was found to be 100–900 times more cost-effective than a computer-based decision tool, under assumptions of constant returns to scale (the conventional assumption) or decreasing returns to scale, but uncertain under increasing returns.

Discussion and general comments

There are lessons to be learned from this review that can potentially be applied to other decision tools (including computer decision support systems), particularly in regard to the reporting of and conduct of evaluation studies. Thus, the provisional reporting guidelines in Chapter 8 are aimed at clinicians keeping up with literature and incorporating research into practice, and at decision tool and medical informatics researchers in general.

Recommendations for clinical practice

Of particular interest to doctors is whether the decision tools reviewed in this monograph improve diagnostic accuracy and have an impact on the management of AAP patients compared with routine clinical practice.

It was previously unclear which, if any, computer-based or algorithm-guided decision tool was useful for clinical practice, as stated in the HTA commissioners' original call for proposals for this project. The meta-analysis on diagnostic accuracies suggested that AAP decision tools included in the review may potentially be useful for confirming a diagnosis of acute appendicitis, but not useful for ruling it out. However, neither of the studies that compared doctors aided by an AAP decision tool with unaided doctors showed sufficient evidence of a difference in diagnostic accuracy between doctors aided by decision tools and those who were not aided. This is an indication that it is insufficient for decision tools to be accurate: doctors given access to them need to find them usable and credible if they are to accept their advice.

The impact study review indicated that the use of a structured checklist alone can help to improve impact on patient outcomes, based on one RCT.[136] With a specificity of 96%, an unaided SHO can confirm a diagnosis of acute appendicitis as confidently as SHOs aided by decision tools. However, the sensitivities of unaided SHOs and aided SHOs were all low, which means that patients cannot be reliably ruled out for the condition, an arguably more important decision for an emergency condition that requires a quick decision. It should be emphasised again that the impact study review is based on only one study, so doctors should view these findings with caution. The clinical use of structured checklists is promising as a way to improve the management of AAP patients, although more research is needed to confirm this. The basic cost-effectiveness comparison in Chapter 7 suggests that a structured data collection sheet is likely to be more cost-effective than a computer-based system such as the PDA version of the Leeds AAP system. The quality of reporting and conduct of evaluation studies of AAP decision tools is generally poor. Doctors need to read them with caution and should not incorporate the findings into their clinical practice, unless the studies have largely conformed to reporting guidelines such as those drafted in Chapter 8. Further recommendations for clinicians were limited by the available literature.

Strengths and limitations of this study

Strengths

- This monograph is the first study to systematically review the accuracy and impact of AAP decision tools and to assess their usability and cost-effectiveness.
- A sufficient number of eligible accuracy studies was identified to conduct a meta-analysis rather than a qualitative systematic review.
- Despite the encouraging results of individual primary studies of decision tool accuracy (some of which demonstrated high sensitivities and specificities), the meta-analyses of error rate ratios showed that AAP decision tools may potentially be useful in confirming a diagnosis (i.e. high specificity), but not useful in ruling it out (low sensitivity).

- Two studies compared doctors aided by an AAP decision tool with unaided doctors. There was insufficient evidence to suggest that aided doctors were more accurate in diagnosing AAP than unaided doctors.
- Although only one eligible impact study was identified, it provided some important results, in particular that the checklist effect is largely responsible for the observed impact of the Leeds AAP system.
- The impact study review identified a gap in the literature, namely the need for more controlled trials to be conducted in the area of AAP decision tools.
- Despite limited data, a basic cost-effectiveness comparison was made and found that the structured data collection sheet is much more cost-effective than a PDA version of the Leeds AAP system, under assumptions of constant returns to scale (the standard assumption). The cost-effectiveness comparison adhered to the basic principles of economics.
- The few studies that did report usage rates demonstrated a high level of heterogeneity (range 10–70%), probably reflecting a wide range in doctors' perceived usability of these tools, an important finding even though the number of studies is small.
- Despite the limitations of a review approach in assessing usability, as mentioned above, numerous primary methods of investigation exist. Outline protocols were written for two of these methods.
- Some of the metaregression findings are quite novel, for example the effect of the type of data set, identity of the evaluator and year of the study as effect modifiers of diagnostic accuracy. The surprising finding was the lack of evidence of an association between prevalence and diagnostic accuracy.
- Deeks and colleagues' recent study on bias in non-randomised studies provided empirical evidence that statistical adjustment methods often do not effectively remove confounding from the results of non-randomised studies.[250] In light of their findings, the decision to include only randomised or quasi-randomised controlled trials in the impact study review seems to have been an apt one.

Limitations

- The focus was on acute appendicitis, although this is also a strength of the review, since it is the AAP condition that requires the most urgent attention. The decision meant that the other important but rarer conditions, acute cholecystitis and perforated peptic ulcer, were ignored. However, emergency surgery is rare now for cholecystitis (see Chapter 1). Perforated peptic ulcer, a once common condition in A&E, is becoming increasingly rare (see Chapter 1).
- A low percentage of the included studies reported the performances of both decision tools and unaided doctors' diagnosis. This could potentially be a manifestation of outcome reporting bias (e.g. studies that analysed the performance of doctors did not report the results because they were more accurate than, or as good as, the decision tool).
- The review included only English-language articles. There is thus potential for language bias, although some authors published their papers in both English and their native language.
- The comparison of doctors' aided decisions using AAP decision tools with unaided doctors' decisions would be a more relevant comparison. However, the limiting factor is the lack of appropriately designed studies.
- Only one study was found to be eligible for the systematic review of impact studies, a cluster randomised trial.[136] There was therefore no scope for analysing subgroups of interest between studies.
- Few studies reported usage rates and the limitations of using a review approach to assess usability were pointed out.
- An insufficient number of data points was available to include more than one covariate at a time in the metaregression for the accuracy review. Lijmer and colleagues' meta-epidemiological study of diagnostic test evaluation studies had a sample size of 218 studies, which allowed them to include all covariates of interest in their metaregression model at the same time.[101]

Recommendations for further research

Additional research is recommended and is described below in approximate order of priority.

Acceptability and determinants of success

There is a need to ascertain the reasons why AAP decision tools have not lived up to their early promise.

Systematic reviews across a range of diseases or tools are needed to assess factors that make decision tools more acceptable to doctors and patients, and determinants of successful decision tools such as the Ottawa Ankle Rules.[243] The impact study review identified only one RCT (Wellwood study[136]) that compared paper-based decision tools with computer-based tools. The format of decision tools may have an effect on their acceptability and usage rates by doctors, which in turn may have an effect on the impact of the tools on clinical decisions, actions and patient outcomes. More such RCTs should be conducted to improve our understanding of the format(s) that doctors find acceptable in a decision tool.

Other types of primary study can also assess the properties of decision tools that would improve their perceived usability and accessibility by healthcare professionals. Two outline protocols and a set of possible study methods were given in the section 'Methods for a primary study on usability and usage rates' (p. 47). Improvements in compliance because of improved usability of decision tools and users' satisfaction with them may result in improved diagnostic accuracy and impact on patient outcomes.

The above research recommendations would contribute to the evidence base of the modified reporting guidelines of CONSORT and STARD outlined in Chapter 8.

Comparison of doctors' aided decisions to unaided decisions

Wellwood and colleagues' study[136] was also the only RCT that compared doctors' aided decisions using AAP decision tools to unaided doctors' decisions. More such RCTs are needed to assess the degree to which this study's findings on aided versus unaided decisions are replicable.

Contributions made by special investigations

The context in which AAP is now managed has changed since the 1970s, when clinical data were all that were available. Special investigations such as computed tomography, ultrasound and other imaging modalities and laboratory tests are now available. Given these new technologies, there may be a need to reassess what factors to include in AAP decision tools and on how to combine the factors meaningfully. As a first step, potential users of AAP decision tools could be recruited to participate in focus group studies, nominal group studies, opinion surveys or discrete choice experiments to identify these factors.

Barriers to economic evaluations

There is a dearth of evidence on the cost-effectiveness of AAP decision tools (or decision tools in general). The barriers and attitudes of medical informatics researchers towards economic evaluations, as mentioned in Chapter 7, need to be identified. A questionnaire survey of medical informatics researchers is one way to elicit what these barriers and attitudes are.

The role of AAP decision tools in primary care

Of all the studies included in the accuracy and impact study reviews, none included primary-care patients. Primary research (using study designs employed by investigations of AAP tools in secondary care from the accuracy review and impact study review) is needed to evaluate the role of decision tools for AAP in GP surgeries.

NSAP and other AAP conditions

There is a need to conduct a systematic review of the empirical literature on the health status of discharged patients, to estimate the prevalence of these patients who actually have appendicitis or other AAP conditions. Unnecessary admissions, or even worse, operations on patients with NSAP pose unnecessary health risks and have resource implications.

Acknowledgements

We would like to thank the following individuals for their help: Ms Sandra Bistrow, Mr A. Botha, Mr Mike Bradburn, Mr Stephen Brearly, Dr Andrew Chivers, Dr Victoria Cornelius, Mr James Crinnion, Ms Tracy Edwards, Mr Hans-Peter Eich, Professor Alastair Gray, Mr James Keating, Ms Helen King, Dr Jonathan Mant, Dr Richard McManus, Ms Rachel Oakes, Professor David Schriger, Dr Robin Snowball, Dr David Spiegelhalter, Ms Ellen Storey and Ms Sally Wellham.

We gratefully acknowledge Dr Peter Donnan (University of Dundee), Professor John Gabbay (Wessex Institute for Health and Development), Professor Brian Haynes (McMaster University), Professor Richard Lilford (University of Birmingham), Dr Ruairidh Milne (University of Southampton), Dr David Spiegelhalter (Cambridge University), Professor Frank Sullivan (University of Dundee) and Dr Johan van der Lei (Erasmus University Rotterdam) for valuable comments on earlier drafts of Appendix 1 of this report.

The Oxford University Vice Chancellors' award fund, the Chevening Scholarship fund, Oxford's Nuffield Department of Medicine, the Wolfson College Bursary Fund and the Wolfson College Senior Tutor's fund provided Joseph Liu with additional financial support to complete this project.

Contribution of authors

Jeremy Wyatt (Professor of Health Informatics) wrote the original grant proposal for the project. Joseph Liu (Senior Research Fellow in Health Informatics) designed the study, designed the data extraction forms, extracted the data, conducted the statistical analysis, interpreted the results and wrote the manuscript with contributions from Jeremy Wyatt, Jonathan Deeks (Professor of Health Statistics), Susan Clamp (Lecturer in Health Informatics), Justin Keen (Professor of Health Politics and Information Management), Pablo Verde (Statistician), Christian Ohmann (Professor of Theoretical Surgery), James Wellwood (Consultant Surgeon), Martin Dawes (Chair of Family Medicine) and Douglas G Altman (Professor of Statistics in Medicine).

References

1. De Dombal FT. *Diagnosis of acute abdominal pain*. New York: Churchill Livingstone; 1991.
2. Barron WF, Perlack RD, Boland JJ. *Fundamentals of economics for environmental managers*. Westport, CT: Quorum Books; 1998.
3. Macpherson G, editor. *Black's medical dictionary*. 39th ed. London: A&C Black; 1999.
4. Colman AM. *A dictionary of psychology*. Oxford: Oxford University Press; 2001.
5. Silen W. *Cope's early diagnosis of the acute abdomen*. 20th ed. New York: Oxford University Press; 2000.
6. Rao PM, Rhea JT, Novelline MD, Mostafavi AA, McCabe CJ. Effect of computer tomography of the appendix on treatment of patients and use of hospital resources. *N Engl J Med* 1998;**3381**:141–6.
7. Phillips KA, Maddala T, Johnson FR. Measuring preferences for health care interventions using conjoint analysis: an application to HIV testing. *Health Ser Res* 2002;**37**:1681–705.
8. Last JM. *A dictionary of epidemiology*. 3rd ed. New York: Oxford University Press; 1995.
9. Jacobs P. *The economics of health and medical care*. 4th ed. Gaithersburg, MD: Aspen; 1996
10. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000;**53**:1119–29.
11. Shortliffe EH, Perreault LE, Wiederhold G, Fagan LME. Glossary. In: *Medical informatics: computer applications in health care and biomedicine*. New York: Springer; 2001. pp. 749–820.
12. Wyatt JC, Liu JLY. Basic concepts in medical informatics. *J Epidemiol Community Health* 2002;**56**:808–12.
13. Jones J, Hunter D. Using the Delphi and nominal group techniques in health services research. In *Qualitative research in health care*. 2nd ed. London: BMJ Publishing; 2000. pp. 40–9.
14. Indar AA, Beckingham IJ. Acute cholecystitis. *BMJ* 2002;**325**:639–43.
15. Clamp SE. *Enhancement of biomedical technology usage in the diagnosis of acute abdominal pain AAP2: final guideline document*. Leeds: University of Leeds; 1996.
16. De Dombal FT, Margulies M. Acute abdominal pain. *Surgery* 1996;**120**:97–102.
17. BMJ Publishing Group. Appendicitis. *Clin Evid* 2002;**7**:387–91.
18. Benjamin IS, Patel AG. Managing acute appendicitis. *BMJ* 2002;**325**:505–6.
19. Spraycar M. *Stedman's medical dictionary*. 26th ed. Baltimore, MD: Williams and Wilkins; 1995.
20. Danesh J, Appleby P, Peto R. How often does surgery for peptic ulceration eradicate *Helicobacter pylori*? *BMJ* 1998;**316**:746–7.
21. Chan FK, To KF, Wu JC, Yung MY, Leung WK, Kwok T, *et al*. Eradication of *Helicobacter pylori* and risk of peptic ulcers in patients starting long-term treatment with non-steroidal anti-inflammatory drugs: a randomised trial. *Lancet* 2002;**ii**:204–6.
22. Metzger J, Styger S, Sieber C, von Flue M, Vogelbach P, Harder F. Prevalence of *Helicobacter pylori* infection in peptic ulcer perforations. *Swiss Med Wkly* 2001;**131**:99–103.
23. Campbell WB, Lee EJK, Van de Sijpe K, Gooding J, Cooper MJ. A 25 year study of emergency surgical admissions. *Ann R Coll Surg Engl* 2002;**84**:273–7.
24. Primatesta P, Goldacre MJ. Appendicectomy for acute appendicitis and for other conditions: an epidemiological study. *Int J Epidemiol* 1994;**23**:155–60.
25. Larner AJ. The aetiology of appendicitis. *British Journal of Hospital Medicine* 1988;**39**:540–2.
26. Irvin TT. Abdominal pain: a surgical audit of 1190 emergency admissions. *Br J Surg* 1989;**76**:1121–5.
27. Pledger G, Stringer MD. Childhood deaths from acute appendicitis in England and Wales 1963-97: observational population based study. *BMJ* 2001;**323**:430–1.
28. Margenthaler J, Schuerer D, Whinney R. Acute cholecystitis. *Clinical Evidence*, Issue 12. London: BMJ Publishing; 2004. pp. 571–80.
29. Beckingham IJ, Bornman PC. Acute pancreatitis. *BMJ* 2001;**322**:595–8.
30. Simpson J, Spiller R. Diverticular disease. *Clin Evid* 2005;**13**:336–43.
31. Williams N, Jackson D, Lambert PC, Johnstone JM. Incidence of non-specific abdominal pain in children during school term: population survey based on discharged diagnoses. *BMJ* 1999;**318**:1455.

32. Fraser S, Smith K, Agarwal M, Bates T. Psychological screening for non specific abdominal pain. *Br J Surg* 1992;**79**:1369–71.

33. De Dombal FT. Acute abdominal pain in the elderly. *J Clin Gastroenterol* 1994;19:331–5.

34. Hwang MY, Glass RM, Molter J. Detecting appendicitis in your children. *JAMA* 1999;**282**:1102.

35. Davenport M. Acute abdominal pain in children. *BMJ* 1996;**312**:498–501.

36. Rothrock SG, Pagane J. Acute appendicitis in children: emergency department diagnosis and management. *Ann Emerg Med* 2000;**36**:39–51.

37. Pena BMG, Mandl KD, Kraus SJ, Fischer AC, Fleisher GR, Lund DP, *et al*. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. *JAMA* 1999;**282**:1041–46.

38. Birnbaum BA, Wilson SR. Appendicitis at the millennium. *Radiology* 2000;**215**:337–48.

39. Eskelinen M, Ikonen J, Lipponen P. Acute appendicitis in patients over the age of 65 years; comparison of clinical and computer based decision making. *International Journal of Biomedical Computing* 1994;**36**:239–49.

40. Eskelinen M, Ikonen J, Lipponen P. The value of history-taking, physical examination, and computer assistance in the diagnosis of acute appendicitis in patients more than 50 years old. *Scand J Gastroenterol* 1995;**30**:349–55.

41. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: how to practise and teach EBM*. 2nd ed. London: Churchill-Livingstone; 2000.

42. Cartwright S, Godlee C. *Churchill's pocketbook of general practice*. Hong Kong: Churchill Livingstone; 1998.

43. Cartwright S, Godlee C. *Churchill's pocketbook of general practice*. London: Churchill Livingstone; 2003.

44. Joint Formulary Committee. *British national formulary*. 44th ed. London: British Medical Association and Royal Pharmaceutical Society of Great Britain; 2002.

45. Eriksson S. Acute appendicitis – ways to improve diagnostic accuracy. *Eur J Surg* 1996;**162**:435–42.

46. Stevens LM, Lynm C. Laparoscopy. *JAMA* 2002;**287**:402.

47. Neumayer L, Kennedy A. Imaging in appendicitis: a review with special emphasis on the treatment of women. *Obstet Gynecol* 2003;**102**:1404–9.

48. Black ER, Bordley DR, Tape TG, Panzer RJ. *Diagnostic strategies for common medical problems* 2nd ed. Philadelphia, PA: American College of Physicians; 1999.

49. Burroughs A, Feagan B, McDonald JE. *Evidence based gastroenterology and hepatology*. London: BMJ Publishing Group; 1999.

50. British Society of Gastroenterology. United Kingdom guidelines for the management of acute pancreatitis. *Gut* 1998;**42**(Suppl 2):S1–13.

51. Adams ID, Chan M, Clifford PC, Cooke WM, Dallos V, de Dombal FT, *et al*. Computer aided diagnosis of acute abdominal pain: a multicentre study. *BMJ* 1986;**293**:800–4.

52. Beasley SW. Can we improve diagnosis of acute appendicitis? *BMJ* 2000;**321**:907–8.

53. Velanovich V, Satava R. Balancing the normal appendectomy rate with perforated appendicitis rate: implications for quality assurance. *Am Surg* 1992;**58**:264–9.

54. Flum DR, Morris A, Koepsell T, Dellinger P. Has misdiagnosis of appendicitis decreased over time? A population-based analysis. *JAMA* 2001;**286**:1748–53.

55. Ohmann C, Yang Q, Franke C. Diagnostic scores for acute appendicitis. *Euro J Surg* 1995;**161**:273–81.

56. Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? *BMJ* 1995;**311**:1539–41.

57. Hollingsworth TH. Using an electronic computer in a problem of medical diagnosis. *Journal of the Royal Statistical Society* 1959;**122**:221–31.

58. Alvarado A. A practical score for the early diagnosis of acute appendicitis. *Ann Emerg Med* 1986;**15**:557–64.

59. Wallis EJ, Ramsay LE, Haq IU, Ghahramani P, Jackson PR, Rowland-Yeo K, *et al*. Coronary and cardiovascular risk estimation for primary prevention: validation of a new Sheffield table in the 1995 health survey population. *BMJ* 2000;**320**:671–6.

60. Brindle P, Emberson J, Lampe F, Walker M, Whincup P, Fahey T, *et al*. Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. *BMJ* 2003;**327**:1267–79.

61. Hense HW. Risk factor scoring for coronary heart disease: prediction algorithms need regular updating. *BMJ* 2003;**327**:1238–9.

62. Liu JLY. *The effectiveness and cost-effectiveness of cardiovascular disease prevention strategies: a brief review and knowledge gaps*. Health Economics Research Centre of the Department of Public Health, Oxford University; 2000.

63. Wong CM, Hedley AJ, Bacon-Shone J, Spiegelhalter DJ, Branicki F, Wong J, *et al*. Diagnosis of acute appendicitis: a robust system to handle variations in the quality of input data. In Hedley AJ, Wong CP, Ho LM, McGhee SM, Johnston J, Leung R, editors. *Proceedings of the Second Asia Pacific Medical Informatics Conference*. Hong Kong: Hong Kong Society of Medical Informatics and Hong Kong Computer Society; 1992. pp. 107–15.

64. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 39–60.

65. Kraemer HC. *Evaluating medical tests: objective and quantitative guidelines*. Newbury Park, CA: Sage; 1992.

66. Feinstein AR. Misguided efforts and future challenges for research on 'diagnostic tests'. *J Epidemiol Community Health* 2002;**56**:330–2.

67. Colditz G. Improving standards of medical and public health research. *J Epidemiol Community Health* 2002;**56**:333–4.

68. Brenner H, Sturmer T, Gefeller O. The need for expanding and re-focusing of statistical approaches in diagnostic research. *J Epidemiol Community Health* 2002;**56**:338–9.

69. Choi BCK. Future challenges for diagnostic research: striking a balance between simplicity and complexity. *J Epidemiol Community Health* 2002;**56**:334–5.

70. Knottnerus JA. Challenges in dia-prognostic research. *J Epidemiol Community Health* 2002;**56**:340–1.

71. Brahams D, Wyatt JC. Decision aids and the law. *Lancet* 1989;**ii**:632–4.

72. Hope T, Savulescu J, Hendrick J. *Medical ethics and law: the core curriculum*. Edinburgh: Churchill Livingstone; 2003.

73. Hart A, Wyatt JC. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. *Med Inform* 1990;**15**:229–36.

74. Rigby M, Forsstrom J, Roberts R, Wyatt J. Verifying quality and safety in health informatics services. *BMJ* 2001;**323**:552–6.

75. Bayes T. An essay towards solving a problem in the doctrine of chances. *Philos Trans R Soc* 1763;**53**:370–418.

76. Spiegelhalter DJ, Knill-Jones R. Statistical and knowledge-based approaches to clinical decision support systems, with an application to gastroenterology. *Journal of the Royal Statistical Society* A 1984;**147**:35–77.

77. Carter JH. Design and implementation issues. In *Clinical decision support systems: theory and practice*. New York: Springer; 1999. pp. 169–97.

78. Deeks JJ. Evaluations of diagnostic and screening tests. In Egger M, Smith GD, Altman DG, editors. *Systematic reviews in health care: meta analysis in context*. 2nd ed. London: BMJ Publishing Group; 2001. pp. 248–84.

79. Deville W. *Evidence in diagnostic research*. The Netherlands: Ponsen and Looijen; 2001.

80. Deville W, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol* 2000;**53**:65–9.

81. Friedman CP, Wyatt JC. *Evaluation methods in biomedical informatics*, 2nd ed. New York: Springer; 2006.

82. McManus RJ, Mant J, Davies MK, Davis RC, Deeks JJ, Oakes RA, *et al*. A systematic review of the evidence for rapid access to chest pain clinics. *Int J Clin Pract* 2002;**56**:29–33.

83. Altman DG. *Practical statistics for medical research*. London: Chapman and Hall; 1991.

84. Cochrane Methods Working Group on Systematic Reviews and Diagnostic Tests. *Recommended methods*. URL: http://som.flinders.edu.au/cochrane/; 1996.

85. Wyatt JC. Decision support systems. *J R Soc Med* 2000;**93**:629–33.

86. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al*. The STARD statement for reporting studies for diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;**49**:7–18.

87. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Chem* 2003;**49**:1–6.

88. Jaeschke R, Guyatt G, Lijmer J. Diagnostic tests. In *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago, IL: AMA Press; 2002. pp. 121–40.

89. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials. Baltimore, MD: Williams and Wilkins; 1996.

90. Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, *et al*. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;**120**:667–76.

91. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993;**13**:313–21.

92. Deeks JJ, Altman DG, Bradburn M. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In Egger M, Smith GD, Altman DG, editors. London: BMJ Publishing Group; 2001. pp. 285–312.

93. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177–88.

94. Egger M, Smith GD, Altman DG, editors. *Systematic reviews in health care: meta-analysis in context*. London: BMJ Publishing Group; 2001.

95. Glasziou P, Irwig L, Bain C, Colditz G. *Systematic reviews in health care: a practical guide*. Cambridge: Cambridge University Press; 2001.

96. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol* 1999;**150**:469–75.

97. Montori V, Guyatt G, Oxman A, Cook D. Summarizing the evidence: fixed effects and random-effects model. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA Press; 2002. pp. 539–45.

98. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. 2nd ed. Boston, MA: Little, Brown; 1991.

99. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.

100. Hellmich M, Abrams KR, Sutton AJ. Bayesian approaches to meta-analysis of ROC curves. *Med Dec Making* 1999;**19**:252–64.

101. Lijmer JG, Mol BM, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen J, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.

102. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–73.

103. Wyatt JC. *A method for developing medical decision-aids applied to ACORN, a chest pain advisor* Doctorate thesis. Oxford: Oxford University; 1991.

104. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: applications and methodological standards. *N Engl J Med* 1985;**313**:793–9.

105. Schmitz PIM, Habbema JDF, Hermans J. The performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods. *Stat Med* 1983;**2**:199–205.

106. Greenland S. Meta-analysis. In Rothman KJ, Greenland S, editors. *Modern epidemiology*. Philadelphia, PA: Lippincott-Raven; 1998. pp. 643–74.

107. Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2001;**2**:463–71.

108. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clin Trials* 1990;**11**:339–52.

109. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;**282**:1054–60.

110. Light R, Pillemer DB. *Summing up: the science of reviewing research*. Cambridge, MA: Harvard University Press; 1984.

111. Egger M, Davey-Smith G, Schneider M, Minder C. Bias in meta analysis detected by a single graphical test. *BMJ* 1997;**315**:629–34.

112. Bond GR, Tully SB, Chan LS, Bradley RL. Use of the MANTRELS score in childhood appendicitis: a prospective study of 187 children with abdominal pain. *Ann Emerg Med* 1990;**19**:1014–18.

113. De Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer aided diagnosis of acute abdominal pain. *BMJ* 1972;**ii**:9–13.

114. Edwards NH. The accuracy of a bayesian computer program for diagnosis and teaching in acute abdominal pain of childhood. *Comput Methods Programs Biomed* 1986;**23**:155–60.

115. Eskelinen M, Ikonen J, Lipponen P. A computer based diagnostic score to aid in diagnosis of acute appendicitis: a prospective study of 1333 patients with acute abdominal pain. *Theoretical Surgery* 1992;**7**:86–90.

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

117. Hallan S, Asberg A, Harald Edna T. Additional value of biochemical tests in suspected acute appendicitis. *Eur J Surg* 1997;**163**:533–8.

118. Hallan S, Asberg A, Harald ET. Estimating the probability of acute appendicitis using clinical criteria of a structured record sheet: the physician against the computer. *Eur J Surg* 1997;**163**:427–32.

119. Horrocks JC, Devroede G, De Dombal FT. Computer aided diagnosis of gastroenterologic diseases in Sherbrooke: preliminary report. *Can J Surg* 1976;**19**:160–4.

120. Izbicki J, Knoefel W, Wilker D, Mandelkow H, Muller K, Siebeck M, *et al*. Accurate diagnosis of acute appendicitis: a retrospective and prospective analysis of 686 patients. *Eur J Surg* 1992;**158**:227–31.

121. Jahn H, Mathiesen F, Neckelmann K, Hovendal C, Bellstrom T, Gottrup F. Comparison of clinical judgement and diagnostic ultrasonography in the diagnosis of acute appendicitis: experience with a score aided diagnosis. *Eur J Surg* 1997;**163**:433–43.

122. Jawaid A, Asad A, Motiei A, Munir A, Bhutto E, Choudry H, *et al*. Clinical scoring system: A valuable tool for decision making cases of acute appendicitis. *J Pak Med Assoc* 1999;**49**:254–9.

123. Kirkeby OJ, Riso C. Use of a computer system for diagnosing acute abdominal pain in a small hospital. *Scand J Gastroenterol* 1987;**22**(Suppl 128):174–6.

124. Kraemer M, Yang Q, Ohmann C. Classifications of subpopulations with a minor and a major diagnostic problem in acute abdominal pain. *Theoretical Surgery* 1993;**8**:6–14.

125. Leaper DJ, Horrocks JC, Staniland JR, De Dombal FT. Computer assisted diagnosis of abdominal pain using 'estimates' provided by clinicians. *BMJ* 1972;**iv**:350–4.

126. Lindberg G, Fenyo G. Algorithmic diagnosis of appendicitis using Bayes' theorem and logistic regression. In Bernado JM, De Groot MH, Lindley DV, Smith FM, editors. *Bayesian statistics* 3. Oxford: Oxford University Press; 1988. pp. 665–8.

127. Macklin CP, Merei JM, Radcliffe GS, Stringer MD. A prospective evaluation of the modified Alvarado score for acute appendicitis in children. *Ann R Coll Surg Engl* 1997;**79**:203–5.

128. Malik A, Wani N. Continuing diagnostic challenge of acute appendicitis: evaluation through modified Alvarado score. *ANZ J Surg* 1998;**68**:504–5.

129. Ohmann C, Franke C, Yang Q. Clinical benefit of a diagnostic score for appendicitis: results of a prospective interventional study. *Arch Surg* 1999;**134**:993–6.

130. Owen TD, William H, Stiff G, Jenkinson LR, Rees BI. Evaluation of the Alvarado score in acute appendicitis. *J R Soc Med* 1992;**85**:87–8.

131. Pesonen E, Eskelinen M, Juhola M. Comparison of different neural network algorithms in the diagnosis of acute appendicitis. *International Journal of Biomedical Computing* 1996;**40**:227–33.

132. Pesonen E. Is neural network better than statistical methods in diagnosis of acute appendicitis? *Med Inform Eur* 1997;**43**:377–81.

133. Saidi R, Ghasemi M. Role of Alvarado score in diagnosis and treatment of suspected acute appendicitis. *Am J Emerg Med* 2000;**18**:230–31.

134. Staniland JR, Clamp SE, De Dombal FT, Solheim K, Hansen S, Ronsen K, *et al*. Presentation and diagnosis of patients with acute abdominal pain: comparisons between Leeds, UK, and Akershus county, Norway. *Ann Chir Gynaecol* 1980;**69**:245–50.

135. Sutton GC. How accurate is computer-aided diagnosis? *Lancet* 1989;**ii**:905–8.

136. Wellwood J, Johannesen S, Spiegelhalter D. How does computer aided diagnosis improve the management of acute abdominal pain? *Ann R Coll Surg Engl* 1992;**74**:40–6.

137. Wong CM, Hedley AJ, Bacon-Shone J, Spiegelhalter DJ, Ma S, Branicki F, *et al*. The determination of the cut-off point in the receiver operating characteristic curve for the management of acute abdominal pain. In McGhee SM, Hedley AJ, Wong CP, Ho LM, editors. *Proceedings of the Third Asia Pacific Medical Informatics Conference*. Hong Kong: Hong Kong Society of Medical Informatics and Hong Kong Computer Society; 1994. pp. 159–61.

138. Horrocks JC, McCann AP, Leaper DJ, De Dombal FT. Computer aided diagnosis: description of an adaptable system and operational experience with 2,034 cases. *BMJ* 1972;**ii**:5–9.

139. Wellwood J, Spiegelhalter DJ. Computers and diagnosis of acute abdominal pain. *British Journal of Hospital Medicine* 1989;**41**:564–7.

140. Aleksander I, Morton H. *An introduction to neural computing*. 2nd ed. London: International Thomson Computer Press; 1995.

141. Ohmann C, Boy O, Yang Q. A systematic approach to the assessment of user satisfaction with health care systems: constructs, models and instruments. *Medical Informatics Europe*. IOS Press; 1997. pp. 781–5.

142. Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago, IL: AMA Press; 2002.

143. Chan AW. *Outcome reporting bias in randomised trials: implications for systematic reviews*. DPhil thesis. Oxford: Oxford University; 2003.

144. Petitti DB. *Meta-analysis, decision analysis and cost-effectiveness analysis: methods for quantitative synthesis in medicine*. 2nd ed. New York: Oxford University Press; 2000. p. 228.

145. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol* 2002;**31**:88–95.

146. Lijmer JG, Bossuyt PM. Diagnostic testing and prognosis: the randomised controlled trial in diagnostic research. In *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 61–80.

147. Randolph A, Haynes B, Wyatt J, Cook D, Guyatt G. Computer decision support systems. In Guyatt G, Rennie D, editors. *Users' guides to the medical*

literature: a manual for evidence-based clinical practice. Chicago, IL: American Medical Association Press 2002; pp. 291–308.

148. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *JAMA* 1998;**280**:1339–46.

149. Eccles M, McColl E, Steen N, Rousseau N, Grimshaw J, Parkin D, *et al*. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. *BMJ* 2002;**325**:941–4.

150. Moher D, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *JAMA* 2001;**285**:1987–91.

151. Wyatt JC, Paterson-Brown S, Fisk NM, Johanson R, Altman DG, Bradburn M. Randomised trials useful to find best methods of enhancing clinical practice. *BMJ* 1999;**318**:1353.

152. Wyatt JC, Paterson-Brown S, Johanson R, Altman DG, Bradburn M, Fisk N. Trials of outreach visits to enhance the use of systematic reviews in 25 obstetric units. *BMJ* 1998;**317**:1041–6.

153. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of clinical trials*. New York: Springer; 1998.

154. Puffer S, Torgerson DJ, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ* 2003;**327**:785–9.

155. Cosby RH, Howard M, Kaczorowski J, Willan AR, Sellors JW. Randomizing patients by family practice: sample size estimation, intracluster correlation and data analysis. *Fam Pract* 2003;**20**:77–82.

156. Kerry SM, Bland M. The intracluster correlation coefficient in cluster randomisation. *BMJ* 1998;**316**:1455–60.

157. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Stat Med* 2002;**21**:3291–315.

158. Altman DG, Bland JM. Units of analysis. *BMJ* 1997;**314**:1874.

159. Kerry SM, Bland JM. Analysis of a trial randomised in clusters. *BMJ* 1998;**316**:54.

160. Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* 2003;**157**:364–75.

161. Friedman CP, Wyatt JC. The design of demonstration studies. In *Evaluation methods in biomedical infomatics*, 2nd ed. New York: Springer; 2006. pp. 188–223.

162. Wyatt JC, Spiegelhalter DJ. Field trials of medical decision-aids: potential problems and solutions. In Clayton P, editor. *Proceeding of the Annual Symposium on Computer Applications in Medical Care*, Washington DC, 17–20 November. New York: McGraw Hill; 1991. pp. 3–7.

163. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *BMJ* 2001;**322**:355–7.

164. De Dombal FT. Objective medical decision making: acute abdominal pain. In: Beneken JW, Thevenin V, editors. *Advances in biomedical engineering*. Amsterdam: IOS Press; 1993.

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

166. Fenyo G, Clamp SE, De Dombal FT, Engstrom L, Hedlund M, Leijonmark CE, *et al*. Computer-aided diagnosis of 233 acute abdominal cases at Nacka Hospital Sweden. *Scand J Gastroenterol* 1987;**22**(Suppl 128):178.

167. Lawrence P, Clifford P, Taylor I. Acute abdominal pain: computer aided diagnosis by non medically qualified staff. *Ann R Coll Surg Engl* 1987;**69**:233–4.

168. De Dombal FT, Leaper DJ, Horrocks JC, Staniland JR, McCann AP. Human and computer aided diagnosis of abdominal pain: further report with emphasis on performance on clinicians. *BMJ* 1974;**i**:376–80.

169. Gruer R, Gunn AA. Medical audit in practice. *BMJ* 1977;**i**:957–8.

170. Scarlett PY, Cooke WM, Clarke D, Bates C, Chan M. Computer aided diagnosis of acute abdominal pain at Middlesbrough General Hospital. *Ann R Coll Engl* 1986;**68**:179–81.

171. Wilson P, Horrocks J, Lyndon P, Yeung C, Page R, De Dombal F. Simplified computer aided diagnosis of acute abdominal pain. *BMJ* 1975;**ii**:73–5.

172. Wilson DH, Wilson PD, Walmsley RG, Horrocks JC, De Dombal FT. Diagnosis of acute abdominal pain in the accident and emergency department. *Br J Surg* 1977;**64**:250–4.

173. Fenyo G, Lindberg G, Blind P, Enochsson L, Oberg A. Diagnostic decision support in suspected acute appendicitis: validation of a simplified scoring system. *Euro J Surg* 1997;**163**:831–8.

174. Gough IR. A study of diagnostic accuracy in suspected acute appendicitis. *ANZ J Surg* 1988;**58**:555–9.

175. Heathfield HA, Pitty D, Hanka R. Evaluating information technology in health care: barriers and challenges. *BMJ* 1998;**316**:1959–61.

176. Moehr JR. Evaluation: salvation or nemesis of medical informatics? *Comput Biol Med* 2002;**32**:113–25.

177. Kaplan B. Evaluating informatics applications – some alternative approaches, and call for methodological pluralism. *Int J Med Inf* 2001;**64**:39–56.

178. Copi IM, Cohen C. *Introduction to logic*. 11th ed. Englewood Cliffs, NJ: Prentice-Hall; 2001.

179. Grimes DA, Schultz KF. Descriptive studies: what they can and cannot do. *Lancet* 2002;**359**:145–9.

180. MacMahon S, Collins R. Reliable assessment of the effects of the effects of treatment on mortality and major morbidity, II: Observational studies. *Lancet* 2001;**357**:455–62.

181. Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: Clinical trials. *Lancet* 2001;**357**:373–80.

182. Collins R, Peto R, Gray R, Parish S. Large-scale randomized evidence: trials and overviews. In *Oxford textbook of medicine*. Oxford: Oxford University Press; 1998. pp. 21–32.

183. UK Institute of Health Informatics. NHS Information Authority – ERDIP. *Review and reference source of methodologies for evaluation of electronic health and patient record projects*. Winchester: UK Institute of Health Informatics; 2001.

184. Rigby M. Health informatics as a tool to improve quality in non-acute care – new opportunities and a matching need for a new evaluation paradigm. *Int J Med Inf* 1999;**56**:141–50.

185. Patel VL, Kaufmann DR, Arocha JF. Emerging paradigms of cognition in medical decision making. *J Biomed Inform* 2002;**35**:52–75.

186. Macintyre S, Petticrew M. Good intentions and received wisdom are not enough. *J Epidemiol Community Health* 2000;**54**:802–3.

187. Wynder EL. Studies in mechanism and prevention: striking a proper balance. *Am J Epidemiol* 1994;**139**:547–9.

188. Mill JS. *System of logic, ratiocinative and inductive*. 5th ed. London: Parker, Son and Bowin; 1862.

189. Wyatt JC. *The effects of manual paper reminders on professional practice and health care outcomes: a systematic review of 22 randomised controlled trials including 41705 patients*. London: Department of Health; 2002.

190. Balas EA, Boren SA. Clinical trials of information interventions. In: Berner ES, editor. *Clinical decision support systems: theory and practice*. New York: Springer; 1999. pp. 199–216.

191. Anderson KM, Wilson PWF, Odell PM, Kannel WB. An updated coronary risk profile: a statement for heart professionals. *Circulation* 1991;**83**:356–62.

192. Hingorani AD, Vallance P. A simple computer programme for guiding management of cardiovascular risk factors and prescribing. *BMJ* 1999;**318**:101–5.

193. Wyatt J. Same information, different decisions: format counts. *BMJ* 1999;**318**:1501–2.

194. Wyatt JC, Wright P. Design should help use of patients' data. *Lancet* 1998;**352**:1375–8.

195. Harvey AC, Moodie PF, Swirsky N, Kirkpatrick JR. An on-line Bayesian program for acute abdominal pain. In: Clayton P, editor. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*. Washington DC, 14–17 October. New York: McGraw Hill; 1983. pp. 863–7.

196. Harvey AC, Moodie PF, Reda JE, Childs NJ, Gray D, Samimi M, *et al*. The acute abdomen in emergency with Hypercard. In: Clayton P, editor. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*. Washington DC, 17–20 November. New York: McGraw Hill; 1991. pp. 156–9.

197. Heathfield HA, Wyatt JC. Philosophies for the design and development of clinical decision support systems. *Methods Inf Med* 1993;**32**:1–8.

198. Pesonen E, Ohmann C, Eskelinen M, Juhola M. Diagnosis of acute appendicitis in two databases. Evaluation of different neighbourhoods with LVQ neural network. *Methods Inf Med* 1998;**37**:59–63.

199. Black N, Murphy M, Lamping D, McKee M, Sanderson C, Askham J, *et al*. Consensus development methods, and their use in creating clinical guidelines. In: Stevens A, Abrams K, Brazier J, Fitzpatrick R, Lilford R, editors. *The advanced handbook of methods in evidence based healthcare*. London: Sage; 2001. pp. 426–48.

200. Kitzinger J. Focus groups with users and providers of health care. In *Qualitative research in health care*. 2nd ed. London: BMJ Publishing; 2000.

201. Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. *BMJ* 2000;**320**:1530–33.

202. Ryan M, McIntosh E, Shackley P. Methodological issues in the application of conjoint analysis in health care. *Health Econ* 1998;**7**:373–8.

203. McFadden D. Conditional logit analysis of qualitative choice behaviour. In: Zarembka P, editor. *Frontiers in econometrics*. New York: Academic Press; 1975. pp. 105–35.

204. Monkman D. Treating dyslipidaemia in primary care: the gap between policy and reality is large in the UK. *BMJ* 2000;**320**:1299–300.

205. British Cardiac Society, British Hyperlipidaemia Association, British Hypertension Society, British Diabetic Association. Joint British recommendations on prevention of coronary heart

disease in clinical practice: summary. *BMJ* 2000;**320**:705–8.

206. Jackson R. Updated New Zealand cardiovascular disease risk–benefit prediction guide. *BMJ* 2000;**320**:709–10.

207. Wyatt JC. Lessons learned from ACORN, an expert system to advise on chest pain. In Barnes BA, Cao D, Qin De, editors. *Proceedings of the Sixth World Conference on Medical Informatics*, Singapore. Amsterdam: North Holland; 1989. pp. 111–15.

208. Jones J, Hunter D. *Qualitative research in health care.* 2nd ed. London: BMJ Publishing; 2000.

209. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CFB, Askham J, *et al*. Consensus development methods, and their use in clinical guideline development. *Health Technol Assess* 1998;**2**(3).

210. Jones J, Hunter D. Qualitative research: consensus methods for medical and health services research. *BMJ* 1995;**311**:376–80.

211. Dolan RJ. *Conjoint analysis: a manager's guide*. Boston, MA: Harvard Business School Publishing; 1990.

212. Ryan M. *Using consumer preferences in health care decision making: the application of conjoint analysis*. London: Office of Health Economics; 1997.

213. Orme B. *Sample size issues for conjoint analysis studies*. Sequin, WA: Sawtooth Software; 1988.

214. Zwerina K, Huber J, Kuhfeld W. *A general method for constructing efficient choice designs*. Durham, NC: Fuqua School of Business, Duke University; 1996.

215. Ryan M. A role for conjoint analysis in technology assessment in health care? *Int J Technol Assess in Health Care* 1999;**15**:443–57.

216. Sheridan WG, Havard T, White AT, Crosby DL. Non-specific abdominal pain: the resource implications. *Ann R Coll Surg Engl* 1992;**74**:181–5.

217. Heath W. Saving money – saving lives. *Health Service Journal* 1986;**96**:1458.

218. Gorannsson J, Lasson A. Decision making in acute abdominal pain – accuracy, costs and availability of various diagnostic methods. *Theoretical Surgery* 1993;**8**:44–52.

219. De Dombal FT. Are improvements in the quality of care incompatible with savings in resources? *Journal of Management in Medicine* 1987;**2**:52–60.

220. Gill BD, Jenkins JR. Cost effective evaluation and management of the acute abdomen. *Surg Clin North Am* 1996;**76**:71–82.

221. Flum DR, Koepsell T. The clinical and economic correlates of misdiagnosed appendicitis: nationwide analysis. *Arch Surg* 2002;**137**:799–804.

222. Drummond MF. *Principles of economic appraisal in health care*. Oxford: Oxford University Press; 1980.

223. Drummond MF, OBrien B, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes*. 2nd ed. Oxford: Oxford University Press; 1997.

224. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost effectiveness in health and medicine*. New York: Oxford University Press; 1996.

225. Gramlich EM. A guide to benefit–cost analysis 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1997.

226. Arrow KJ. *Social choice and individual values*. 2nd ed. New York: Wiley; 1963.

227. Dasgupta AK, Pearce DW. Social welfare functions. In *Cost–benefit analysis: theory and practice*. London: Macmillan; 1978. pp. 70–96.

228. Sen AK. *Choice, welfare and measurement*. Cambridge, MA: Harvard University Press; 1997.

229. Hausman DM, McPherson MS. Social choice theory. In *Economic analysis and moral philosophy*. Cambridge: Cambridge University Press; 1996. pp. 166–79.

230. Cookson R. Willingness to pay methods in health care: a sceptical view. *Health Econ* 2003;**12**:891–4.

231. Hoch JS, Briggs AH, Willan AR. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Econ* 2002;**11**:415–30.

232. Lipsey RG, Courant PN. *Economics*. 11th ed. New York: Harper Collins; 1996.

233. Elbasha EH, Messonnier ML. Cost-effectiveness analysis and health care resource allocation: decision rules under variable returns to scale. *Health Econ* 2004;**13**:21–35.

234. Briggs AH, Gray AM. Handling uncertainty in economic evaluations of healthcare interventions. *BMJ* 1999;**319**:635–8.

235. Chapman RH, Stone PW, Sandberg EA, Bell CM, Neumann PJ. A comprehensive league table of cost utility ratios and a sub-table of panel worthy studies. *Med Decis Making* 2000;**20**:451–67.

236. Weinstein MC. From cost-effectiveness ratios to resource allocation: where to draw the line? In: Sloan F, editor. *Valuing health care: costs, benefits, and effectiveness of pharmaceuticals and other medical technologies*. Cambridge: Cambridge University Press; 1995. pp. 77–98.

237. Dexter PR, Perkins S, Overhage JM, Maharry K, Kohler RB, McDonald CJ. A computerized reminder system to increase the use of preventive care for hospitalized patients. *N Engl J Med* 2001;**345**:965–70.

238. Whitten PS, Mair FS, Haycox A, May CR, Williams TL, Hellmich S. Systematic review of cost effectiveness studies of telemedicine interventions. *BMJ* 2002;**324**:1434–7.

239. Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the panel on cost effectiveness in health and medicine. *JAMA* 1996;**276**:1253–8.

240. Moher D, Altman DG, Schultz KF, Elbourne DR. Opportunities and challenges for improving the quality of reporting clinical research: CONSORT and beyond. *Can Med Assoc J* 2004;**171**:349–50.

241. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;**328**:702–8.

242. Ioannidis JPA, Evans SJW, Gotzsche PC, O'Neill RT, Altman DG, Schultz K, *et al*. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 2004;**141**:781–8.

243. Bachmann LM, Esther K, Koller MT, Steurer J, ter Riet G. Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid foot: systematic review. *BMJ* 2003;**326**:417–9.

244. Donawa M. FDA final guidance on software validation. *Medical Device Technology* 2002;**13**:20–4.

245. Friedman CP, Abbas UL. Is medical informatics a mature science? A review of measurement practice in outcome studies of clinical systems. *Int J Med Inf* 2003;**69**:261–72.

246. Docherty M, Smith R. The case of structuring the discussion of scientific papers. Much the same as that for structuring abstracts. *BMJ* 1999;**318**:1224–5.

247. Liu JLY, Wyatt JC, Altman DG. Decision tools: focus on the problem, not the solution. *BMC Medical Informatics and Decision Making* 2006;**6**(4).

248. Pope C, Mays N. *Qualitative research in health care*. 2nd ed. London: BMJ Publishing; 2000.

249. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment. *Health Technol Assess* 1998;**2**(16).

250. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;**7**(27).

251. Kennedy I. *The Report of the Public Inquiry into children's heart surgery at the Bristol Royal Infirmary 1984–1995: learning from Bristol*. Bristol: Bristol Royal Infirmary Inquiry; 2001.

252. Randolph AD, Haynes RB, Wyatt JC, Cook DJ, Guyatt GH. Users' guides to the medical literature: XVIII. How to use an article evaluating the clinical impact of a computer-based clinical decision support system. *JAMA* 1999;**282**:67–74.

253. Wyatt JC. Clinical data systems, Part 3: Development and evaluation. *Lancet* 1994;**344**:1682–8.

254. Eccles M, Steen N, Grimshaw J, Thomas L, McNamee SJ, Soutter J, *et al*. Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial. *Lancet* 2001;**357**:1406–9.

255. O'Connor AM, Rostom A, Fiset V, Tetroe J, Entwistle V, Llewellyn-Thomas H, *et al*. Decision aids for patients facing health treatment or screening decisions: systematic review. *BMJ* 1999;**319**:731–4.

256. Leteurtre S, Martinot A, Duhamel A, Proulx F, Grandbastien B, Cotting J, *et al*. Validation of the paediatric logistic organ dysfunction (PELOD) score: prospective, observational, multicentre study. *Lancet* 2003;**362**:192–7.

257. Murray LS, Teasdale GM, Murray GD, Jennett B, Miller JD, Pickard JD, *et al*. Does prediction of outcome alter patient management? *Lancet* 1993;**341**:1487–91.

258. Gremy F. Information systems evaluation and subjectivity. *Int J Med Inf* 1999;**56**:13–23.

259. Bryce FP, Neville RG, Crombie IK, Clark RA, McKenzie P. Controlled trial of an audit facilitator in diagnosis and treatment of childhood asthma in general practice. *BMJ* 1995;**310**:838–42.

260. Holtzman J, Bjerke T, Kane R. The effects of clinical pathways for renal transplantation on patient outcomes and length of stay. *Med Care* 1998;**36**:826–34.

261. Lilford RJ, Kelly M, Baines A, Cameron S, Cave M, Guthrie K, *et al*. Effect of using protocols on medical care; randomised controlled trial of three methods of antenatal history taking. *BMJ* 1992;**305**:1181–4.

262. Wallace P, Haines A, Harrison R, Barber J, Thompson S, Jacklin P, *et al*. Joint teleconsultations (virtual outreach) versus standard outpatient appointments for patients referred by their general practitioner for a specialist opinion: a randomised trial. *Lancet* 2002;**359**:1961–8.

263. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. *JAMA* 1995;**274**:700–5.

264. Hoffer EP, Barnett GO. Computer-aided instruction in medicine: 16 years of MGH experience. In: Salamon R, Blum B, Jorgensen M, editors. *MEDINFO 1986*. Amsterdam: Elsevier North-Holland; 1986.

265. Ledley R, Lusted L. Reasoning foundations of medical diagnosis. *Science* 1959;**130**:9–21.

266. Habbema JDF, Eijkemans R, Krijnen P, Knottnerus JA. Analysis of data on the accuracy of diagnostic tests. In: Knottnerus JA, editor. *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 117–43.

267. Warner HR, Toronto AF, Veasy LG, Stephenson RS. A mathematical approach to medical diagnosis: application to congenital heart disease. *JAMA* 1961;**177**:177–83.

268. Boyle JA, Grieg WR, Franklin DA, Harden R, Buchanan WW, McGirr EM. Construction of a model for computer assisted diagnosis: an application to the problem of non toxic goitre. *QJM* 1966;**35**:565–88.

269. Stern RB, Maxwell JD, Knill-Jones RP, Thompson RP, Williams R. Use of computer-assisted model in diagnosis of drug sensitivity jaundice. *BMJ* 1973;**i**:767–9.

270. Albert A. On the use and computation of likelihood ratios in clinical chemistry. *Clin Chem* 1982;**28**:1113–19.

271. Aho AV, Ullmann JD, Hopcroft JE. *Data structures and algorithms*. London: Addison–Wesley; 1983.

272. Marshall RJ. The use of classification and regression trees in clinical epidemiology. *J Clin Epidemiol* 2001;**54**:603–9.

273. Musen MA, Shahar Y, Shortliffe EH. Clinical decision support systems. *Medical informatics: computer applications in health care and biomedicine*. 2nd ed. New York: Springer; 2001. pp. 573–609.

274. Shortliffe EH. Computer based consultations in clinical therapeutics: the MYCIN system. *Computers in Biomedical Research* 1975;**8**:303–20.

275. Yu VL, Buchanan BG, Shortcliffe EH, Wraith SM, Davis R, Scott AC, *et al*. Evaluating the perfomance of a computer based consultant. *Comput Programs Biomed* 1979;**9**:95–102.

276. Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* 2000;**19**:541–61.

277. Staniland JR, Ditchburn J, De Dombal FT. Clinical presentation of acute abdomen: study of 600 patients. *BMJ* 1972;**3**:393–8.

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

279. Wong CM, Hedley AJ, Bacon-Shone J, Branicki F, Wong J, Leaper DJ, *et al*. Development of clinical decision support for the management of acute abdominal pain. In Hedley AJ, Wong CP, editors. *Proceedings of the First Asia Pacific Medical Informatics Conference*. Hong Kong: Hong Kong Society of Medical Informatics and Hong Kong Computer Society; 1990. pp. 170–83.

280. Arnbjörnsson E. Scoring system for computer-aided diagnosis of acute appendicitis. *Ann Chir Gynaecol* 1985;**74**:159–66.

281. Anatol T, Holder Y. A multivariate analysis of childhood abdominal pain in Trinidad. *J R Coll Surg Edinb* 1995;**40**:99–103.

282. Bjerregaard B, Brynitz S, Holst Christensen J, Jess P, Kalaja E, Lund Kristensen J, *et al*. The reliability of medical history and physical examination in patients with acute abdominal pain. *Methods Inf Med* 1983;**22**:15–18.

283. Blazadonakis M, Moustakis V, Charisis G. Deep assessment of machine learning techniques using patient treatment in acute abdominal pain in children. *Artif Intell Med* 1996;**8**:527–42.

284. Bohner H, Yang Q, Franke C, Verreet PR, Ohmann C. Simple data from history and physical examination help to exclude bowel obstruction and to avoid radiographic studies in patients with acute abdominal pain. *Eur J Surg* 1998;**164**:777–84.

285. Browder W, Smith J, Vivoda L, Nicols R. Nonperforative appendicitis: a continuing surgical dilemma. *J Infect Dis* 1989;**159**:1088–94.

286. Crossley R. Hospital admissions for abdominal pain in childhood. *J R Soc Med* 1982;**75**:772–6.

287. Davenport PM, Morgan AG, Darnborough A, De Dombal FT. Can preliminary screening of dyspeptic patients allow more effective use of investigational techniques? *BMJ* 1985;**290**:217–20.

288. De Dombal FT, Horrocks JC, Staniland JR, Guillou P. Construction and uses of a data base of clinical information concerning 600 patients with acute abdominal pain. *Proc R Soc Med* 1971;**64**:978.

289. De Dombal FT, Horrocks JC. Use of receiver operating characteristic (ROC) curves to evaluate computer confidence threshold and clinical performance in the diagnosis of appendicitis. *Methods Inf Med* 1978;**17**:157–61.

290. De Dombal FT. Diagnosis of acute abdominal pain. *Tidsskr Nor Laegeforen* 1980;**100**:1852–5.

291. De Dombal FT. Computer aided diagnosis of acute abdominal pain: the British experience. *Rev Epidemiol Sante* 1984;**32**:50–6.

292. De Dombal FT, Barnes S, Dallos V, Kumar PS, Sloan J, Chan M, *et al*. How should computer aided decision support systems present their predictions to the practising surgeon? *Theoretical Surgery* 1992;**7**:111–16.

293. Dickson JAS, Edwards N, Jones A. A computer assisted diagnosis of acute abdominal pain in childhood. *Lancet* 1985;**i**:1389–90.

294. Edwards FH, Davis RS. Use of a Bayesian algorithm in the computer assisted diagnosis of appendicitis. *Surgery, Gynecology and Obstetrics* 1984;**158**:219–22.

295. Eskelinen M, Ikonen J, Lipponen P. Contributions of history taking, physical examination and computer assistance to diagnosis of acute small bowel obstruction. A prospective study of 1333 patients with acute abdominal pain. *Scand J Gastroenterol* 1994;**29**:715–21.

296. Eskelinen M, Ikonen J, Lipponen P. Sex specific diagnostic scores for acute appendicitis. *Scand J Gastroenterol* 1994;**29**:59–66.

297. Fathi-Torbaghan M, Meyer D. MEDUSA: a fuzzy expert system for medical diagnosis of acute abdominal pain. *Methods Inf Med* 1994;**33**:522–9.

298. Gallego M, Fadrique B, Nieto M, Calleja S, Fernandez-Acenero M, Ais G, *et al*. Evaluation of ultrasonography and clinical diagnostic scoring in suspected appendicitis. *Br J Surg* 1998;**85**:37–40.

299. Graff L, Russell J, Seashore J, Tate J, Elwell A, Prete M, *et al*. False negative and false positive errors in abdominal pain evaluation: failure to diagnose acute appendicitis and unnecessary surgery. *Acad Emerg Med* 2000;**7**:1244–55.

300. Graham DF. Computer aided prediction of gangrenous and perforating appendicitis. *BMJ* 1977;**ii**:1375–7.

301. Graham DF, Wyllie FJ. Prediction of gall-stone pancreatitis by computer. *BMJ* 1979;**1**:515–17.

302. Gunn AA. The acute abdomen: the role of computer assisted diagnosis. *Baillière's Clinical Gastroenterology* 1991;**5**:639–65.

303. Ikonen JK, Rokkanen PU, Gronroos P, Kataja JM, Nykanen P, De Dombal FT, *et al*. Presentation and diagnosis of acute abdominal pain in Finland: a computer aided study. *Ann Chir Gynaecol* 1983;**72**:332–6.

304. McAdam WAF, Brock MB, Armitage T, Davenport P, Chan M, De Dombal FT. Twelve years' experience of computer aided diagnosis in a district general hospital. *Ann R Coll Surg Engl* 1990;**72**:140–6.

305. Ohmann C, Moustakis V, Yang Q, Lang K. Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. *Artif Intell Med* 1996;**8**:23–36.

306. Orient JM. Evaluation of abdominal pain: clinicians' performance compared with three protocols. *South Med J* 1986;**79**:793–9.

307. Paterson-Brown S, Vipond MN, Simms K, Gatzen C, Thompson J, Dudley H. Clinical decision making and laparoscopy versus computer prediction in the management of the acute abdomen. *Br J Surg* 1989;**76**:1011–13.

308. Pesonen E, Ikonen JK, Juhola M, Eskelinen M. Parameters for a knowledge base for acute appendicitis. *Methods Inf Med* 1994;**33**:220–6.

309. Puppe B, Ohmann C, Goos K, Puppe F, Mootz O. Evaluating four diagnostic methods with acute abdominal pain cases. *Methods Inf Med* 1995;**34**:361–8.

310. Sturman M, Perez M. Computer assisted diagnosis of acute abdominal pain. *Compr Ther* 1989;**15**:26–35.

311. Talwar S, Talwar R, Prasad P. Continuing diagnostic challenge of acute appendicitis: evaluation through modified Alvarado score: comment. *ANZ J Surg* 1999;**69**:821.

312. Teicher I, Landa B, Cohen M, Kabnick LS, Wise L. Scoring system to aid in diagnoses of appendicitis. *Ann Surg* 1983;**198**:753–9.

313. Wade D, Morrow S, Balsara Z, Burkhard T, Goff W. Accuracy in ultrasound in the diagnosis of acute appendicitis compared with the surgeon's clinical impression. *Arch Surg* 1993;**128**:1039–46.

314. Zielke A, Hasse C, Sitter H, Rothmund M. Influence of ultrasound on clinical decision making in acute appendicitis: a prospective study. *Eur J Surg* 1998;**164**:201–9.

Appendix 1

The definition and scope of decision tools

This appendix is based on a working paper by Liu and colleagues.[247]

Decision support systems as a solution to the gap between knowledge and practice in healthcare

Many problems facing healthcare systems today are caused not by a lack of knowledge but by the gap between what we know and what we do in the face of staff shortage, economic pressures and rising public demand.[251] Systematic reviews or RCTs published in prestigious journals may help to establish the effectiveness of drugs or procedures, but are not enough to ensure that the knowledge is actually used.

The process from innovation to routine clinical use is a complex one. For example, in cardiovascular disease prevention, despite the systematic reviews, evidence-based guidelines and decision tools (e.g. the Joint British Charts), there is continuing evidence to suggest that these approaches have not yet changed actual clinical practice.[204,205] The Leeds AAP system, which estimates patient-specific diagnostic probabilities and underwent extensive development and testing over decades,[51] is scarcely used today. Many factors appear to influence physician uptake of these systems, and the guidelines on which they are based.[85] For example, some health professionals are unaware of, or simply forget, guideline recommendations, while others fail to follow them because of patient choice or peer pressure.

Hundreds of decision support systems (DSSs) and other computerised aids have been developed to assist patient management. In trials, some are effective at narrowing knowledge gaps, improving clinical practice and patient outcomes,[148] but many others are not (e.g. computer-based guidelines on the management of angina and asthma[149]). Why do doctors and other health professionals often fail to adopt the effective DSSs into routine clinical practice? Some developers construct technologically advanced systems with little relevance to the real world, while others create DSSs without first determining whether a clinical need exists.[252,253] The authors believe that there should be a move away from this technology-driven approach to one that entails identifying and using the most effective method to manage knowledge, regardless of whether a high-tech PDA or a low-tech paper reminder is used.

Computerised DSSs (also called decision aids[12]) are fundamentally no different from paper algorithms, nomograms, reminders or other aids to clinical decision-making, because they all aim to improve the appropriateness of clinical actions and patient health outcomes. However, this is an important class of health technology, for which a consistent nomenclature is needed. The authors therefore suggest the generic term ‘decision tool’ to demonstrate that these decision-making aids, which may seem very different from a technical perspective, are conceptually the same from a clinical viewpoint. Examples of decision tools that improve clinical practice include reminders for doctors,[237,254] patient information/support leaflets (e.g. O’Connor[255]) and predictive scores [e.g. the Paediatric Logistic Organ Dysfunction (PELOD) score,[256] the Ottawa Ankle Rules[243] and the Glasgow Coma Scale[257]). Computer-based reminder systems have been shown to be effective in increasing the use of preventive care in both inpatient and outpatient settings.[148,237] Some empirical evidence suggests that DSSs can have more impact than paper-based guidelines and checklists.[85]

Problems with current decision support systems

Although good evidence exists for the clinical benefit of some DSSs, there are also numerous examples of failures and difficulties, for various reasons.

First, current DSSs are rarely based on the best available knowledge. They should incorporate rigorous evidence, such as knowledge derived from well-designed, relevant studies or a large patient database. Secondly, there is usually insufficient emphasis on the need for the health professional or patient to capture high-quality clinical data for the DSS. Thirdly, the development of DSSs is too often technology led. Their true role, of improving decisions and actions about individual patients, is frequently ignored and applies regardless of the technology. A closely related issue is that the most appropriate method should be selected to overcome demonstrated barriers to change,[85] avoiding what Gremy has termed the "idolatry of technology" by those working in medical informatics.[258] Some barriers require education or organisational change to abolish them, not a DSS at all.[85] Fourthly, health technology assessment methods (such as studies on accuracy or impact, systematic reviews and economic analyses) are frequently misapplied.[238,258] Correct application of these methods is necessary to evaluate their impact on clinical practice and their cost-effectiveness.[162] The cost-effectiveness of computer DSSs compared with paper-based decision tools is seldom studied, and was missing from a recent large study on computerised reminders in US hospitals.[237]

A fifth problem eluding the technologists is failure to address broader legal and ethical issues. For example, health professionals using DSSs should always apply their own clinical judgement in the context of the patient and the encounter, and not unthinkingly follow its advice. The system should be designed to treat its user as a 'learned intermediary' and not act as a black box.[71,73] A sixth problem is that developers and users of DSSs too often fail to appreciate that effectiveness and cost-effectiveness will vary according to the user and their context. Finally, DSS developers should become more aware of regulatory issues. Although DSSs are currently exempt from regulation, unlike the closed-loop systems that measure patient variables and automatically adjust a drug infusion device, for example, this may change.[244]

Some of the above failures follow from insufficient clinical and patient involvement, owing partly to the failure to recognise the role of different kinds of DSS and their underlying similarities. However, this position is likely to change as more DSSs are used and some cause medical errors.

The proposed term, decision tool, includes these systems. Those developing DSSs may reject the blanket category, claiming significant differences between subclasses of these systems (e.g. how a specific tool works or how it is developed), in the same way that a chemist will recognise differences between the individual drugs that form a single therapeutic class. However, from the clinical and health policy perspective, such differences are largely irrelevant, as is often the case with drugs from the same class. The authors also disagree with the technologists[175] and believe that there is essentially no difference between evaluating a new drug or a new decision tool. While qualitative methods are necessary to help to elucidate the barriers to change or requirements for a decision tool or reasons for failure, there is no alternative to an RCT to quantify reliably the tool's impact on clinical decisions, actions or patient outcomes.

A definition of decision tool

The following definition of the decision tool is adopted here:

> A 'decision tool' is an active knowledge resource that uses patient data to generate case-specific advice, which supports decision-making about individual patients by health professionals, the patients themselves or others concerned about them.

This definition is an updated and more general version of Wyatt and Spiegelhalter's 1991 definition of computer decision aids: "active knowledge systems which use two or more items of patient data to generate case-specific advice").[162]

Decision tools have four important characteristics:

1. **Target decision maker: is the tool designed to aid a clinical decision by a health professional and/or patient?**
   This characteristic highlights the importance of shared decision-making between health professionals and their patients. Decision aids for health professionals and patients are both included. If the patient is unable to make an informed decision (e.g. someone in a coma or a child), then a carer or relative familiar with his or her condition is an appropriate proxy.
2. **Target decision: do the decisions concern an individual patient?**
   The focus is on decisions about an identified individual patient, rather than on groups of patients (e.g. to support health policy) or on hypothetical patients (e.g. for teaching purposes).
3. **Knowledge component: does the tool use patient data and knowledge to generate an interpretation that aids clinical decision-making?**
   A decision tool must contain some knowledge to help a health professional or patient use patient data to generate an interpretation or aid to decision-making. Examples include (a) explicit

advice, such as a printed recommendation for a course of action, (b) interpretation, such as an asterisk meaning 'this result is abnormal' or a predicted probability of death for a patient on the intensive care unit, and (c) reminders or alerts, such as 'This patient is allergic to penicillin'."

4. **Timing: is the tool used before the health professional or patient takes the relevant decision?**
   A tool used retrospectively, after the relevant decision has been taken, is excluded. Tools that interpret patient data such as test results after a clinical encounter can be classified as decision tools if their output is used during the next encounter.

Examples of tools which fit within or lie outside the definition

Table 12 provides examples of tools that are and are not decision tools using the above definition. A care pathway (example 3) is a preprinted record designed to aid health professionals in recording data and interpreting them as well as in making decisions (fulfilling characteristics 1 and 4) about an individual patient (characteristic 2). It is a knowledge resource for health professionals that enables them actively to use patient data to make decisions (characteristic 3). Clearly, care pathways are decision tools.

Some examples of aids that are not decision tools include distance learning material used away from patients (example 12) and imaging investigations/laboratory tests (examples 15 and 16), which are not knowledge resources (characteristic 3). However, an algorithm or other tool to support interpretation of the results of such tests is a decision tool. For example, a sheet summarising test results would be included if it included knowledge on how to interpret the results and obtain predictions that inform patient management. Some examples depend on the user and current task. For example, a computer-based simulator used to help physicians to develop their diagnostic skills would not be a decision tool if the data they enter are not about a patient they are managing (example 14). However, it is a decision tool if they enter data about a real patient. The typology in *Figure 17* demonstrates the diversity of decision tools.

Conclusion

It is argued that DSSs and other computer-based, paper-based and even mechanical clinical decision aids are members of a wider family, have called decision tools. By viewing decision tools as a group their role in healthcare becomes clearer, which will encourage clinical involvement in developing such tools and evaluating their impact on clinical practice. The excessive emphasis on technology to date has resulted in a disproportionate amount of research conducted by informatics experts and computer scientists, many of whom do not appreciate the crucial need for input from clinicians and epidemiologists in the development and testing of these tools. The identification of decision tools as a coherent and important category of health technology should encourage the sharing of lessons between decision tool developers and users, reduce the frequency of decision tool projects focusing only on technologies, and reduce silo thinking by those in clinical and informatics disciplines. The focus of evaluation should thus become more clinical. It is not sufficient to evaluate the accuracy of computer-based decision tools compared with routine clinical practice or a gold standard.[162] Rather, their impact should be evaluated against other computer-based, paper-based or even mechanical tools, to identify the most cost-effective tool for each clinical problem. It is unlikely that the most cost-effective option will always be computer based.

How should adoption of this decision tool mindset be encouraged? Authors and editors should be encouraged to use the term in titles and abstracts. The present authors propose the inclusion of 'decision tool' as a new Medical Subject Heading (MeSH) term to aid the identification of studies for clinical and research purposes. A joint clinician and decision tool developers' network should be established, with an infrastructure including e-mail lists, web support materials, conferences and a coordinating resource centre. Finally, a case should be made for a multidisciplinary R&D programme on decision tools, jointly supported by clinical and informatics funding bodies.

TABLE 12 *Aids to clinical decisions*

Tool	Characteristics of decision tools			
	1. User **Designed to aid clinical decisions by health professional or patient?**	**2. Target decision** **Decisions about a real individual patient?**	**3. Knowledge component** **Does tool use knowledge to assist interpretation or aid clinical decision-making**	**4. Timing** **Is tool used before health professional or patient makes the relevant decision?**
1. Computerised reminder system for preventive care (e.g. Dexter *et al*[237])	Yes	Yes	Yes	Yes
2. Paper reminder to check for sign X, take into account symptom Y or take action Z when seeing a patient (e.g. Bryce *et al*[259] and Eccles *et al*[254])	Yes	Yes	Yes	Yes
3. Care pathway (e.g. Holtzman *et al*[260])	Yes	Yes	Yes	Yes
4. Tool to enhance shared decision-making (http://www.shared-decision-making.org)	Yes	Yes	Yes	Yes
5. Computerised patient interviewing (checklist for patient to complete, after which the data are presented to the doctor in summary form (e.g. Lilford *et al.*[261])	Yes	Yes	Yes	Yes
6. Sheet for doctor giving definitions of clinical findings in AAP or advice on how to elicit them	Yes	Yes	Yes	Yes
7. Nomograms	Yes	Yes	Yes	Yes
8. Joint British Societies coronary risk prediction charts[205]	Yes	Yes	Yes	Yes
9. Telemedicine system[262]	Yes	Yes	Yes	Yes
10. Information leaflet for patient with AAP	Yes	Yes	Yes	Yes
11. Sheet summarising results of special investigations with advice on interpreting results	Yes	Yes	Yes	Yes
12. Distance learning material used away from patients (e.g. on a course, or self-study) (see examples in Davis *et al.*[263])	Yes	No	Yes	No
13. Monthly performance feedback report (i.e. giving doctors feedback about their performance on previous groups of patients)	Yes	No	No	Yes
14. Computer simulator to help doctors to develop their diagnostic skills (e.g. Hoffer and Barnett[264])	Yes	No	Yes	Yes
15. Imaging investigation, e.g. ultrasonography, computed tomography	Yes	Yes	No	Yes
16. Laboratory test, e.g. white cell count, C-reactive protein	Yes	Yes	No	Yes

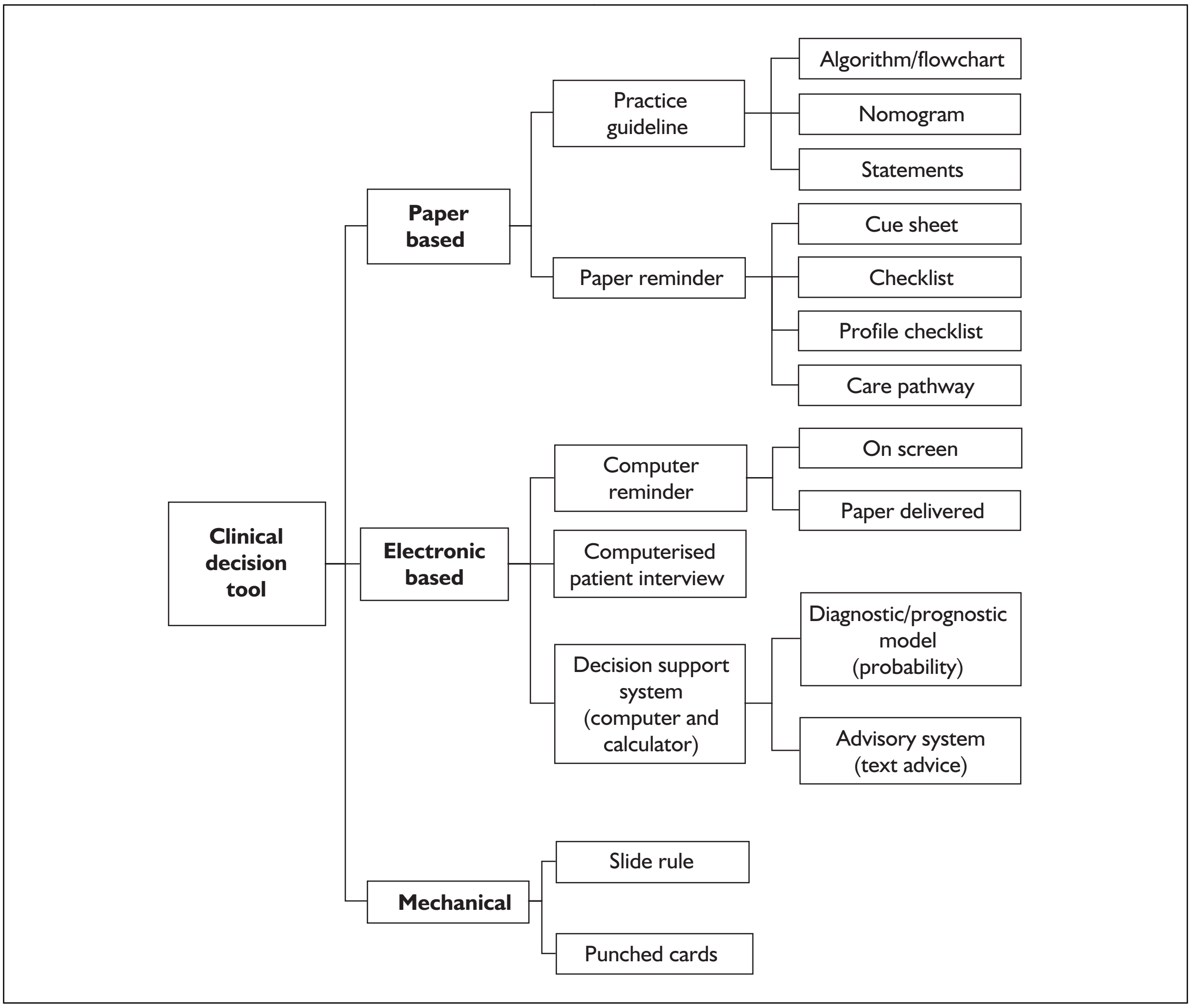

FIGURE 17 *Typology of clinical decision tools*

Appendix 2

Common reasoning methods for decision tools

This appendix examines the more common reasoning methods used by decision tools.

Statistical methods

Bayes' theorem

Bayes' theorem describes how the probability that an individual has disease changes when the result of a diagnostic test is obtained, dependent on the performance characteristics of the test. The theorem originally appeared in a landmark paper in 1763, published posthumously in the *Philosophical Transactions of the Royal Society*.[75] The potential for its application in medicine and health care was first recognised by Ledley and Lusted in 1959.[265] The theorem can be represented as follows:

$$P(D+|S+) = \frac{P(S+|D+) \times (P(D+)}{P(S+|D)P(D+)+P(S|D-)P(D-)} = \frac{\text{Sensitivity} \times \mathrm{P(D+)}}{\text{Sensitivity} \times P(D+) + (1 - \text{Specificity}) \times P(D-)}$$

where $D+$ is the presence of a disease or target condition, $S+$ is the presence of a sign, symptom or prognostic factor, or a positive diagnostic test result, $D-$ is the absence of disease or target condition, and $S-$ is the absence of a sign, symptom or prognostic factor, or a negative diagnostic test result.[266] $P(D+|S+)$ is the probability of having a target condition given the presence of a sign or positive result from a diagnostic test. It is also known as the post-test probability. $P(D+)$ is the prior probability of a target condition, estimated from an empirical study (e.g. a cross-sectional study estimating the prevalence of a disease in a community), or the educated guess or informed belief of a doctor about his or her surgery's case-mix, before a patient is examined for signs and symptoms. Bayes' theorem is more conveniently expressed in terms of odds rather than probabilities:

$$\frac{P(D+|S+)}{P(D-|S+)} = \frac{P(S+|D+)}{P(S+|D-)} \times \frac{P(D+)}{P(D-)}$$

where $\frac{P(S+|D+)}{P(S+|D-)}$ is the likelihood ratio for the result of test S.

Or, in simpler terminology:

$$\text{Odds } (D+|S+) = LR(S) \times \text{Odds}(D+)$$

Where $LR(S)$ is the likelihood for the result of test S.

A nomogram is available that converts pretest probabilities and likelihood ratios into post-test probabilities.[41]

Bayes' theorem can be extended in a simple way to combine multiple pieces of diagnostic information. The post-test probability obtained from the first test can serve as the prior probability for the next test. However, this approach (coined naïve Bayes) has been noted to give overoptimistic predictions owing to double counting of diagnostic information when the individual test results that are being combined are not independent.

Early examples of decision tools that made use of Bayes' theorem include aids for diagnosing congenital heart disease,[267] goitre,[268] jaundice[269] and acute abdominal pain.[113]

Logistic regression extensions of Bayes' theorem

The problem of double counting diagnostic information has been tackled by using logistic regression models, accounting for correlations between pieces of diagnostic information in the same way that confounding factors are adjusted for in epidemiological studies. The links between Bayes' theorem and logistic regression models can be observed by re-expressing the theorem using weights of evidence or log likelihood ratios:

$$\ln \frac{P(D+|S_i+)}{P(D-|S_i+)} = \ln \frac{P(D+)}{P(D-)} + \sum_{i=1}^{N} \ln \frac{P(S_i+|D+)}{P(S_i+|D-)}$$

The first term on the right-hand side of the equation, $\ln \frac{P(D+)}{P(D-)}$, is the natural log of the prior odds of the target condition, because $P(D-) = 1 - P(D+)$. The second term,

$$\sum_{i=1}^{N} \ln \frac{P(S_i+D+)}{P(S_i+D-)}$$

is the sum of the weights of evidence (or log likelihood ratios) from each sign or symptom. The discussion of Bayes' theorem so far has assumed that signs, symptoms and/or test results are independent. Adjustments for correlations between diagnostic items can be made by estimating a parameter β for each test, which either shrinks or enlarges the likelihood ratio for each test:

$$\ln \frac{P(D+|S_i+)}{P(D-|S_i+)} = \ln \frac{P(D+)}{P(D-)} + \sum_{i=1}^{N} \beta_i \frac{P(S_i+|D+)}{P(S_i+|D-)}$$

Two approaches have been described that use logistic regression to estimate the β parameters and fit these models. Spiegelhalter and Knill-Jones described a two-stage procedure that involves a logistic regression model of the unadjusted weights of evidence to obtain adjusted weights of evidence or adjusted likelihood ratios.[76] They used this method to construct a Bayesian decision tool, GLADYS, a diagnostic system for patients with dyspepsia. Albert described how the models can be fitted directly using logistic regression using an offset to include directly the pretest log odds as a fixed term in the model.[270] Albert's approach allows estimates of the adjusted likelihood ratios for combinations of tests results to be obtained.

Different model fitting approaches can be used to select terms for inclusion in these models. Typically, bivariate analyses are used to identify symptoms and signs associated with disease and then multivariate modelling (e.g. logistic regression or discriminant analysis) removes redundant symptoms through the use of stepwise or another variable selection procedure to select the symptoms to keep in the final model. The first step here (often called univariate analysis) is not necessary and can be omitted.

In some situations the parameter estimates (estimated regression coefficients) are replaced by a scoring system derived from the model, and its accuracy and impact as a diagnostic tool are tested. Eskelinen and colleagues' logistic regression models for diagnosing AAP conditions provide an example of such a tool.[115] It is expected that discrimination would improve as more terms (signs and symptoms) are added. Hence, a decision tool with two or more terms is expected to be better than a simple diagnostic test.

It is worth noting that Bayes' theorem has been developed beyond this simple application to create a whole approach to statistical analysis known as Bayesian inference. In Bayesian inference the concept of a prior probability is extended to the specification of prior distributions for every parameter included in a statistical model, likelihood ratios are replaced by likelihood functions, and post-test probabilities are replaced by posterior distributions obtained by combining prior belief distributions with the likelihood functions. Throughout the article, methods that use simple naïve Bayesian probability updating are referred to as Bayesian methods; however, it is noted that they are not based on fully Bayesian methods of the nature described above. The term logistic regression methods will be used to describe approaches that make adjustments for correlations between diagnostic signs and symptoms.

Discrimination rules

Discrimination rules employ statistical methods to produce a rule that can be used to allocate individuals to the group in which they are most likely to belong. For example, it is possible to produce predictions of group membership (disease present or disease absent) from a logistic model, categorising those with a probability of the disease greater than 0.5 into the disease-present group and those with probabilities less than 0.5 into the disease-absent group. Such an approach differs in two respects from the application of logistic regression as described above. First, there is no direct way of altering predictions to allow for differences in pretest probabilities. Secondly, the certainty with which a person is allocated to a particular group is lost. Discriminant analysis may also be based on multivariate normal statistical methods and not logistic functions, and so categorise people into more than two groups.

Clinical algorithms

An algorithm is a process for carrying out a complex task broken down into simple decision and action steps.[12] Clinical algorithms can be represented as paper-based flowcharts or computer programs typically written in standard high-level programming languages such as PASCAL, C or BASIC. Wyatt[103] cited examples of effective decision tools that used algorithms in various settings, such as by American paramedics, British physicians in a tertiary referral centre and primary-care physicians in developing countries. Clinical algorithms have various limitations, including the tendency to assume that all the data specified in the algorithm are available, space restrictions if the algorithm is paper based, and the practical difficulties of breaking down many clinical problems into a set of discrete decisions,[103] that are programmable in high-level languages (preferably as reusable 'objects' of code or abstract data types[271]). These problems are not unique to medicine, but relate to the development of algorithms to solve real-world problems in general. The above limitations have resulted in the use of "more sophisticated symbolic reasoning approaches" by some decision tool developers.[103] Optimal clinical algorithms or pathways can be developed using statistical methods known as classification and regression trees (CART).[272]

Expert systems

According to Shortliffe and colleagues, an expert system is a computer program that simulates human thought processes "to provide the kind of problem analysis and advice that the expert might provide."[11] In a 1956 conference attended by psychologists and computer scientists at Dartmouth College, the scope and potential for computing software to mimic human intelligence and behaviour was discussed and the field of artificial intelligence (AI) was born.[103] As a result, LISP, a symbolic programming language aimed at solving AI problems, was developed. A symbolic programming language is "a language designed to support the representation of knowledge and semantic relationships, while de-emphasising numerical computations".[11] MYCIN was an early expert system developed in the 1970s constructed using AI principles and was designed to manage patients with infections, particularly before conclusive results from cultures are obtained.[273,274] Given the lack of consensus and knowledge in this area at the time, the developers of MYCIN felt that traditional statistical methods or clinical algorithms were 'inadequate' and used symbolic programming (LISP) to develop their system. MYCIN used a large number of what computer scientists call 'production rules', IF–THEN statements that "relate observations to associated inferences that can be drawn".[273] In an evaluation of MYCIN, its accuracy was comparable to that of infectious disease experts.[275] However, as Wyatt pointed out, "it was slow and unwieldy, requiring much data to be input and providing rather limited explanations of its reasoning."[103] MYCIN was never used in a real clinical setting and its development was discontinued in 1980, but as Musen and colleagues pointed out, it provided a basis for the research and development of expert systems in various fields in the 1980s.[273] Wyatt discussed in detail the advantages and disadvantages of decision tools that use symbolic reasoning.[103]

Machine learning

Machine learning can be either supervised or unsupervised.[77] In the former, a system is provided with a sample of data and instructions on how to identify and classify patterns within the data by a trainer. In the latter, a system is also provided with data, but is left to identify patterns without external assistance, a form of cluster analysis. There are various types of machine learning method: decision trees, artificial neural networks and genetic algorithms. For all three methods, internal weights within the system are adjusted during training until a prespecified level of performance is attained. Although experts can evaluate the decision tree generated by a machine learning system, such a system often cannot provide understandable reasons for the advice it generates. Experts cannot evaluate the reasoning behind the classifiers generated by neural networks and genetic algorithms, because they are essentially 'black boxes'.[73] On the technical side, Schwarzer and colleagues' review concluded that they suffered from many problems, such as the fitting of models that are implausible and the tendency for neural networks to understate misclassification errors.[276] For introductory discussions on artificial neural networks and genetic algorithms, see Musen and colleagues,[273] Carter[77] and Wyatt.[193] For more in-depth coverage in this area, see Aleksander and Morton.[140]

Some comments on the potential of different reasoning methods

Some of the more common reasoning methods used in decision tools were briefly outlined: Bayesian probability updating, discrimination rules (such as logistic regression models and discriminant analysis), clinical algorithms, expert systems and machine learning systems (directed trees, artificial neural networks and genetic algorithms). Tools that are promising are ones that are simple to use, have clinical credibility, and have been shown to be effective in aiding decisions on patient care. Doctors are more likely to accept Bayesian probability updating and discrimination rules because their underlying reasoning methods are easier to understand and make clinical sense. Expert systems and machine learning systems suffer from the ‘black box’ problem mentioned earlier, as well as other problems including legal and ethical issues and the tendency of such systems to underestimate probabilities of misclassification.[276] Rigorously conducted evaluations can be used to compare the relative performance of these different types of decision tool.[71,73]

Appendix 3
Search terms used

Medline search strategies

MEDLINE search 1

1. explode "Abdomen-Acute"/ all subheadings
2. acute
3. abdom*
4. pain*
5. acute near abdom* near pain*
6. #1 or #5
7. (explode "appendicitis"/ all subheadings) or ("appendectomy"/ all subheadings) or ("appendix"/ all subheadings)
8. appendic* or appendec* or appendicec*
9. #6 or #7 or #8
10. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model* or decision support system* or computer*
11. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
12. expert system* or neural network* or artificial intellig* or machine learning or Bayes*
13. (explode "decision-support-systems-clinical"/ all subheadings) or (explode "decision-support-systems-management"/ all subheadings) or (explode "decision-support-techniques"/ all subheadings) or (explode "artificial-intelligence"/ all subheadings)
14. (explode "decision-making-computer-assisted"/ all subheadings) or (explode "medical-informatics"/ all subheadings) or (explode "information-systems"/ all subheadings) or (explode "decision-making"/ all subheadings)
15. (explode "Reminder-Systems"/ all subheadings) or (explode "Hospital-Information-Systems"/ all subheadings) or (explode "Management-Information-Systems"/ all subheadings) or (explode "Medical-Records-Systems-Computerized"/ all subheadings)
16. (explode "Computers"/ all subheadings) or information system* or informatic*
17. #10 or #11 or #12 or #13 or #14 or #15 or #16
18. (explode "Sensitivity-and-Specificity"/ all subheadings) or (explode "Predictive-Value-of-Tests"/ all subheadings) or (explode "ROC-Curve"/ all subheadings)
19. specificit* or sensitivit* or false negative* or predictive value* or likelihood ratio* or accuracy
20. #18 or #19
21. #20 or #17
22. #21 and #9

MEDLINE search 2

1. explode "Abdomen-Acute"/ all subheadings
2. acute
3. abdom*
4. pain*
5. acute near abdom* near pain*
6. #1 or #5
7. (explode "appendicitis"/ all subheadings) or ("appendectomy"/ all subheadings) or ("appendix"/ all subheadings)
8. appendic* or appendec* or appendicec*
9. #6 or #7 or #8
10. abdom*
11. pain*
12. abdom* near pain*
13. #12 not #9
14. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model* or decision support system* or computer*
15. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
16. expert system* or neural network* or artificial intellig* or machine learning or Bayes*
17. (explode "decision-support-systems-clinical"/ all subheadings) or (explode "decision-support-systems-management"/ all subheadings) or (explode "decision-support-techniques"/ all subheadings) or (explode "artificial-intelligence"/ all subheadings)
18. (explode "decision-making-computer-assisted"/ all subheadings) or (explode "medical-informatics"/ all subheadings) or (explode "information-systems"/ all subheadings) or (explode "decision-making"/ all subheadings)
19. (explode "Reminder-Systems"/ all subheadings) or (explode "Hospital-Information-Systems"/ all subheadings) or (explode "Management-Information-Systems"/ all subheadings) or (explode "Medical-Records-Systems-Computerized"/ all subheadings)
20. (explode "Computers"/ all subheadings) or information system* or informatic*
21. #14 or #15 or #16 or #17 or #18 or #19 or #20

22. (explode "Sensitivity-and-Specificity"/ all subheadings) or (explode "Predictive-Value-of-Tests"/ all subheadings) or (explode "ROC-Curve"/ all subheadings)
23. specificit* or sensitivit* or false negative* or predictive value* or likelihood ratio* or accuracy
24. #22 or #23
25. #21 or #24
26. #13 and #25

MEDLINE search 3

1. non?specific
2. abdominal
3. pain*
4. non?specific abdominal pain*
5. NSAP
6. cholecystitis
7. explode "gallbladder-diseases"/ all subheadings
8. perfora*
9. ulcer*
10. intestin*
11. perfora* near (ulcer* or intestin*)
12. pancreatitis
13. explode "pancreatitis"/ all subheadings
14. diverticulitis
15. explode "diverticulitis"/ all subheadings
16. explode "diverticulitis-colonic"/ all subheadings
17. explode "diverticulum"/ all subheadings
18. explode "diverticulum-stomach"/ all subheadings
19. diverticular
20. disease*
21. diverticular disease*
22. small
23. bowel
24. obstruction*
25. small bowel obstruction*
26. intestin*
27. obstruction*
28. intestin* obstruction*
29. explode "intestinal-obstruction"/ all subheadings
30. gyn?ecologic
31. disease*
32. gyn?ecologic disease*
33. ectopic
34. pregnan*
35. ectopic pregnan*
36. explode "pregnancy-ectopic"/ all subheadings
37. colic
38. explode "colic"/ all subheadings
39. pelvic
40. inflammatory
41. disease*
42. pelvic inflammatory disease*
43. salpingi*
44. explode "Salpingitis"/ all subheadings
45. #4 or #5 or #6 or #7 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18
46. #21 or #25 or #28 or #29 or #32 or #35 or #36 or #37 or #38 or #42 or #43 or #44
47. #45 or #46
48. explode "Abdomen-Acute"/ all subheadings
49. acute
50. abdom*
51. pain*
52. acute near abdom* near pain*
53. #48 or #52
54. (explode "appendicitis"/ all subheadings) or ("appendectomy"/ all subheadings) or ("appendix"/ all subheadings)
55. appendic* or appendec* or appendicec*
56. #53 or #54 or #55
57. abdom*
58. pain*
59. abdom* near pain*
60. #59 or #56
61. #47 not #60
62. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model* or decision support system* or computer*
63. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
64. expert system* or neural network* or artificial intellig* or machine learning or Bayes*
65. (explode "decision-support-systems-clinical"/ all subheadings) or (explode "decision-support-systems-management"/ all subheadings) or (explode "decision-support-techniques"/ all subheadings) or (explode "artificial-intelligence"/ all subheadings)
66. (explode "decision-making-computer-assisted"/ all subheadings) or (explode "medical-informatics"/ all subheadings) or (explode "information-systems"/ all subheadings) or (explode "decision-making"/ all subheadings)
67. (explode "Reminder-Systems"/ all subheadings) or (explode "Hospital-Information-Systems"/ all subheadings) or (explode "Management-Information-Systems"/ all subheadings) or (explode "Medical-Records-Systems-Computerized"/ all subheadings)
68. (explode "Computers"/ all subheadings) or information system* or informatic*
69. #62 or #63 or #64 or #65 or #66 or #67 or #68
70. #61 and #69

MEDLINE search 4

1. non?specific

2. abdominal
3. pain*
4. non?specific abdominal pain*
5. NSAP
6. cholecystitis
7. explode "gallbladder-diseases"/ all subheadings
8. perfora*
9. ulcer*
10. intestin*
11. perfora* near (ulcer* or intestin*)
12. pancreatitis
13. explode "pancreatitis"/ all subheadings
14. diverticulitis
15. explode "diverticulitis"/ all subheadings
16. explode "diverticulitis-colonic"/ all subheadings
17. explode "diverticulum"/ all subheadings
18. explode "diverticulum-stomach"/ all subheadings
19. diverticular
20. disease*
21. diverticular disease*
22. small
23. bowel
24. obstruction*
25. small bowel obstruction*
26. intestin*
27. obstruction*
28. intestin* obstruction*
29. explode "intestinal-obstruction"/ all subheadings
30. gyn?ecologic
31. disease*
32. gyn?ecologic disease*
33. ectopic
34. pregnan*
35. ectopic pregnan*
36. explode "pregnancy-ectopic"/ all subheadings
37. colic
38. explode "colic"/ all subheadings
39. pelvic
40. inflammatory
41. disease*
42. pelvic inflammatory disease*
43. salpingi*
44. explode "Salpingitis"/ all subheadings
45. #4 or #5 or #6 or #7 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18
46. #21 or #25 or #28 or #29 or #32 or #35 or #36 or #37 or #38 or #42 or #43 or #44
47. #45 or #46
48. explode "Abdomen-Acute"/ all subheadings
49. acute
50. abdom*
51. pain*
52. acute near abdom* near pain*
53. #48 or #52
54. (explode "appendicitis"/ all subheadings) or ("appendectomy"/ all subheadings) or ("appendix"/ all subheadings)
55. appendic* or appendec* or appendicec*
56. #53 or #54 or #55
57. abdom*
58. pain*
59. abdom* near pain*
60. #59 or #56
61. #47 not #60
62. (explode "Sensitivity-and-Specificity"/ all subheadings) or (explode "Predictive-Value-of-Tests"/ all subheadings) or (explode "ROC-Curve"/ all subheadings)
63. specificit* or sensitivit* or false negative* or predictive value* or likelihood ratio* or accuracy
64. #62 or #63
65. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model* or decision support system* or computer*
66. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
67. expert system* or neural network* or artificial intellig* or machine learning or Bayes*
68. (explode "decision-support-systems-clinical"/ all subheadings) or (explode "decision-support-systems-management"/ all subheadings) or (explode "decision-support-techniques"/ all subheadings) or (explode "artificial-intelligence"/ all subheadings)
69. (explode "decision-making-computer-assisted"/ all subheadings) or (explode "medical-informatics"/ all subheadings) or (explode "information-systems"/ all subheadings) or (explode "decision-making"/ all subheadings)
70. (explode "Reminder-Systems"/ all subheadings) or (explode "Hospital-Information-Systems"/ all subheadings) or (explode "Management-Information-Systems"/ all subheadings) or (explode "Medical-Records-Systems-Computerized"/ all subheadings)
71. (explode "Computers"/ all subheadings) or information system* or informatic*
72. #65 or #66 or #67 or #68 or #69 or #70 or #71
73. #64 not #72
74. #61 and #73

EMBASE search strategies

EMBASE search 1

1. explode "acute-abdomen"/ all subheadings
2. acute*
3. abdom*
4. pain*
5. acute* near abdom* near pain*

6. explode "appendicitis"/ all subheadings
7. explode "acute-appendicitis"/ all subheadings
8. explode "appendectomy"/ all subheadings
9. explode "appendix"/ all subheadings
10. appendic*
11. appendec*
12. appendicec*
13. appendic* or appendec* or appendicec*
14. #1 or #5 or #6 or #7 or #8 or #9 or #13
15. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model* or decision support system* or computer*
16. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
17. expert system* or neural network* or artificial intellig* or machine learning or Bayes* or information system* or informatic*
18. explode "decision-support-system"/ all subheadings
19. explode "artificial-intelligence"/ all subheadings
20. explode "computer-assisted-diagnosis"/ all subheadings
21. explode "computer"/ all subheadings
22. explode "computer-analysis"/ all subheadings
23. explode "computer-assisted-drug-therapy"/ all subheadings
24. explode "computer-assisted-therapy"/ all subheadings
25. explode "computer-interface"/ all subheadings
26. explode "computer-model"/ all subheadings
27. explode "computer-network"/ all subheadings
28. explode "computer-prediction"/ all subheadings
29. explode "computer-program"/ all subheadings
30. explode "computer-system"/ all subheadings
31. explode "digital-computer"/ all subheadings
32. explode "microcomputer"/ all subheadings
33. explode "mathematical-computing"/ all subheadings
34. explode "artificial-neural-network"/ all subheadings
35. explode "online-system"/ all subheadings
36. explode "decision-making"/ all subheadings
37. explode "information-system"/ all subheadings

38 explode "medical-decision-making"/ all subheadings

39. explode "medical-informatics"/ all subheadings
40. explode "reminder-system"/ all subheadings
41. explode "hospital-information-system"/ all subheadings
42. #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24
43. #25 or #26 or #27 or #28 or #29 or #30 or #31 or #32 or #33 or #34
44. #35 or #36 or #37 or #38 or #39 or #40 or #41
45. #42 or #43 or #44
46. specificit* or sensitivit* or false negative* or predictive value* or likelihood ratio* or accuracy
47. explode "sensitivity-and-specificity"/ all subheadings
48. explode "prediction-and-forecasting"/ all subheadings
49. explode "roc-curve"/ all subheadings
50. explode "receiver-operating-characteristic"/ all subheadings
51. #46 or #47 or #48 or #49 or #50
52. #51 or #45
53. #52 and #14

EMBASE search 2

1. explode "acute-abdomen"/ all subheadings
2. acute*
3. abdom*
4. pain*
5. acute* near abdom* near pain*
6. explode "appendicitis"/ all subheadings
7. explode "acute-appendicitis"/ all subheadings
8. explode "appendectomy"/ all subheadings
9. explode "appendix"/ all subheadings
10. appendic*
11. appendec*
12. appendicec*
13. appendic* or appendec* or appendicec*
14. #1 or #5 or #6 or #7 or #8 or #9 or #13
15. explode "abdominal-pain"/ all subheadings
16. abdom*
17. pain*
18. abdom* near pain*
19. #18 or #15
20. #19 not #14
21. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model* or decision support system* or computer*
22. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
23. expert system* or neural network* or artificial intellig* or machine learning or Bayes* or information system* or informatic*
24. explode "decision-support-system"/ all subheadings
25. explode "artificial-intelligence"/ all subheadings
26. explode "computer-assisted-diagnosis"/ all subheadings
27. explode "computer"/ all subheadings
28. explode "computer-analysis"/ all subheadings
29. explode "computer-assisted-drug-therapy"/ all subheadings
30. explode "computer-assisted-therapy"/ all subheadings

31. explode "computer-interface"/ all subheadings
32. explode "computer-model"/ all subheadings
33. explode "computer-network"/ all subheadings
34. explode "computer-prediction"/ all subheadings
35. explode "computer-program"/ all subheadings
36. explode "computer-system"/ all subheadings
37. explode "digital-computer"/ all subheadings
38. explode "microcomputer"/ all subheadings
39. explode "mathematical-computing"/ all subheadings
40. explode "artificial-neural-network"/ all subheadings
41. explode "online-system"/ all subheadings
42. explode "decision-making"/ all subheadings
43. explode "information-system"/ all subheadings
44. explode "medical-decision-making"/ all subheadings
45. explode "medical-informatics"/ all subheadings
46. explode "reminder-system"/ all subheadings
47. explode "hospital-information-system"/ all subheadings
48. #21 or #22 or #23 or #24 or #25 or #26 or #27 or #28 or #29 or #30
49. #31 or #32 or #33 or #34 or #35 or #36 or #37 or #38 or #39 or #40
50. #41 or #42 or #43 or #44 or #45 or #46 or #47
51. #48 or #49 or #50
52. specificit* or sensitivit* or false negative* or predictive value* or likelihood ratio* or accuracy
53. explode "sensitivity-and-specificity"/ all subheadings
54. explode "prediction-and-forecasting"/ all subheadings
55. explode "roc-curve"/ all subheadings
56. explode "receiver-operating-characteristic"/ all subheadings
57. #52 or #53 or #54 or #55 or #56
58. #57 or #51
59. #58 and #20

EMBASE search 3

1. non?specific
2. abdominal
3. pain
4. non
5. specific
6. abdominal
7. pain
8. non?specific abdominal pain or non specific abdominal pain
9. NSAP
10. cholecystitis
11. explode "gallbladder-disease"/ all subheadings
12. perfora*
13. ulcer*
14. intestin*
15. perfora* near (ulcer* or intestin*)
16. pancreatitis
17. explode "pancreatitis"/ all subheadings
18. diverticulitis
19. explode "diverticulitis"/ all subheadings
20. explode "colon-diverticulosis"/ all subheadings
21. explode "diverticulosis"/ all subheadings
22. diverticular
23. disease*
24. diverticular disease*
25. diverticulosis
26. small
27. bowel
28. obstruction*
29. small bowel obstruction*
30. intestin*
31. obstruction*
32. intestin* obstruction*
33. explode "intestine-obstruction"/ all subheadings
34. gyn?ecologic
35. disease*
36. gyn?ecologic disease*
37. ectopic
38. pregnan*
39. ectopic pregnan*
40. explode "ectopic-pregnancy"/ all subheadings
41. colic
42. explode "colic"/ all subheadings
43. pelvic
44. inflammatory
45. disease*
46. pelvic inflammatory disease*
47. salpingi*
48. explode "salpingitis"/ all subheadings
49. explode "pelvic-inflammatory-disease"/ all subheadings
50. #8 or #9 or #10 or #11 or #15 or #16 or #17 or #18 or #19 or #20 or #21
51. #24 or #25 or #29 or #32 or #33 or #36 or #39 or #40 or #41 or #42
52. #46 or #47 or #48 or #49
53. #50 or #51 or #52
54. explode "acute-abdomen"/ all subheadings
55. acute*
56. abdom*
57. pain*
58. acute* near abdom* near pain*
59. explode "appendicitis"/ all subheadings
60. explode "acute-appendicitis"/ all subheadings
61. explode "appendectomy"/ all subheadings
62. explode "appendix"/ all subheadings
63. appendic*
64. appendec*
65. appendicec*
66. appendic* or appendec* or appendicec*

67. #54 or #58 or #59 or #60 or #61 or #62 or #66
68. explode "abdominal-pain"/ all subheadings
69. abdom*
70. pain*
71. abdom* near pain*
72. #71 or #68
73. #72 or #67
74. #53 not #73
75. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model* or decision support system* or computer*
76. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
77. expert system* or neural network* or artificial intellig* or machine learning or Bayes* or information system* or informatic*
78. explode "decision-support-system"/ all subheadings
79. explode "artificial-intelligence"/ all subheadings
80. explode "computer-assisted-diagnosis"/ all subheadings
81. explode "computer"/ all subheadings
82. explode "computer-analysis"/ all subheadings
83. explode "computer-assisted-drug-therapy"/ all subheadings
84. explode "computer-assisted-therapy"/ all subheadings
85. explode "computer-interface"/ all subheadings
86. explode "computer-model"/ all subheadings
87. explode "computer-network"/ all subheadings
88. explode "computer-prediction"/ all subheadings
89. explode "computer-program"/ all subheadings
90. explode "computer-system"/ all subheadings
91. explode "digital-computer"/ all subheadings
92. explode "microcomputer"/ all subheadings
93. explode "mathematical-computing"/ all subheadings
94. explode "artificial-neural-network"/ all subheadings
95. explode "online-system"/ all subheadings
96. explode "decision-making"/ all subheadings
97. explode "information-system"/ all subheadings
98. explode "medical-decision-making"/ all subheadings
99. explode "medical-informatics"/ all subheadings
100. explode "reminder-system"/ all subheadings
101. explode "hospital-information-system"/ all subheadings
102. #75 or #76 or #77 or #78 or #79 or #80 or #81 or #82 or #83 or #84
103. #85 or #86 or #87 or #88 or #89 or #90 or #91 or #92 or #93 or #94
104. #95 or #96 or #97 or #98 or #99 or #100 or #101
105. #102 or #103 or #104
106. #105 and #74

EMBASE search 4

1. non?specific
2. abdominal
3. pain
4. non
5. specific
6. abdominal
7. pain
8. non?specific abdominal pain or non specific abdominal pain
9. NSAP
10. cholecystitis
11. explode "gallbladder-disease"/ all subheadings
12. perfora*
13. ulcer*
14. intestin*
15. perfora* near (ulcer* or intestin*)
16. pancreatitis
17. explode "pancreatitis"/ all subheadings
18. diverticulitis
19. explode "diverticulitis"/ all subheadings
20. explode "colon-diverticulosis"/ all subheadings
21. explode "diverticulosis"/ all subheadings
22. diverticular
23. disease*
24. diverticular disease*
25. diverticulosis
26. small
27. bowel
28. obstruction*
29. small bowel obstruction*
30. intestin*
31. obstruction*
32. intestin* obstruction*
33. explode "intestine-obstruction"/ all subheadings
34. gyn?ecologic
35. disease*
36. gyn?ecologic disease*
37. ectopic
38. pregnan*
39. ectopic pregnan*
40. explode "ectopic-pregnancy"/ all subheadings
41. colic
42. explode "colic"/ all subheadings
43. pelvic
44. inflammatory
45. disease*
46. pelvic inflammatory disease*
47. salpingi*

48. explode "salpingitis"/ all subheadings
49. explode "pelvic-inflammatory-disease"/ all subheadings
50. #8 or #9 or #10 or #11 or #15 or #16 or #17 or #18 or #19 or #20 or #21
51. #24 or #25 or #29 or #32 or #33 or #36 or #39 or #40 or #41 or #42
52. #46 or #47 or #48 or #49
53. #50 or #51 or #52
54. explode "acute-abdomen"/ all subheadings
55. acute*
56. abdom*
57. pain*
58. acute* near abdom* near pain*
59. explode "appendicitis"/ all subheadings
60. explode "acute-appendicitis"/ all subheadings
61. explode "appendectomy"/ all subheadings
62. explode "appendix"/ all subheadings
63. appendic*
64. appendec*
65. appendicec*
66. appendic* or appendec* or appendicec*
67. #54 or #58 or #59 or #60 or #61 or #62 or #66
68. explode "abdominal-pain"/ all subheadings
69. abdom*
70. pain*
71. abdom* near pain*
72. #71 or #68
73. #72 or #67
74. #53 not #73
75. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model* or decision support system* or computer*
76. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
77. expert system* or neural network* or artificial intellig* or machine learning or Bayes* or information system* or informatic*
78. explode "decision-support-system"/ all subheadings
79. explode "artificial-intelligence"/ all subheadings
80. explode "computer-assisted-diagnosis"/ all subheadings
81. explode "computer"/ all subheadings
82. explode "computer-analysis"/ all subheadings
83. explode "computer-assisted-drug-therapy"/ all subheadings
84. explode "computer-assisted-therapy"/ all subheadings
85. explode "computer-interface"/ all subheadings
86. explode "computer-model"/ all subheadings
87. explode "computer-network"/ all subheadings
88. explode "computer-prediction"/ all subheadings
89. explode "computer-program"/ all subheadings
90. explode "computer-system"/ all subheadings
91. explode "digital-computer"/ all subheadings
92. explode "microcomputer"/ all subheadings
93. explode "mathematical-computing"/ all subheadings
94. explode "artificial-neural-network"/ all subheadings
95. explode "online-system"/ all subheadings
96. explode "decision-making"/ all subheadings
97. explode "information-system"/ all subheadings
98. explode "medical-decision-making"/ all subheadings
99. explode "medical-informatics"/ all subheadings
100. explode "reminder-system"/ all subheadings
101. explode "hospital-information-system"/ all subheadings
102. #75 or #76 or #77 or #78 or #79 or #80 or #81 or #82 or #83 or #84
103. #85 or #86 or #87 or #88 or #89 or #90 or #91 or #92 or #93 or #94
104. #95 or #96 or #97 or #98 or #99 or #100 or #101
105. #102 or #103 or #104
106. specificit* or sensitivit* or false negative* or predictive value* or likelihood ratio* or accuracy
107. explode "sensitivity-and-specificity"/ all subheadings
108. explode "prediction-and-forecasting"/ all subheadings
109. explode "roc-curve"/ all subheadings
110. explode "receiver-operating-characteristic"/ all subheadings
111. #106 or #107 or #108 or #109 or #110
112. #111 not #105
113. #74 and #112

CINAHL search strategies

CINAHL search 1

1. explode "Abdomen-Acute"/ all topical subheadings / all age subheadings
2. abdom*
3. pain*
4. abdom* near pain*
5. acute
6. abdomen
7. acute abdomen
8. appendic*
9. appendec*
10. appendicec*
11. appendic* or appendec* or appendicec*
12. explode "Appendicitis"/ all topical subheadings / all age subheadings

13. explode "Appendix"/ all topical subheadings / all age subheadings
14. explode "Appendectomy"/ all topical subheadings / all age subheadings
15. #1 or #4 or #7 or #11 or #12 or #13 or #14
16. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model or decision support system* or computer*
17. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
18. expert system* or neural network* or artificial intellig* or machine learning or Bayes*
19. explode "Decision-Making"/ all topical subheadings / all age subheadings
20. explode "Computing-Methodologies"/ all topical subheadings / all age subheadings
21. explode "Health-Information-Systems"/ all topical subheadings / all age subheadings
22. explode "Management-Information-Systems"/ all topical subheadings / all age subheadings
23. explode "Information-Science"/ all topical subheadings / all age subheadings
24. explode "Information-Systems"/ all topical subheadings / all age subheadings
25. explode "Artificial-Intelligence"/ all topical subheadings / all age subheadings
26. explode "Computerized-Patient-Record"/ all topical subheadings / all age subheadings
27. explode "Microcomputers"/ all topical subheadings / all age subheadings
28. explode "Checklists"/ all topical subheadings / all age subheadings
29. explode "Algorithms"/ all topical subheadings / all age subheadings
30. explode "Clinical-Assessment-Tools"/ all topical subheadings / all age subheadings
31. #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23
32. #24 or #25 or #26 or #27 or #28 or #29 or #30
33. #31 or #32
34. specificit* or sensitivit* or false negative* or predictive value* or likelihood ratio* or accuracy
35. explode "Sensitivity-and-Specificity"/ all topical subheadings / all age subheadings
36. explode "Predictive-Validity"/ all topical subheadings / all age subheadings
37. ROC curve
38. receiver operating characteristic curve
39. receiver operator characteristic curve
40. #34 or #35 or #36 or #37 or #38 or #39
41. #40 or #33
42. #15 and #41

CINAHL search 2

1. appendicitis not (acute appendicitis)
2. ((explode "Abdomen-Acute"/ all topical subheadings / all age subheadings) or (acute abdomen)) not acute abdominal pain
3. non?specific abdomin* pain*
4. NSAP
5. cholecystitis
6. explode "Cholecystitis"/ all topical subheadings / all age subheadings
7. gall?bladder disease
8. explode "Gallbladder-Diseases"/ all topical subheadings / all age subheadings
9. perfora* near (ulcer* or intestin*)
10. pancreatitis
11. explode "Pancreatitis"/ all topical subheadings / all age subheadings
12. diverticulitis
13. diverticular disease*
14. explode "Diverticulum"/ all topical subheadings / all age subheadings
15. explode "Diverticulitis"/ all topical subheadings / all age subheadings
16. small bowel obstruct*
17. intestin* obstruction*
18. explode "Intestinal-Obstruction"/ all topical subheadings / all age subheadings
19. gyn?ecologic* disease*
20. ectopic pregnanc*
21. explode "Pregnancy-Ectopic"/ all topical subheadings / all age subheadings
22. colic
23. explode "Colic"/ all topical subheadings / all age subheadings
24. pelvic inflammatory disease*
25. salpingi*
26. explode "Salpingitis"/ all topical subheadings / all age subheadings
27. explode "Pelvic-Inflammatory-Disease"/ all topical subheadings / all age subheadings
28. explode "Salpingitis"/ all topical subheadings / all age subheadings
29. #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10. or #11 or #12 or #13 or #14
30. #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24 or #25 or #26 or #27 or #28
31. #29 or #30
32. explode "Abdomen-Acute"/ all topical subheadings / all age subheadings
33. abdom*
34. pain*
35. abdom* near pain*
36. acute
37. abdomen
38. acute abdomen

39. appendic*
40. appendec*
41. appendicec*
42. appendic* or appendec* or appendicec*
43. explode "Appendicitis"/ all topical subheadings / all age subheadings
44. explode "Appendix"/ all topical subheadings / all age subheadings
45. explode "Appendectomy"/ all topical subheadings / all age subheadings
46. #32 or #35 or #38 or #42 or #43 or #44 or #45
47. #31 not #46
48. checklist* or algorith* or slide rule* or calculator* or scor* or practice guideline* or progno* model or decision support system* or computer*
49. decision tree* or decision analy* or decision aid* or decision tool* or advisory system* or nomogram*
50. expert system* or neural network* or artificial intellig* or machine learning or Bayes*

51 explode "Decision-Making"/ all topical subheadings / all age subheadings

52. explode "Computing-Methodologies"/ all topical subheadings / all age subheadings
53. explode "Health-Information-Systems"/ all topical subheadings / all age subheadings
54. explode "Management-Information-Systems"/ all topical subheadings / all age subheadings
55. explode "Information-Science"/ all topical subheadings / all age subheadings
56. explode "Information-Systems"/ all topical subheadings / all age subheadings
57. explode "Artificial-Intelligence"/ all topical subheadings / all age subheadings
58. explode "Computerized-Patient-Record"/ all topical subheadings / all age subheadings
59. explode "Microcomputers"/ all topical subheadings / all age subheadings
60. explode "Checklists"/ all topical subheadings / all age subheadings
61. explode "Algorithms"/ all topical subheadings / all age subheadings
62. explode "Clinical-Assessment-Tools"/ all topical subheadings / all age subheadings
63. #48 or #49 or #50 or #51 or #52 or #53 or #54 or #55
64. #56 or #57 or #58 or #59 or #60 or #61 or #62
65. #63 or #64
66. specificit* or sensitivit* or false negative* or predictive value* or likelihood ratio* or accuracy
67. explode "Sensitivity-and-Specificity"/ all topical subheadings / all age subheadings
68. explode "Predictive-Validity"/ all topical subheadings / all age subheadings
69. ROC curve
70. receiver operating characteristic curve
71. receiver operator characteristic curve
72. #66 or #67 or #68 or #69 or #70 or #71
73. #72 or #65
74. #73 and #47

INSPEC search strategy

1. acute near abdom* near pain*
2. acute abdomen
3. appendicitis
4. abdominal pain
5. NSAP
6. cholecystitis
7. perfora* near ulcer
8. perfora* near intestin*
9. pancreatitis
10. diverticular disease
11. small bowel near obstruction*
12. acute gyn?ecological disease* or ectopic pregnanc*
13. colic*
14. salpingitis or pelvic inflammatory disease*
15. non specific abdominal pain or non?specific abdominal pain

* 16. #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15

SIGLE search strategy

1. acute near abdom* near pain*
2. acute abdomen
3. appendicitis
4. abdominal pain
5. NSAP
6. cholecystitis
7. perfora* near ulcer
8. perfora* near intestin*
9. pancreatitis
10. diverticular disease
11. small bowel near obstruction*
12. acute gyn?ecological disease* or ectopic pregnanc*
13. colic*
14. salpingitis or pelvic inflammatory disease*
15. #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14

DH-Data search strategy

1. abdom*
2. pain*

3. appendici*
4. appendec*
5. appendicec*
6. NSAP
7. (abdom* near pain*) or appendici* or appendec* or appendicec* or NSAP

HEALTH-CD search strategy

1. abdom*
2. pain*
3. appendici*
4. appendec*
5. appendicec*
6. NSAP
7. (abdom* near pain*) or appendici* or appendec* or appendicec* or NSAP

CENTRAL search strategy

appendici* OR acute abdomen OR (abdomin* near pain)

Appendix 4

Data collection forms

RAPT eligibility criteria form for review questions 1 (accuracy studies) and 2 (impact studies)

*Paper ID*________________ *Reviewer ID*________________ *Date*________________

A	Did patients have previously undiagnosed acute abdominal pain (AAP) lasting ≤7 days from onset? If AAP, was a decision tool or unaided doctors' decisions studied? **If no to either question, go to OUT**	Yes □ No □ Unclear □ Yes □ No □ Unclear □
B	Did the study evaluate one of the following aspects of a decision tool and/or unaided doctors' decision? 1. diagnostic accuracy of decision tool 2. diagnostic accuracy of unaided doctors' decisions 3. accuracy of decisions to order tests 4. accuracy or appropriateness of referral of patients 5. accuracy or appropriateness of other clinical management decisions 6. impact of decision tools/unaided doctors on decisions 7. impact of decision tools/unaided doctors on actions 8. impact of decision tools/unaided doctors on patient outcomes 9. impact of decision tools/unaided doctors on use of health care resources 10. Other: Specify ____________ **Stop & consult Steering Gp**	Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □
C	If study was on diagnostic accuracy, was/were reference standard(s) specified? If yes, what type of study was it? (check one of the following) retrospective/prospective cohort □ case–control □ If study was an impact study, was the design one of the following? (check one of the following) Randomised □ Quasi-randomised □	Yes □ No □ Unclear □ Check one Yes □ No □ Unclear □ Check one
D	Were the following measures/results reported or calculable from data in study? • For accuracy studies, sensitivity, specificity, likelihood ratio or similar (e.g. 2 × 2 tables, area under ROC curves, +ve/–ve predictive values) against reference standard • For impact studies, measures of impact of decision tool on patient outcomes, decisions, actions and/or use of health care resources, e.g. relative risk, absolute risk, odds ratio, quantity or cost of resources used	Yes □ No □ Unclear □ Yes □ No □ Unclear □
In	Eligible if yes to both of A **and** one of B **and** one of C **and** one of D (Check 1)	Eligible □ Ineligible □ Unclear □

Out	Reason if excluded (Check as appropriate) Not AAP ≤7 days □ Not decision tool □ Inappropriate study question (i.e. none of **B**) □ No original data □ No appropriate results □ No reference standard (accuracy study only) □ Non-randomised study (impact study only) □ Other ______________________

Quality assessment form for study question 1 (accuracy studies)

Paper ID	**Reviewer ID**	**Date**
Type of data set	Training set □ Test set □	
Reference Std 1	What is reference standard for disease positives? Specify:	
	Was it the same for everyone?	Yes □ No □ Unclear □
	What is reference standard for disease negatives? Specify:	
	Was it the same for everyone?	Yes □ No □ Unclear □
Reference Std 2	What is reference standard for disease positives? Specify:	
	Was it the same for everyone?	Yes □ No □ Unclear □
	What is reference standard for disease negatives? Specify:	
	Was it the same for everyone?	Yes □ No □ Unclear □
Incorporation bias	Does reference standard exclude the output of the decision tool?	Yes □ No □ Unclear □
Blinding	Was reference standard allocated blind to decision tool results?	Yes □ No □ Unclear □
	Was the decision tool user blind to the reference standard?	Yes □ No □ Unclear □
Partial verification/ work-up bias	Was there an attempt to compare all results of unaided doctors' decisions or decision tools to reference standard, and vice versa?	Yes □ No □ Unclear □
	Have all results been compared to the same standard? Please explain:	Yes □ No □ Unclear □

Selection of the study sample	a) Consecutive or random selection of cases sampled b) Single relevant clinical population or separate +ve and –ve groups	Consecutive □ Random □ Unclear □ Other: ____________ Single □ Separate □ Unclear □
Subgroups	Were subgroups analysed separately? (e.g. children vs adults) Were subgroups prospectively defined? If Yes then check as appropriate: Gender □ Age □ Healthcare setting □ Type of admission (A&E, ward admission or surgical clinic) □ Country □ Type of decision tool □ Other: ____________	Yes □ No □ Unclear □ Yes □ No □ Unclear □
Completeness State numbers not percentages	No. of patients originally considered for inclusion No. of patients eligible No. of patients included at start No. of patients with full test results No. of patients lost No. of patients included in analysis Reason for exclusion from analysis: Specify: ____________	Stated □ Not stated □
Indeterminate outputs	How were indeterminate scores/outputs of DTs handled in analysis? Check as appropriate: Excluded □ Treated as positive □ Treated as negative □ Sensitivity analysis □ Other □ Unclear □	
Treatment paradox	Were patients treated for their AAP before DTs were used? If so then explain: ____________	Yes □ No □ Unclear □
Type of study	What type of study design was used? Retrospective or prospective data?	Cohort □ Case-control □ Unclear □ Other □ specify____________ Retrospective □ Prospective □ Unclear □
COMMENTS		

Quality assessment form for study question 2 (impact studies)

Paper ID	**Reviewer ID**	**Date**
Allocation	What was the reported method of allocating decision tool(s)? Describe method:	Randomisation □ Quasi-randomisation □ Unclear □ Other □
Allocation concealment and contamination	What is the allocation unit? If patient, could investigators have been aware of allocation of each patient before trial entry?	Patient □ Doctor □ Team □ Other ________________ Yes □ No □ Unclear □
Unit of analysis error	In cluster trial, did analysis take clustering into account?	Yes □ No □ Unclear □ NA □
Analysis by "intention to provide or communicate information" State numbers not percentages	No. of patients originally considered for inclusion No. of patients eligible No. of patients included at start No. of patients with data available No. of patients lost No. of patients included in analysis Reason(s) for exclusion from analysis: Specify:	Grp 1 Grp2 Grp 3 Grp 4 Stated □ Not stated □
Blinding	Was assessment of outcome 1 (specify _______________) blind to use of DT? Was assessment of outcome 2 (specify _______________) blind to use of DT? Was assessment of outcome 3 (specify _______________) blind to use of DT? Was assessment of outcome 4 (specify _______________) blind to use of DT?	Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □ Yes □ No □ Unclear □
Ceiling/floor effects	What was the baseline performance %?	Outcome 1: ____________ Performance: _________ % Outcome 2: ____________ Performance: _________ % Outcome 3: ____________ Performance: _________ % Outcome 4: ____________ Performance: _________ %

		Grp 1	Grp 2	Grp 3	Grp 4
Co-interventions (4 if yes, 6 if no)	Printed materials e.g. cue sheet	□	□	□	□
	Checklist (form with spaces for data)	□	□	□	□
	Education or seminar	□	□	□	□
	Algorithm, nomogram in addition to DT	□	□	□	□
	Access to extra advice from other doctors	□	□	□	□
	Outreach visit	□	□	□	□
	Audit and feedback	□	□	□	□
	Other: __________________________				
Other	Was there a sample size calculation?	Yes □ No □ Unclear □			

Study details for review questions 1 (accuracy studies) and 2 (impact studies)

Paper ID	**Reviewer ID**	**Date**
Study details		
Type of study	Accuracy □ Impact □ Both □	
Healthcare setting	Ward □ Surgical dept □ A&E □ Unspecified or other secondary care □ Primary care □	
If secondary care, academic status of hospital	University or teaching hospital □ Non-teaching nor university affiliated □ Some of both □ Not clear □ Multicentre □	
Decision tool user	Doctor □ nurse □ researcher □ Patient □ other ____________ If decision tool user is doctor, fill in numbers for the following: *Doctor's seniority:* Consultant (Attending) ____________ Registrar (Chief Resident) ____________ SHO (Resident) ____________ HO (Intern) ____________ *Doctor's training:* FRCS or Board certified ____________ Not FRCS nor Board certified ____________	
Healthcare system	Fee for service □ HMO □ insurance □ NHS model □	
Country	UK □ US □ Germany □ other ____________	
Medical condition(s) causing AAP studied (% and no. of patients)	acute appendicitis ____________ acute cholecystitis ____________ small bowel obstruction ____________ gynaecological ____________ acute pancreatitis ____________ renal colic ____________ perforated peptic ulcer ____________ cancer ____________ diverticular disease ____________ dyspepsia ____________ NSAP ____________ other (specify) ____________	
Patients' characteristics (as reported in paper)	*Age:* ____________ *Gender* (nos): Male ________ Female ________ *Ethnicity* (nos and %s): *Mean duration of AAP* ____________	
Were non-operated, non-admitted and discharged cases followed up?	*Non-operated* Yes □ No □ Unclear □ Method ____________ *Non-admitted* Yes □ No □ Unclear □ Method ____________ *Discharged* Yes □ No □ Unclear □ Method ____________	

Description of decision tool(s) studied (check boxes as appropriate)	checklist □ slide rule □ scoring system □ prognostic model □ cue sheet □ expert system □ nomogram □ clinical algorithm □ pre-prog. calculator □ practice guideline □ decision support system □ advisory system □ other (specify) ____________
Name of decision tool	
How was decision tool developed?	Univariate □ Multivariate □ Unclear □ State method ____________
Method of presenting results from decision tool	Probability as % □ Graph □ Prose report □ Probability 0 to 1 □ Raw score □ Other ____________
Purpose and timing of use	Education □ Clinical audit □ Immediate decision-making □ Research on use as decision-making aid □ other ____________
Was evaluator of tool also its developer?	Yes □ No □ Unclear □ Not Applicable □
Describe diagnostic work-up	Describe:

Comments on DT

Year in which study was performed ____________

No. of patients undergoing laparotomy ____________

No. of patients undergoing laparoscopic surgery ____________

No. of patients undergoing peritoneal lavage ____________

Rate of usage of X-rays ____________ %/no. of patients

Rate of usage of C-reactive protein ____________ %/no. of patients

Rate of usage of nuclear scan ____________ %/no. of patients

Rate of usage of ultrasound ____________ %/no. of patients

Rate of usage of CT scan ____________ %/no. of patients

Rate of usage of computed tomography or MRI ____________ %/no. of patients

Rate of usage of blood tests ____________ %/no. of patients

Rate of usage of urine tests ____________ %/no. of patients

Rate of usage of other special investigations ____________ %/no. of patients

Referral rate to trainee/qualified senior surgeons ____________ %/no. of patients

Did study use training set and test set data?	Training set □ Test set □ Unclear □ **For training-set data, fill in the following:** *Design:* consecutive □ random □ retrospective □ unclear □ other ______________ *Target decision:* Diagnosis of AAP □ Diagnosis of Acute Appendicitis □ Other: ______________________________ *Sample characteristics:* Age ____________________ Gender (nos): male ______ female ______ Sample size ______________ *Setting:* Ward □ Surgical dept □ A&E □ Other/unspecified secondary care □ Primary care □ *Casemix* (% or number of patients as appropriate): acute appendicitis __________ acute cholecystitis __________ small bowel obstruction ________ gynaecological __________ acute pancreatitis __________ renal colic __________ perforated peptic ulcer ________ cancer __________ diverticular disease __________ dyspepsia __________ NSAP ______________ other (specify) ____________________ **For <u>test-set</u> data, fill in the following:** Properties of test data set: split sample □ jack-knife □ new prospective data □ unclear □ other □ (describe): ______________________________ Test centre: same as for training set □ new centre □ unclear □ Decision tested: ______________________________
GENERAL COMMENTS	

Paper no.	**Reviewer ID** ________________ **Date** ________________

Results from studies of accuracy

Measure(s): Decision / Action / Outcome / Healthcare resources

Whole Training set sample / Whole Test set sample / Training set subgroup / Test set subgroup (circle one):

If subgroup, describe: ____

Definitions of Positive and Negative Results:

		Reference Standard	
		positive	negative
Decision Tool / Unaided Doctors (choose one)	positive		
DT name:	negative		

Notes: (What other information is available? Where from?)

Needs statistical check □

Measure(s): Decision / Action / Outcome / Healthcare resources

Whole Training set sample / Whole Test set sample / Training set subgroup / Test set subgroup (circle one):

If subgroup, describe: ____

Definitions of Positive and Negative Results:

		Reference Standard	
		positive	negative
Decision Tool / Unaided Doctors (choose one)	positive		
DT name:	negative		

Notes: (What other information is available? Where from?)

Needs statistical check □

Measure(s): Decision / Action / Outcome / Healthcare resources

Whole Training set sample / Whole Test set sample / Training set subgroup / Test set subgroup (circle one):

If subgroup, describe: ____

Definitions of Positive and Negative Results:

		Reference Standard	
		positive	negative
Decision Tool / Unaided Doctors (choose one)	positive		
DT name:	negative		

Notes: (What other information is available? Where from?)

Needs statistical check □

Copy this page if further results needed. Please attach original paper with area highlighted where results have been extracted.
If 2 × 2 not easy or possible, then mark for statistical review.

Paper no.	**Reviewer ID** ________________ **Date** ________________

Results from impact studies

Comparison no. ____________

Groups compared (use labelling in coding instructions sheet):

Describe comparison (e.g. decision tool vs. no intervention)

Outcome no. ____________ Type of outcome: Decision / Action / Outcome / Resources

Describe outcome measure: ____

Results for outcome measure in natural units (report intervention group first):

	Baseline period		Post-intervention period	
Group	No. with event	Total observed	No. with event	Total observed
(intervention)				
(control)				

Total observed: no. of cases in group who were completely monitored for that outcome.
No. with event: no. of cases in group in which specified outcome occurred. For continuous variables, mean and standard deviation.

If data for 2 × 2 table are not available but the following effect measure(s) are reported, then fill in:

Risk difference: ______________ Relative risk: ______________ Odds ratio: ______________

Statistical significance: ______________ Reported by author or calculated by reviewer?

If unit of analysis error, was appropriate adjustment made (e.g. measure of intra-cluster correlation): Yes / No

Statistical test used: ______________ Comments (e.g. one / two-tailed test):

Further comments:

Outcome no. ______ Type of outcome: Decision / Action / Outcome / Resources

Describe outcome measure: ______________________________________

Results in natural units (report intervention group first):

	Baseline period		Post-intervention period	
Group	No. with event	Total observed	No. with event	Total observed
(intervention)				
(control)				

Total observed: no. of cases in group who were completely monitored for that outcome.
No. with event: no. of cases in group in which specified outcome occurred. For continuous variables, mean and standard deviation.

If data for 2 × 2 table are not available but the following effect measure(s) are reported, then fill in:

Risk difference: ______________ Relative risk: ______________ Odds ratio: ______________

Statistical significance: ______________ Reported by author or calculated by reviewer?

If unit of analysis error, was appropriate adjustment made (e.g. measure of intra-cluster correlation): Yes / No

Statistical test used: ________________ Comments (e.g. one / two-tailed test):

Further comments:

Copy this page if further results needed. Please attach original paper with area highlighted where results have been extracted.

Appendix 5

Definition of quality indicators for accuracy review

Theoretically speaking, the best reference standard for those with the target disorder is the histopathological examination of the excised appendix (the appendix for this monograph, since the focus is on acute appendicitis) for both those with and those without the target disorder. This is unethical, so the quality of the reference standards used in the included studies was classified as follows:

1. histopathology for those with the target disorder and final diagnosis for those without, with postdischarge follow-up of the latter: *good*
2. histopathology for those with the target disorder and those without: *fair*
3. final diagnosis for both those with and those without the target disorder (with standard criteria): *fair*
4. final diagnosis for both those with and those without the target disorder (not standardised or unclear criteria): *poor*.

Reference standard (1) above is probably as close to approximating a gold standard as possible for the purposes of this study. The follow-up of non-admitted discharged patients is important, as patients may appear to be healthy or free of a serious cause of AAP when they are not.

The potential for incorporation bias was classified as follows:

1. the decision tool (DT) or a component of the DT is part of reference standard for both those with and without the target disorder: *high*
2. the DT or a component of the DT is part of reference standard for those with the target disorder only: *moderate*
3. the DT or a component of the DT is part of reference standard for those without the target disorder only: *moderate*
4. the DT or a component of the DT is not part of reference standard for the target disorder: *none*.

The potential for differential verification bias was classified as moderate if all decision tool results were compared with reference standard 1 above, since it most closely approximates a gold standard for acute appendicitis, and high if decision tool results were compared with the other types of reference standard.

Other quality indicators included in the metaregression analysis included potential for partial verification bias (whether reference standard observations were obtained from all patients in a study), completeness of data (proportion of eligible patients who actually participated in the study), type of study (cohort versus retrospective), treatment paradox (whether patient was treated for target condition after diagnosis by decision tool, but before testing using the reference standard) and method of dealing with indeterminate output or missing data (e.g. excluded, treated as positive, treated as negative or sensitivity analysis).

Appendix 6

Studies included in the review of accuracy studies

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Alvarado, 1986,[58] USA	Weighted scoring system (Alvarado score)	*Training set* Consecutive recruitment Retrospective cohort Surgical department admission 145 males 126 females (six records did not report gender) Mean age 25.3 (SD 15.9) years 81.9% acute appendicitis (227/277) *n* = 277 *Test set* No test set	Sensitivity Specificity	– –	93% (211/227) 52% (26/50)	Cut-off ≥6 appendicitis Cut-off<6 not appendicitis
Bond, 1990,[12] USA	Weighted scoring system (Alvarado score)	*Training set* See Alvarado, 1986[58] *Test set* Consecutive recruitment Prospective cohort A&E admission Gender distribution not reported Children: age range 2–17 years 61.4% appendicitis (116/189) *n* = 189	Sensitivity Specificity	– –	90% (104/116) 71% (52/73)	Cutoff ≥7 appendicitis Cutoff<7 not appendicitis
Bond, 1990,[112] USA	Discrimination rule (using discriminant analysis)	*Training set* Consecutive recruitment Prospective cohort A&E admission Gender distribution not reported Children: age range 2–17 years 61.4% appendicitis (116/189) *n* = 189 *Test set* No test set	Sensitivity Specificity	– –	88% (102/116) 75% (55/73)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
De Dombal, 1972,[113] UK	Bayesian (Leeds)	*Training set:* See Horrocks, 1972[138] for training-set study (1050 patient records obtained retrospectively and 884 patients recruited prospectively), which reported crude accuracies only and insufficient data for sensitivity and specificity to be calculated See Staniland *et al*., 1972[277] which describes database used to train Leeds AAP system (n = 600 patients) *Test set:* Consecutive recruitment Prospective cohort Surgical department admission Age and gender distribution not reported 28% appendicitis (85/304) n = 304	Sensitivity Specificity	88% (75/85) 86% (188/219) Unaided senior clinicians	99% (84/85) 96% (208/217)	For unaided doctors, “occasionally two registrars saw case simultaneously – Both diagnoses then being entered”. This explains n = 322 For computer system, in two cases, computer could not compute result because clinician could not decide on data. Hence n = 302
Edwards, 1986,[114] UK	Bayesian (Sheffield)	*Training set* Authors mentioned use of a training set, but did not give any information or reference *Test set:* Representative sample of patients (“well over 95%” of all admitted AAP patients) Prospective cohort Surgical department admission Children: age and gender distribution not reported 33% appendicitis (114/344) n = 344	Sensitivity Specificity	80% (91/114) 78% (179/230) Junior doctors and medical students	81% (92/114) 78% (180/230)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Eskelinen, 1992,[115] Finland	Logistic regression	*Training set*				
		Representative sample of patients	Sensitivity	93% (251/270)	88% (237/270)	
		Prospective cohort	Specificity	86% (914/1063)	88% (935/1063)	
		Secondary care (unspecified type)	LR+	6.64	7.33	
		Mean age 38 (SD 22.1) years	LR-	0.08	0.14	
		636 males, 697 females	PV+	62%	69%	
		20.3% appendicitis (270/1333)	PV–	98%	96%	
		n = 1333			(Cut-off score 55.0)	
		Test set No test set				
Fenyo, 1987,[116] Sweden	Logistic regression	*Training set*				
		Criteria for inclusion unclear	Sensitivity	–	–	
		Case–control study	Specificity	–	–	
		Surgical department admission				
		Age and gender distribution not reported				
		42% appendicitis				
		n = 259				
		Test set				
		Consecutive recruitment	Sensitivity	–	90% (231/256)	
		Prospective cohort	Specificity	–	91% (525/574)	
		Surgical department admission				
		Median age 24 (range 15–86) years				
		358 males, 472 females				
		30.8% appendicitis (256/830)				
		n = 830				
		45.8% appendicitis (164/358)	Sensitivity	–	86% (141/164)	
			Specificity	–	94% (182/194) (Men)	
		19.5% appendicitis (92/472)	Sensitivity	–	95% (87/92)	
			Specificity	–	83% (315/380) (Women)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Hallan, 1997,[117] Norway	Weighted scoring system (modified Alvarado Score)	*Training set* See Alvarado, 1986[58] *Test set* Consecutive recruitment Prospective cohort Surgical ward admission (GP referred) Median age 22 (range 3–86) years 142 males, 115 females 38.1% appendicitis (98/257) *n* = 257	Sensitivity Specificity	– –	63.3% (62/98) 81.1% (129/159)	Random split sample 20 times, derived tool 20 time averages Cut-offs for Alvarado score ≤6
Hallan, 1997;[117] Hallan, 1997,[118] Norway	Logistic regression	*Training set* Consecutive recruitment Prospective cohort Surgical ward admission (GP referred) Median age 22 (range 3–86) years 142 males, 115 females 38.1% appendicitis (98/257) *n* = 257 *Test set* No test-set validation reported	Sensitivity Specificity AUROC curve	– – 0.809 (95% CI: 0.793 to 0.824)	85% (83/98) 84% (134/159) 0.813 (95% CI 0.797 to 0.829)	Random split sample 20 times, derived tool 20 time averages Authors concluded that "surgeons' probability estimates are as good as those of an acceptable computer model"
Horrocks, 1976,[119] Canada	Bayesian (Leeds)	*Training set* See Horrocks, 1972[138] for training-set study, which did not report estimates of sensitivity or specificity See Staniland *et al*., 1972[277] which describes a database used to train, Leeds AAP system (*n* = 600 patients)	Sensitivity Specificity	– –	– –	
		Test set "Series of unselected real-life patients" Retrospective cohort Secondary care (unspecified type) Age and gender distribution not reported 32.7% appendicitis (34/104) *n* = 104	Sensitivity Specificity	– –	91% (31/34) 99% (69/70)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Izbicki, 1992,[120] Germany	Discrimination rule	*Training set* Consecutive recruitment Prospective cohort Emergency ward admission Mean age 31.5 (range 11–88) years 66 males, 84 females 36% appendicitis (54/150) n=150 *Test set* No test-set validation study reported	Sensitivity Specificity	– –	89% (48/54) 54% (52/96)	Score cut-off >2
Jahn, 1997,[121] Denmark	Discrimination rule	*Training set* Consecutive recruitment Prospective cohort Surgical department admission Mean age 27 (SD 18 years), range 4–98 years 118 males, 104 females 42% acute appendicitis (94/222) *n* = 222 *Test set* No test-set validation study	Sensitivity Specificity	64% (60/94) 58% (74/128)	63% (59/94) 94% (120/128)	Score cut-off +16 ≥+16 appendicitis <+16 no appendicitis

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Jawaid, 1999,[122] Pakistan	Logistic regression	*Training set* Criteria for inclusion unclear Retrospective cohort of patients who had undergone appendicectomy Surgical ward admission All patients were older than 15 years (age distribution not reported) 270 males, 131 females 87.5% acute appendicitis (351/401) *n* = 401	Sensitivity Specificity	– –	Excluded Excluded Training-set data excluded (see comments)	Performance data for training-set study excluded from systematic review because patients were not unselected. Same score cut-off was used for training set and test-set studies.
		Test set Consecutive recruitment Prospective cohort of patients Surgical ward admission Age and gender distribution not reported 81.8% acute appendicitis (81/99) *n* = 99	Sensitivity Specificity	– –	78% (274/351) 88% (44/50)	
Kirkeby, 1987,[123] Norway	Bayesian (Leeds)	*Training set* See Horrocks, 1972[138] and Staniland, 1972[277] for details of training-set study, which did not report accuracy data				Assumed authors followed Leeds criteria for reference standard
		Test set Consecutive recruitment Prospective cohort Surgical department admission Small county hospital Age and gender distribution not reported 23% appendicitis (18/77) *n* = 77	Sensitivity Specificity	78% (14/18) 39% (23/59)	72% (13/18) 53% (31/59)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Kraemer, 1993[124] Germany	Bayesian (Leeds)	*Training set* See Horrocks, 1972,[138] and Staniland, 1972,[277] for details of training-set study, which did not report accuracy data *Test set* Consecutive recruitment Prospective cohort Multicentre surgical ward (six surgical departments) 17% appendicitis 510 men, 576 women; 116 children ≤ 14 years (gender breakdown not given) Age distribution of men and women reported by age bands (15–49 years and ≥ 50 years) 17% appendicitis (213/1254) *n* = 1254	Sensitivity Specificity	– –	53% (111/210) 90% (938/1043)	
Leaper, 1972,[125] UK	Bayesian (Leeds)	*Training set* See Horrocks, 1972,[138] and Staniland, 1972,[277] for details of training-set study, which did not report accuracy data *Test set* Consecutive recruitment Prospective cohort Surgical department admission Age and gender distribution not reported 26% appendicitis (122/472) *n* = 472	Sensitivity Specificity	87% (106/122) 87% (303/350) Senior doctors	99% (120/121) 96% (333/347)	Four cases excluded in the computer diagnosis because clinicians unable to decide on what data to input

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Leaper, 1972,[125] UK	Bayesian (Leeds)	*Training set* Six clinicians' "personal estimates" of diagnoses and other information when presented with profiles of data from actual AAP patients'				Leeds AAP system but used database of doctors' "personal estimates" of AAP patients' diagnoses
		Test set Consecutive recruitment Prospective cohort Surgical department admission Age and gender distribution not reported 25% appendicitis (121/487) *n* = 487	Sensitivity Specificity	– –	92% (111/121) 95% (347/366)	
Lindberg and Fenyo, 1988[126] Sweden	Logistic regression	*Training set* Consecutive recruitment Prospective cohort Surgical department admission Age and gender distribution not reported 29% appendicitis (219/746) *n* = 746	Sensitivity Specificity	– –	No accuracy data reported	
		Test set Consecutive recruitment Prospective cohort Surgical department admission Age and gender distribution not reported 28% appendicitis (27/96) *n* = 96	Sensitivity Specificity	– –	89% (24/27) 93% (64/69)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Macklin, 1997,[127] UK	Weighted scoring system (modified Alvarado Score)	*Training set* See Alvarado 1986[58] *Test set* Consecutive recruitment Prospective cohort Surgical department Children: median age 10 years (range 4–14) years 54 boys, 64 girls 20% appendicitis (38/188) *n*=118	Sensitivity Specificity	– –	76% (29/38) 79% (63/80) (All patients)	Modified score: omitted 'left shift of neutrophil maturation', so maximum score of 9 instead of 10 Cut-off at 7: ≥7 appendicitis <7 not appendicitis
		20% appendicitis (20/54)	Sensitivity Specificity	– –	80% (16/20) 82% (28/34) (Boys)	As above
		28% appendicitis (18/64)	Sensitivity Specificity	– –	72% (13/18) 76% (35/46) (Girls)	As above
Malik, 1998,[128] India	Weighted scoring system (modified Alvarado Score)	*Training set* See Alvarado, 1986[58] *Test set* Consecutive recruitment Prospective cohort Surgical department Age range not reported 53 men, 41 women, 12 children 78% appendicitis (83/106) *n* = 106	Sensitivity Specificity	– –	83.1% (69/83) 30.4% (7/23)	
		87% appendicitis (46/53)	Sensitivity Specificity	– –	78% (36/46) 57% (4/7) (Men)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
		66% appendicitis (27/41)	Sensitivity Specificity	– –	89% (24/27) 86% (12/14) (Women)	
		83% appendicitis (10/12)	Sensitivity Specificity		90% (9/10) 50% (1/2) (Children)	
Ohmann, 1999,[129] Germany and Austria	Discrimination rule	*Training set* See Ohmann 1995 (developed using stepwise linear regression of data from prospective German study with 1254 patients)				Aided doctors' diagnostic accuracies were compared with unaided doctors' accuracies
		Test set Consecutive recruitment Prospective cohort Surgical department admission Age and gender distribution not reported 21.2% appendicitis (315/1484) *n* = 1484	Sensitivity Specificity	91.5% (174/190) 86.4 (552/639)	95.5% (107/112) 78.1% (379/485)	
Owen, 1992,[130] UK	Weighted scoring system (modified Alvarado score)	*Training set* See Alvarado, 1986[58]				
		Test set Consecutive recruitment Prospective cohort Surgical department admission Age distribution not reported 75 men, 70 women, 70 children (no gender breakdown for children) 58% appendicitis (124/215) *n* = 215	Sensitivity Specificity	– –	93% (115/124) 87% (79/91)	
		64% appendicitis (48/75))	Sensitivity Specificity	– –	96% (46/48) 96% (26/27) (Men)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
		47% appendicitis (33/70)	Sensitivity Specificity	– –	94% (31/33) 78% (29/37) (Women)	
		61% appendicitis (43/70)	Sensitivity Specificity	– –	88% (38/43) 89% (24/27) (Children)	
Pesonen, 1996,[131] Finland	Neural networks: 1. Binary adaptive resonance theory (ART1) 2. Kohanen self-organising map (SOM) 3. Learning vector Quantisation (LVQ) 4. Back propagation (BP)	*Training set* Representative sample of patients (as claimed by authors) Prospective cohort Random split sample Secondary care (unspecified type) Age distribution not reported 217 males, 237 females (n = 454) 20% appendicitis (92/454) Prevalence for split sample not reported so prevalence of entire sample used as proxy	Sensitivity Specificity	– –	– –	Four groups of parameters were used to test the neural networks: (1) "Four most important clinical signs in diagnosis of acute appendicitis", (2) "17 clinical signs", (3) 14 clinical history parameters, and (4) all 31 parameters above. Sensitivity and specificity are reported for group 1 only, since from the point of view of usability, using four clinical signs is better than 31 Unaided doctors were senior surgeons
		Test set Representative sample of patients (as claimed by authors) Prospective cohort Random split sample Secondary care (unspecified type) Age distribution not reported 218 males 239 females 20% acute appendicitis (92/457) n = 457	ART1: Sensitivity Specificity SOM: Sensitivity Specificity LVQ: Sensitivity Specificity BP: Sensitivity Specificity	 93% (86/92) 84% (307/365) 93% (86/92) 84% (307/365) 93% (86/92) 84% (307/365) 93% (86/92) 84% (307/365)	 79% (73/92) 88% (321/365) 55% (51/92) 83% (303/365) 87% (80/92) 90% (329/365) 83% (76/92) 92% (336/365)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Pesonen, 1997,[132] Finland	Discrimination rule	*Training set* Consecutive recruitment Prospective cohort Secondary care (unspecified type) Age and gender distribution not reported Random split sample Prevalence of appendicitis not reported *n* = 717 *Test set* Consecutive/prospective Secondary care Random split sample Prevalence of appendicitis not reported *n* = 374	AUROC curve	– –	Training set: 0.92 Test set: 0.94	
Saidi, 2000,[133] Iran	Weighted scoring system (Alvarado score)	*Training set* See Alvarado, 1986[58] *Test set* Consecutive recruitment Prospective cohort Emergency ward admission Age range not specified 49 males, 79 females, of whom 20 were children (age range for children not specified) 35% appendicitis (45/128) for adults *n* = 128	Sensitivity Specificity	– –	94% (45/48) 88% (70/80)	Whole sample Cut-off ≥6
			Sensitivity Specificity	– –	96% (25/26) 96% (22/23)	Men Cut-off ≥6
			Sensitivity Specificity	– –	90% (17/19) 80% (48/60)	Women Cut-off ≥6
			Sensitivity Specificity	– –	83% (5/6) 93% (13/14)	Children Cut-off ≥6

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Staniland, 1980,[134] Norway	Bayesian (Leeds)	*Training set* See Horrocks, 1972,[138] and Staniland, 1972,[277] for details of training-set study, which did not report accuracy data *Test set* Representative sample of patients (as claimed by authors; see comment) Prospective cohort Surgical ward admission Age and gender distribution not reported 32.5% appendicitis (102/313) *n* = 313	Sensitivity Specificity	– –	81% (83/102) 78% (164/211)	Authors "reprogrammed [computer] with data from Norwegian patients and re-analysed" the data. Crude accuracy rates went up, but not enough data were provided in the paper to calculate sensitivity, specificity, etc., for this comparison
Sutton, 1989,[135] Scotland	Bayesian (CADA)	*Training set for CADA* see Gunn, 1976[278] *Test set for CADA* Consecutive recruitment Prospective cohort Secondary care (unspecified type) Age and gender distribution not reported 15.5% appendicitis (777/4998) *n* = 4998	Sensitivity Specificity	81% (442/547) 90% (4028/4451)	81% (444/547) 88% (3896/4451)	
Wellwood, 1992,[136] UK	Bayesian (Leeds)	*Training set* See Horrocks, 1972[138] and Staniland, 1972[277] for details of training-set study, which did not report accuracy data *Test set* Consecutive recruitment Prospective cohort Surgical ward admission Age and gender distribution not reported 6% appendicitis (324/5193) *n* = 5193	Sensitivity Specificity	79% (103/130) 92% (1706/1855)	63% (82/130) 95% (1757/1855)	

continued

Study, country	Decision tool: reasoning method (name of tool, if any)	Study design, setting and sample characteristics (Training set and test set)	Metrics	Results		Comments
				Doctors	Decision tools	
Wong, 1990;[279] Wong, 1992;[63] Wong, 1994[137] Hong Kong	Logistic regression	*Training set* Random sample of patients selected Prospective cohort Surgical department admission Age distribution reported in detail for entire sample and for male and females (in 10-year age bands from 10–19 ≥70 years) 149 males, 217 females 42% appendicitis *n* = 397 *Test set* Not reported	Sensitivity Specificity	– –	88% (64/73) 68% (221/324)	

LR, likelihood ratio; PV, predictive value.

Appendix 7

Summary table of excluded studies in the review of accuracy studies

Study, country	Reasons for exclusion
Arnbjörnsson, 1985,[280] Sweden	Only patients who had undergone appendicectomy were included, therefore not unselected patients
Anatol, 1995,[281] Trinidad	No appropriate results were reported (only crude accuracy)
Bjerregaard, 1983,[282]	No appropriate results were reported (kappas only)
Blazadonakis, 1996,[283] Greece	No appropriate results were reported (only crude accuracy)
Bohner, 1998,[284] Germany	Focus was on bowel obstruction as cause of AAP, not acute appendicitis
Browder, 1989,[285] USA	No appropriate results
Crossley, 1982,[286] UK	Excluded because Rutter scale was used to diagnose emotional disturbance. Study examined association between appendicitis and emotional disturbance. No appropriate results were presented
Davenport, 1985,[287] UK	Included patients with AAP for more than 7 days. Focus was on dyspepsia, not acute appendicitis. No new data were reported
De Dombal, 1971,[288] UK	No analysable results were presented
De Dombal, 1978,[289] UK	Selected sample: only patients finally classified as appendicitis or NSAP were included
De Dombal, 1980,[290] UK	No new data: review paper and commentary of previous studies by De Dombal's group
De Dombal, 1984[291] UK	No new data were reported: review of work done by De Dombal's group up to 1982
DeDombal, 1992,[292] UK	No relevant results calculable: only crude accuracies presented
Dickson, 1985[293] UK	No relevant results calculable: only crude accuracies presented
Edwards, 1984,[294] USA	Patients were not recruited in a consecutive and unselected manner: discharged patients were excluded
Eskelinen, 1992,[115] Finland	No decision tool was reported, only individual clinical signs
Eskelinen, 1994,[295] Finland	Study focused on scoring system for small bowel obstruction, not acute appendicitis
Eskelinen, 1994,[296] Finland	Duplicate of Eskelinen, 1992[115]
Eskelinen, 1994,[39] Finland	Duplicate of Eskelinen, 1992[115]
Eskelinen, 1995,[40] Finland	Duplicate of Eskelinen, 1992[115]
Fathi-Torbaghan, 1994,[297] Germany	No appropriate results: crude accuracy only. Unclear how patients were recruited
Fenyo, 1987,[166] Sweden	Sensitivity and specificity not calculable from data presented
Fraser, 1992,[32] UK	Study was on NSAP, not acute appendicitis
Gallego, 1998,[298] Spain	Selected sample: "Patients in whom ultrasonography could not be performed and those with a clear diagnosis of acute appendicitis were excluded"
Gough, 1988,[174] Australia	Not AAP less than 7 days
Graff, 2000,[299] USA	Objective of study was not evaluation of accuracy or impact of decision tool, although Alvarado scores were reported
Graham, 1977,[300] UK	Selected sample (case–control study): patients with proven appendicitis compared with those with perforated appendicitis
Graham, 1979,[301] UK	Not AAP less than 7 days

continued

Study, country	Reasons for exclusion
Gunn, 1976,[278] UK	No appropriate results: crude accuracy only
Gunn, 1991,[302] UK	No new results: author reported results from other studies. Contains useful information on usage rates and economic data
Horrocks, 1972,[138] UK	No appropriate accuracy data reported (training-set study for Leeds AAP system)
Ikonen, 1983,[303] Finland	Selected sample. Inappropriate results
Jawaid, 1999,[122] Pakistan	Training-set study excluded because accuracy data were obtained from selected patients (retrospective cohort of patients who had undergone appendicectomy)
Lawrence, 1987,[167] UK	No appropriate results: crude accuracy only
McAdam, 1990,[304] UK	No appropriate results: crude accuracy only
Ohmann, 1996,[305] Germany	No appropriate results: crude accuracy only
Orient, 1986,[306] USA	Focus was on NSAP, not acute appendicitis
Paterson-Brown, 1989,[307] UK	No appropriate results
Pesonen, 1998,[198] Finland	Duplicate of Pesonen, 1996[131]
Pesonen, 1994,[308] Finland	No decision tool was evaluated: described parameters in detail for database used in other papers by Eskelinen and Pesonen. Also discussed difficulties in distinguishing acute appendicitis from NSAP
Puppe, 1995,[309] Germany	No appropriate results: crude accuracy only
Staniland, 1972,[277] UK	No appropriate results reported. Study described training-set data for Leeds AAP system
Sturman, 1989,[310] America	No appropriate results: crude accuracy only. Note: this is an expert-assigned score. The authors "developed and tested several novel diagnostic algorithms to encode knowledge based information about each sign and symptom in matrix of several thousand numbers that represent relative importance an expert in abdominal pain assigns to each disease for each question asked"
Talwar, 1999,[311] India	No new results
Teicher, 1983,[312] US	Not a study of consecutively recruited and unselected patients (case–control study)
Wade, 1993,[313] USA	Study question inappropriate (although Alvarado score was reported, study assessed accuracy of ultrasonography)
Wilson, 1975,[171] UK	No calculable results from a 2 × 2 table
Zielke, 1998,[314] Germany	No decision tool reported

Appendix 8

Summary table of decision tool characteristics

Reasoning method (tool name, if any)	Relevant studies	Purpose of DT (Clinical decisions, audit, education; decisions tested; who uses it)	Type of DT (Checklist, algorithm, score, decision support system, etc; medium)	Information used (No. of items used in DT; signs, symptoms, history, biochemical tests (list); time to complete tool)	Reasoning method and development of DT (Bayes, cluster, logistic, neural net, logic; method of development)	Output format (Score, probability, graph, advice, etc.)
Weighted scoring system (Alvarado score)	Alvarado 1986;[58] Hallan, 1997;[117] Macklin, 1997;[127] Malik, 1998;[128] Owen, 1992;[130] Saidi, 2000;[133] Bond, 1990[112]	To use a simple scoring system as an aid to clinical decisions on patients suspected of having acute appendicitis	Scoring system and checklist	Eight items used: three symptoms, three signs and two laboratory tests. *Items* — *Value* Migration — 1 Anorexia–acetone — 1 Nausea–vomiting — 1 Signs: Tenderness in right lower quadrant — 2 Rebound pain — 1 Elevated temperature — 1 Laboratory: Leucocytosis — 2 Shift to the left of neutrophil maturation — 1 **Total score — 10** In some studies, because left shift of neutrophil not available in laboratory, it is omitted and a nine-point modified Alvarado score used 1–4 appendicitis very unlikely 5–6 observe patient 7–8 probably appendicitis 9–10 appendicitis very probable Did not mention time to complete, but authors said score is much simpler than computer methods	Weighted scoring system and checklist Developed using Bayesian methods but designed for use as a weighted scoring system. For each indicant, 2 × 2 table is used to compute sensitivity, specificity, predictive values and probability estimate. Initially 11 indicants were analysed. It was unclear what criteria the author used to select these indicants. Diagnostic weight was computed for each indicant (obtained by dividing total number of patients by number of true-positive or true-negative tests). Eight items were found to be useful diagnostically, with the more important items given a value of 2 and the less important ones given a value of 1	Raw score. As the authors point out, an advantage of the score is that it does not require the use of a computer

continued

Reasoning method (tool name, if any)	Relevant studies	Purpose of DT (Clinical decisions, audit, education; decisions tested; who uses it)	Type of DT (Checklist, algorithm, score, decision support system, etc; medium)	Information used (No. of items used in DT; signs, symptoms, history, biochemical tests (list); time to complete tool)	Reasoning method and development of DT (Bayes, cluster, logistic, neural net, logic; method of development)	Output format (Score, probability, graph, advice, etc.)
Discrimination rule	Bond, 1990[112]	To use discriminant function to discriminate patients with appendicitis from those without	Discriminant analysis	16 items considered, from which seven items were found best to separate patients with and without acute appendicitis	Discriminant analysis using stepwise selection procedure to identify discriminant function with seven statistically significant items that provided the best model	Score
Logistic regression	Eskelinen, 1992[115]	To aid doctors in the diagnosis of acute appendicitis	Scoring system derived from patient data (recorded using predefined structured data collection sheet) and results of special investigations entered into computer	22-item clinical history, 14 clinical signs on physical examination of patients and three special investigations (body temperature, leucocyte count, urine). For full list see Tables 1 and 2 in Eskelinen, 1992.[115] Time to report not given	Stepwise logistic regression of significant factors identified on bivariate analysis. Then diagnostic score obtained by adding up regression coefficients from variables in logistic regression	Score
Logistic regression	Fenyo,1987[116]	To aid doctors in the diagnosis of acute appendicitis in adults	Scoring system derived from patient data using Bayesian weights of evidence as outlined by Spiegelhalter.[76] Score is computed from scoring sheet by user	19 items used in scoring sheet. For full list of items, see Table II in Fenyo's paper. About 1 minute is needed to compute the score	Used database of 259 patients collected for training purposes. Incidence of 19 items (selected from Leeds form of about 36 items) was measured by gender in those with and without appendicitis. Sensitivity (Se), specificity (Sp) and weight of evidence ($10 \times \log_e[Se/(1-Sp)]$) were calculated for each item. Scores were then obtained by combining item-specific weights for men and women and estimated probability of appendicitis	Score

continued

Reasoning method (tool name, if any)	Relevant studies	Purpose of DT (Clinical decisions, audit, education; decisions tested; who uses it)	Type of DT (Checklist, algorithm, score, decision support system, etc; medium)	Information used (No. of items used in DT; signs, symptoms, history, biochemical tests (list); time to complete tool)	Reasoning method and development of DT (Bayes, cluster, logistic, neural net, logic; method of development)	Output format (Score, probability, graph, advice, etc.)
Logistic regression	Hallan, 1997[117] Hallan, 1997[118]	To aid physicians in the diagnosis of acute appendicitis in patients suspected of having the condition	Scoring system	The optimum model used the following six variables (in decreasing order of importance): classic migration of pain, rebound tenderness in right lower quadrant, aggravation of pain by coughing, guarding, nausea, and duration of pain	Used logistic regression to estimate function discriminating acute appendicitis from other diseases. Initial accuracy "tested as follows: weights of function estimated on randomly drawn sample constituting about half the patients; using these weights probability estimates then treated as scores and AUROC curve was calculated. Done 20 times"	Probability treated as score: AUROC curve
Discrimination rule	Izbicki, 1992[120]	To aid physicians in improving their accuracy in preoperative diagnosis of acute appendicitis	Scoring system	Seven items used: gender (female = 0, male = 1), white cell count > 11×10^9 l^{-1} (no = 0, yes = 1), localised guarding (no = 0, yes = 1), rebound tenderness in right lower quadrant (no = 0, yes = 1), shift of pain from epigastrium to right lower quadrant (no = 0, yes = 1), history of complaints <24 hours and recurrent complaints (yes = 0, no = 1)	Seven-item additive score. Authors did not state in text how the seven items were selected, but judging from bivariate analysis in Tables III and IV in their paper, items were included if their prevalences were significantly different between those with acute appendicitis, those not operated on and those with negative appendicectomy, provided cells within table were sufficiently large. The score was obtained by adding up the contribution from the items, each of which was assigned a value of 1 or 0	Score

continued

Reasoning method (tool name, if any)	Relevant studies	Purpose of DT (Clinical decisions, audit, education; decisions tested; who uses it)	Type of DT (Checklist, algorithm, score, decision support system, etc; medium)	Information used (No. of items used in DT; signs, symptoms, history, biochemical tests (list); time to complete tool)	Reasoning method and development of DT (Bayes, cluster, logistic, neural net, logic; method of development)	Output format (Score, probability, graph, advice, etc.)
Discrimination rule	Jahn, 1997[121]	To aid physicians in the diagnosis of acute appendicitis in patients suspected of having the condition	Computer-generated scoring system	Eleven items used: total white cell count, migration of pain to right lower quadrant, gradual onset of pain, increasing intensity of pain, pain aggravated by movement, pain aggravated by coughing, anorexia, vomiting, indirect tenderness (Rovsing's movement), muscle spasm and gender	Bivariate scoring system Scoring system derived from 21 items (not clear how authors selected these items). Bivariate analysis using χ^2 test to distinguish those with appendicitis from those without. Items not significant were discarded. 11 factors found to be significant and used in additive scoring system	Score
Logistic regression	Jawaid, 1999[122]	To aid clinical decision-making in patients with suspected acute appendicitis in author's local setting in Pakistan	Scoring system	Ten items used: gender, location of initial pain (epigastric), migration of pain to right lower quadrant, anorexia, vomiting, fever, guarding, rebound tenderness, leucocytosis and neutrophilia	"Using Bayesian probability the negative and positive weighting for each factor was calculated using following formulae:" +ve weight= 10 × ln[Se/(1-Sp)] –ve weight= 10 x ln[(1-Se)/Sp] "When factor present +ve weight was given and when absent –ve weight assigned. Weights were rounded off to nearest integer, applied to 401 files [in training set] and summated [to get the score]." Score ranged from –83 to +8. The ten items used in the score were those that best distinguished patients with acute appendicitis from those with negative appendicectomy (using χ^2 tests)	Raw score

continued

Reasoning method (tool name, if any)	Relevant studies	Purpose of DT (Clinical decisions, audit, education; decisions tested; who uses it)	Type of DT (Checklist, algorithm, score, decision support system, etc; medium)	Information used (No. of items used in DT; signs, symptoms, history, biochemical tests (list); time to complete tool)	Reasoning method and development of DT (Bayes, cluster, logistic, neural net, logic; method of development)	Output format (Score, probability, graph, advice, etc.)
Bayesian (Leeds)	Adams, 1986;[51] De Dombal, 1972,[113] 1992;[292] Horrocks, 1972,[138] 1976;[119] Leaper, 1972;[125] Kirkeby, 1987;[123] Kraemer, 1993;[124] Staniland, 1980;[134] Wellwood, 1989,[139] 1992;[136] Wilson, 1975[171]	To aid doctors in the diagnosis of conditions causing AAP in routine clinical settings	Decision support system run on various computer platforms from mainframes, Commodore Pet and Apple computers in the early days, to the latest 2002 versions (run on Pentium PCs and palmtops)	From Figure 3 in Horrocks 1972,[138] there were 36 items in data collection form. For full list of signs, symptoms, history and other items used, see figure in Horrocks, 1972.[138] The authors claim in the discussion that "the diagnostic 'turnaround time' from collection of data from the patient to provision of probabilities by the computer was of the order of a few minutes only [mentioned 20 minutes in another section of the paper], perfectly acceptable in the clinical situation." Wellwood and Spiegelhalter, 1989[139] reported that doctors took 3 minutes to complete the forms and 5 minutes to use computer, but did not mention what the overall turnaround time was. Compliance rates in doctors using the tool in some participating centres were low because of need to improve speed and ease of use	The system uses a Bayesian approach After patient information has been entered into computer, the program "compares information for each new patient with data from 6000 patients of known diagnoses using Bayesian probabilistic analysis … The computer program provides an assessment of the chance of the patient under study having each of the diagnoses listed in [a computer print-out]"[139] data stored in the computer on known diagnoses were naturally less extensive in earlier versions. In the original 1972 version, the training set of the system consisted of 600 patients.[125] Original system described in detail by Horrocks, 1972[138]	List of diagnoses with percentage problems

continued

Reasoning method (tool name, if any)	Relevant studies	Purpose of DT (Clinical decisions, audit, education; decisions tested; who uses it)	Type of DT (Checklist, algorithm, score, decision support system, etc; medium)	Information used (No. of items used in DT; signs, symptoms, history, biochemical tests (list); time to complete tool)	Reasoning method and development of DT (Bayes, cluster, logistic, neural net, logic; method of development)	Output format (Score, probability, graph, advice, etc.)
Logistic regression	Lindberg, 1988[126]	To aid doctors in the diagnosis of acute appendicitis using a simple scoring system	Scoring system tabulated in a simple pocket chart	The regression analysis demonstrated that only ten out of 22 indicants from a structured questionnaire were significant at the $p < 0.01$ level and included in the score The ten indicants were: gender, white blood cell count, duration of pain, progression of pain, relocation of pain, vomiting, aggravation with coughing, rebound tenderness, rigidity and pain outside the right lower quadrant. The authors said that an "easy-to-use algorithm … has been developed", but did not mention time to complete	Bayesian model that does not assume independence between variables The authors expressed "Bayes' theorem as additive model using logarithms of likelihood ratios and to subject this model to a logistic regression analysis." "For each symptom and sign a score reflecting its weight of evidence for diagnosis of appendicitis calculated from the likelihood ratio … These 'raw' scores were then subjected to a stepwise logistic regression analysis with acute appendicitis as dependent variable. The resulting β-values served to adjust the raw score"	Score
Discrimination rule	Ohmann, 1999[129]	To aid doctors in the diagnosis of acute appendicitis	16-point scoring system derived from patient data	Eight items used: tenderness in right lower quadrant (4.5 points), rebound tenderness (2.5 points), no micturition difficulties (2.0 points), steady pain (2.0 points), leucocyte counts (1.5 points), age <50 years (1.5 points), relocation of pain to right lower quadrant (1.0 points) and rigidity (1.0 points). Did not report time to complete tool	Stepwise logistic regression on a prospective German database (n = 1254) to identify statistically significant factors and independent evaluation on a Dutch database (n = 1346)	Score

continued

Reasoning method (tool name, if any)	Relevant studies	Purpose of DT (Clinical decisions, audit, education; decisions tested; who uses it)	Type of DT (Checklist, algorithm, score, decision support system, etc; medium)	Information used (No. of items used in DT; signs, symptoms, history, biochemical tests (list); time to complete tool)	Reasoning method and development of DT (Bayes, cluster, logistic, neural net, logic; method of development)	Output format (Score, probability, graph, advice, etc.)
Discrimination rule	Pesonen, 1997[132]	To aid doctors in the diagnosis of acute appendicitis	Statistical model	Nine items used: colour, location of tenderness, scar, mass, rebound, guarding, rigidity, distension and leucocyte count. Did not mention time to complete	Logistic regression model Data randomly split into two samples: training set with 717 patients and test set with 347 patients. Training set used to develop model and test set used to evaluate accuracy	Only reported AUROC curve. Did not report data that could be used to compute sensitivity and specificity
Neural networks	Pesonen, 1996,[131] 1997,[132] 1998[198]	To aid doctors in the diagnosis of acute appendicitis	Neural networks: four self-learning networks were tested: ART1, SOM, LVQ and BP	17 items used: age, gender, mood, colour, location of tenderness, scar, distension, abdominal movement, mass, rebound, guarding, rigidity, Murphy's positive, bowel sounds, renal tenderness, rectal digital tenderness, leucocyte count. Four groups of parameters used in Pesonen, 1996 study:[131] (1) Four most important clinical signs, (2) 17 clinical signs, (3) 14 clinical history parameters and (4) 31 clinical signs and clinical history. Did not mention time to complete	Neural network Data randomly split into two samples: training set with 717 patients and test set with 347 patients. Training set used to develop model and test set used to evaluate accuracy	
Bayesian (Sheffield)	Edwards, 1986[114]	To aid doctors in the diagnosis of acute appendicitis in children	Decision support system. Program's "written in BASIC and run on a BBC B microcomputer"	35 signs and symptoms; did not specify what these signs and symptoms were. Patient data were recorded by junior doctors or medical students onto forms and then entered into the computer. Did not report time to complete tool	"Program used the prior Bayesian probabilities for about 35 signs and symptoms derived from a different population to classify children with 'acute abdomens' into one of two categories of appendicitis or non-specific abdominal pain (NSAP)". Authors did not give any information or references on this population (presumably the training-set sample)	Probabilities of appendicitis and NSAP summarised as graphs

continued

Reasoning method (tool name, if any)	**Relevant studies**	**Purpose of DT (Clinical decisions, audit, education; decisions tested; who uses it)**	**Type of DT (Checklist, algorithm, score, decision support system, etc; medium)**	**Information used (No. of items used in DT; signs, symptoms, history, biochemical tests (list); time to complete tool)**	**Reasoning method and development of DT (Bayes, cluster, logistic, neural net, logic; method of development)**	**Output format (Score, probability, graph, advice, etc.)**
Bayesian (CADA)	Sutton, 1989;[135] Gunn, 1976[278]	To aid doctors in the diagnosis of acute abdominal pain	Decision support system. BASIC program run on a Cromenco Z2-D or Z2-H microcomputer	Did not mention number of items used or the type of items included	Bayesian approach Author mentioned that development and methods were described in another paper on the Lothian CADA study, but the paper referred to was by Gunn, 1976.[278] The Gunn study appears to have used the Leeds AAP system and was based in the A&E department of Bangor General Hospital	Unclear
Logistic regression	Wong, 1990,[279] 1992,[63] 1994[137]	To aid doctors in the diagnosis of acute appendicitis in patients with APP	Calculator-based scoring system	Eight items: shifting of pain, duration of pain 6–24 hours, progressive worsening, fever, rebound tenderness, guarding, history of illness and site of tenderness	Bayesian scoring system Modified form of Leeds AAP form to collect data (42 items). Bivariate analysis (using χ^2 statistic) to select eight items that were significantly different between those with appendicitis and those without (at $p < 0.01$). Difference in log LR between symptom present/absent was used as weight of evidence (as opposed to the use of log LR). Difference is log OR. Lack of conditional independence adjusted using logistic regression (similar to Spiegelhalter and Knill Jones method[76])	Score from a calculator-based program

Appendix 9

Quality assessment of studies in the systematic review of accuracy studies of decision tools

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Alvarado, 1986,[58] USA	Weighted scoring system (Alvarado score)	*Design and recruitment* Retrospective cohort study with consecutive sampling of all cases of patients admitted to hospital for suspected appendicitis *Description of participants* Age and gender breakdown and mean duration of AAP were reported *Completeness of data* Patients' records with incomplete clinical information were excluded: 277 out of 305 eligible patients were included in the study	*Reference standard based on* Histopathology of excised appendix for disease positives and final diagnosis for disease negatives (unclear whether standardised criteria were used). No postdischarge follow-up Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to decision tool results. DT user was not blind to reference standard *Verification bias* All DT results and unaided doctors' decisions were compared to reference standard. Not all DT results and not all unaided doctors' decisions were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnosis *Type of data set* Article described a development study, so only training-set data were used *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were not reported, but 2 × 2 table could be constructed from the article *Handling of indeterminate scores/missing data* Excluded
Bond, 1990[112] USA	Weighted scoring system (Alvarado score)	*Design and recruitment* Prospective cohort study with consecutive sampling of all cases of patients admitted to hospital for suspected appendicitis *Description of participants* Age range of children was reported, but the gender breakdown was not given *Completeness of data* All 189 eligible patients were included in the study	*Reference standard based on* Histopathology of excised appendix for disease positives. Postdischarge follow-up for disease negatives Reference standard in effect based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was not part of the reference standard for disease negatives. *Blinding* Unclear whether reference standard was allocated blind to decision tool results. Unclear whether decision tool user was blind to reference standard *Verification bias* All decision tool results were compared to reference standard. All decision tool results were in effect compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnosis *Type of data set* Article described a test-set validation study of the Alvarado score using a prospective data set at a different centre from where it was developed *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Bond, 1990[112] USA	Discrimination rule	*Design and recruitment* Prospective cohort study with consecutive sampling of all cases of patients admitted to hospital for suspected appendicitis *Description of participants* Age range of children was reported, but the gender breakdown was not given *Completeness of data* All 189 eligible patients were included in the study	*Reference standard based on* Histopathology of excised appendix for disease positives. Postdischarge follow-up for disease negatives Reference standard in effect based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was not part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to decision tool results. Unclear whether decision tool user was blind to reference standard *Verification bias* All decision tool results were compared to reference standard. All decision tool results were in effect compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnosis *Type of data set* Article described a development study of a discriminant function using training-set data *Identity of tool developer* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear
De Dombal, 1972,[113] UK	Bayesian (Leeds)	*Design and recruitment* Prospective cohort study with consecutive recruitment of AAP patients admitted to hospital *Description of participants* Age and gender breakdown was not reported *Completeness of data* 302 out of 304 eligible patients were included in the study	*Reference standard based on* Final diagnosis for disease positives (most made at operation, some used biochemical evidence) and final diagnosis for disease negatives (negative laparotomy or discharged home). No post-discharge follow-up Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to decision tool results. Unclear whether decision tool user was blind to reference standard *Verification bias* All decision tool results were compared to reference standard. Not all decision tool results were compared to the same reference standard, so there is scope for differential verification bias	*Comparison* Authors compared performance of DT with unaided doctors' diagnosis *Type of data set* A test-set validation study was conducted, using a new prospective data set at the centre where the tool was developed *Identity of tool evaluator* Evaluator of DT was also its developer	*Reporting of results* 2 × 2 tables were reported, so sensitivities and specificities could be computed, even though authors only reported crude accuracies *Handling of indeterminate scores/missing data* Patient records with incomplete clinical information excluded

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Edwards, 1986,[114] UK	Bayesian (Sheffield)	*Design and recruitment* Prospective cohort study with representative sample of patients ("well over 95% of possible data points", according to authors) *Description of participants* Age and gender breakdown of patients was not reported *Completeness of data* 344 out of 366 eligible patients were included in the study	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis, recovery without surgery or histopathological confirmation of normal excised appendices for disease negatives. No postdischarge follow-up Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT was part of the reference standard for disease negatives. *Blinding* Unclear whether reference standard observations were conducted blind to decision tool results. Decision tool user was blind to reference standard *Verification bias* Not all decision tool results were compared to reference standard. Not all decision tool results were compared to same reference standard	*Comparison* Authors compared performance of DT with unaided doctors' diagnosis *Type of data set* A test-set validation study was conducted, using prospective data at a different centre *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were not reported, but 2 × 2 table could be constructed from the article *Handling of indeterminate scores/missing data* Excluded
Eskelinen, 1992,[115] Finland	Logistic regression	*Design and recruitment* Prospective cohort study with representative sample of patients *Description of participants* Age and gender breakdown was reported *Completeness of data* All 1333 patients included at the start of the study were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives with standard criteria (OMGE). No postdischarge follow-up Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard, so there is scope for differential verification bias	*Comparison* Authors compared performance of DT with unaided doctors' diagnostic accuracy *Type of data set* Article described a development study of logistic regression model using training-set data *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were reported, but no 2 × 2 table could be constructed from data reported in the article *Handling of indeterminate scores/missing data* Unclear

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Fenyo, 1987,[116] Sweden	Logistic regression	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients *Description of participants* Age and gender breakdown was reported *Completeness of data* 830 out of 867 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and follow-up of patients' records 1–2 years after admission to check final diagnosis Reference standard in effect based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was not part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. All DT results were in effect compared to the same reference standard	*Comparison* Authors compared diagnostic accuracy of DT with aided doctors' diagnostic accuracy *Type of data set* Article described a development and test-set validation study of a Bayesian scoring system. A test-set validation study was conducted, using a new prospective data set at the centre where the tool was developed *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were reported together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear *Comments* Details of training-set study not reported in quality assessment tables because of insufficient information, e.g. no accuracy data were reported

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Hallan, 1997,[117] Norway	Weighted scoring system (modified Alvarado score)	*Design and recruitment:* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Median age, age range and gender breakdown of patients reported. Median duration of AAP also reported *Completeness of data* 257 out of 309 eligible patients were included in the study	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives (unclear whether standardised criteria were used) Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to decision tool results. Unclear whether decision tool user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnosis *Type of data set* A test-set validation study was conducted, using a random split sample *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were reported together with 2 × 2 table *Handling of indeterminate scores/missing data* Excluded
Hallan, 1997;[117] Hallan, 1997,[118] Norway	Logistic regression	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Median age, age range and gender breakdown of patients reported. Median duration of AAP also reported *Completeness of data* 257 out of 309 eligible patients were included in the study	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives (unclear whether standardised criteria were used) Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to decision tool results. Unclear whether decision tool user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnosis *Type of data set* Article described a development and test-set validation study of a logistic regression scoring system. A test-set validation study was conducted, using a random split sample *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Excluded *Comments* Details of training-set study not reported in quality assessment tables because of insufficient information, e.g. no accuracy data were reported

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Horrocks, 1976,[199] Canada	Bayesian (Leeds)	*Design and recruitment* Retrospective cohort study of representative sample of unselected patients *Description of participants* Age and gender breakdown was not reported *Completeness of data* All 104 eligible patients were included in the analysis	*Reference standard based on* Final diagnosis for disease positives (most determined at operation, some used biochemical evidence), and final diagnosis for disease negatives (negative laparotomy or discharged home) Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* A test-set validation study was conducted, using prospective data at a different centre *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were not reported, but 2 × 2 table could be constructed from the article *Handling of indeterminate scores/missing data* Excluded
Izbicki, 1992,[120] Germany	Discrimination rule	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Age and gender breakdown was reported *Completeness of data* All 150 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. Postdischarge follow-up for disease negatives Reference standard in effect based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was not part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. All DT results were in effect compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* Article described a development study using training-set data *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were not reported, but 2 × 2 table could be constructed from the article *Handling of indeterminate scores/missing data* Unclear

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Jahn, 1997,[121] Denmark	Discrimination rule	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Age and gender breakdown was reported *Completeness of data* All 222 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. Postdischarge follow-up for disease negatives Reference standard in effect based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was not part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. DT user was not blind to reference standard *Verification bias* All DT results were compared to reference standard. All DT results were in effect compared to the same reference standard	*Comparison* Authors compared performance of DT with unaided doctors' diagnostic accuracy *Type of data set* Article described a development study using training-set data *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Sensitivity analysis
Jawaid, 1999,[122] Pakistan	Logistic regression	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Age and gender breakdown was reported *Completeness of data* 99 out of 126 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. No postdischarge follow-up for disease negatives Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was not blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* Article described a test-set validation study using a prospective data set at the same centre *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were not reported, but 2 × 2 table could be constructed from the article *Handling of indeterminate scores/missing data* Excluded *Comments* Authors also reported a training-set study but it was excluded because patients were not unselected

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Kirkeby, 1987,[123] Norway	Bayesian (Leeds)	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with AAP *Description of participants* Age and gender breakdown was not reported *Completeness of data* All 77 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. No postdischarge follow-up for disease negatives Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors compared performance of DT with unaided doctors' diagnostic accuracy *Type of data set* Article described a test-set validation study using a prospective data set at a different centre *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were not reported, but 2 × 2 table could be constructed from the article *Handling of indeterminate scores/missing data* Treated as negative
Kraemer, 1993,[124] Germany	Bayesian (Leeds)	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with AAP *Description of participants* Age and gender breakdown was not reported *Completeness of data* All 1254 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. No postdischarge follow-up for disease negatives Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* Article described a test-set validation study using a prospective data set at a different centre *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were not reported, but 2 × 2 table could be constructed from the article *Handling of indeterminate scores/missing data* Unclear

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Leaper, 1972,[125] UK	Bayesian (Leeds)	*Design and recruitment* Prospective cohort study with consecutive recruitment of AAP patients admitted to hospital *Description of participants* Age and gender breakdown was not reported *Completeness of data* 468 out of 472 eligible patients were included in the analysis	*Reference standard based on* Final diagnosis for disease positives (most made at operation, some used biochemical evidence) and final diagnosis for disease negatives (negative laparotomy or discharged home). No postdischarge follow-up Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All decision tool results were compared to reference standard. Not all decision tool results were compared to the same reference standard, so there is scope for differential verification bias	*Comparison* Authors compared performance of DT with unaided doctors' diagnosis *Type of data set* A test-set validation study was conducted, using a new prospective data set at the centre where the tool was developed *Identity of tool evaluator* Evaluator of the DT was also its developer	*Reporting of results* 2 × 2 tables were reported, so sensitivities and specificities could be computed, even though authors only reported crude accuracies *Handling of indeterminate scores/missing data* Patient records with incomplete clinical information excluded

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Lindberg, 1988,[126] Sweden	Logistic regression	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Age and gender breakdown was reported *Completeness of data* All 96 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. Follow-up of patients' records 1–2 years after admission to check final diagnosis Reference standard in effect based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was not part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. All DT results were in effect compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* Article described a development and test-set validation study of a Bayesian scoring system. A test-set validation study was conducted, using a new prospective data set at the centre where the tool was developed *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear *Comments* Details of training-set study not reported in quality assessment tables because of insufficient information, e.g. no accuracy data were reported

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Macklin, 1997,[127] UK	Weighted scoring system (modified Alvarado score)	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients admitted to hospital for suspected appendicitis *Description of participants* Age and gender breakdown of patients was reported *Completeness of data* All 118 eligible patients were included in the study	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. No postdischarge follow-up for disease negatives Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to decision tool results. Decision tool user was blind to reference standard *Verification bias* All decision tool results were compared to reference standard. Not all decision tool results were compared to the same reference standard	*Comparison* Authors compared performance of DT with unaided doctors' diagnosis, but reported only crude accuracies of unaided doctors, without the data needed to derive sensitivity and specificity *Type of data set* Article described a test-set validation study of the Alvarado score using a prospective data set at a different centre from where it was developed *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Malik, 1998,[128] India	Weighted scoring system (modified Alvarado score)	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients for suspected appendicitis *Description of participants* Age and gender breakdown of patients was reported *Completeness of data* All 106 eligible patients were included in the study	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. No postdischarge follow-up for disease negatives Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to decision tool results. Decision tool user was blind to reference standard *Verification bias* All decision tool results were compared to reference standard. Not all decision tool results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnosis *Type of data set* Article described a test-set validation study of the Alvarado score using a prospective data set at a different centre from where it was developed *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Ohmann, 1999,[129] Germany and Austria	Logistic regression	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Age and gender breakdown was reported *Completeness of data* All 96 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. Postdischarge follow-up for disease negatives Reference standard in effect based on a standard set of criteria	*Incorporation bias* The DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was not part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. All DT results were in effect compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* A test-set validation study was conducted, using a prospective data set at a different centre from where the tool was developed *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear *Comments* Details of training-set study not reported in quality assessment tables because of insufficient information, e.g. no accuracy data were reported

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Owen, 1992,[130] UK	Weighted scoring system (modified Alvarado score)	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Age distribution was not reported. Gender breakdown was reported for adults but not for children *Completeness of data* All 215 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. No postdischarge follow-up for disease negatives Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* A test-set validation study was conducted, using a prospective data set at a different centre from where the tool was developed *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were reported together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear
Pesonen, 1996,[131] Finland	Four neural network algorithms developed elsewhere	*Design and recruitment* Prospective cohort study with representative sample of patients with suspected appendicitis *Description of participants* Age distribution was not reported. Gender breakdown was reported *Completeness of data* 457 out of 669 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives with standard criteria (OMGE). No postdischarge follow-up Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors compared performance of DT with unaided doctors' diagnostic accuracy *Type of data set* A test-set validation study was conducted, using a random split sample *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Pesonen, 1997,[132] Finland	Discrimination rule	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Age and gender breakdown was not reported. *Completeness of data* 1064 out of 1333 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives with standard criteria (OMGE). No postdischarge follow-up Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors compared performance of DT with unaided doctors' diagnostic accuracy *Type of data set* A test-set validation study was conducted, using a random split sample *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear
Saidi, 2000,[133] Iran	Weighted scoring system (Alvarado score)	*Design and recruitment* Prospective cohort study with consecutive recruitment of patients with suspected appendicitis *Description of participants* Age distribution was not reported. Gender breakdown was reported for adults but not for children *Completeness of data* All 128 eligible patients were included in the analysis	*Reference standard based on* Histopathology of excised appendix for disease positives, and final diagnosis for disease negatives. Postdischarge follow-up for disease negatives Potential for differences in reference standard as it is not based on a standard set of criteria. Reference standard in effect based on a standard set of criteria	*Incorporation bias* The DT or component(s) of the DT was not part of the reference standard for disease positives. The DT or component(s) of the DT was not part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. All DT results were in effect compared to the same reference standard	*Comparison* Authors compared performance of DT with unaided doctors' diagnostic accuracy *Type of data set* A test-set validation study was conducted, using a prospective data set at a different centre from where the tool was developed *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity of DT were reported, together with 2 × 2 table *Handling of indeterminate scores/missing data* Unclear

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Staniland, 1980,[134] Norway	Bayesian (Leeds)	*Design and recruitment* Prospective cohort study of representative sample of unselected patients *Description of participants* Age and gender breakdown was not reported *Completeness of data* 313 out of 675 eligible patients were included in the analysis	*Reference standard based on* Final diagnosis for disease positives (most determined at operation, some used biochemical evidence), and final diagnosis for disease negatives (negative laparotomy or discharged home) Potential for differences in reference standard as it is not based on a standard set of criteria	*Incorporation bias* Unclear whether the DT or component(s) of the DT was part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* Not all DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* A test-set validation study was conducted, using prospective data at different centre *Identity of evaluator* Evaluator of DT was also its developer	*Reporting of results* Sensitivity and specificity of DT were not reported, but 2 × 2 table could be constructed from the article *Handling of indeterminate scores/missing data* Excluded
Sutton, 1989,[135] Scotland	Bayesian (CAD-A)	*Design and recruitment* Prospective cohort study with consecutive recruitment of unselected patients with AAP *Description of participants* Age and gender breakdown was not reported *Completeness of data* 4998 out of 5340 eligible patients were included in the analysis	*Reference standard based on* Final diagnosis for disease positives and final diagnosis for disease negatives Potential for differences in reference standard as it is unclear whether a standard set of criteria is applied	*Incorporation bias* The DT or component(s) of the DT was part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* Unclear whether all DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors compared performance of DT with unaided doctors' diagnostic accuracy *Type of data set* A test-set validation study was conducted, using prospective data at different centre *Identity of evaluator* Evaluator of DT was not its developer	*Reporting of results* Sensitivity and specificity were reported, and 2 × 2 table could be constructed from data reported in the article *Handling of indeterminate scores/missing data* Excluded

continued

Study, country	Reasoning method (tool name, if any)	Study design and sample selection	Reference standard	Bias	Testing and comparison of decision tool	Reporting of results/comments
Wong, 1990;[279] Wong, 1992;[63] Wong, 1994,[137] Hong Kong	Logistic regression	*Design and recruitment* Prospective cohort study with representative sample of unselected patients with AAP *Description of participants* Age and gender breakdown was reported *Completeness of data* 397 out of 411 eligible patients were included in the analysis	*Reference standard based on* Final diagnosis for disease positives and final diagnosis for disease negatives Potential for differences in reference standard as it is unclear whether a standard set of criteria is applied	*Incorporation bias* The DT or component(s) of the DT was part of the reference standard for disease positives. The DT or component(s) of the DT was part of the reference standard for disease negatives *Blinding* Unclear whether reference standard was allocated blind to DT results. Unclear whether DT user was blind to reference standard *Verification bias* All DT results were compared to reference standard. Not all DT results were compared to the same reference standard	*Comparison* Authors did not compare performance of DT with unaided doctors' diagnostic accuracy *Type of data set* Article described a development study using training-set data *Identity of evaluator* Evaluator of tool was also its developer	*Reporting of results* Sensitivity and specificity were reported, and 2 × 2 table could be constructed from data reported in the article *Handling of indeterminate scores/missing data* Unclear

OMGE, Organisation Mondiale de Gastro-Entérologie (World Organisation of Gastroenterology).

Appendix 10

Summary table of excluded studies in the review of impact studies

Study, country	**Reasons for exclusion**
Adams, 1986,[51] UK	Not a randomised or quasi-randomised trial
Clifford, 1986,[165] UK	Not a randomised or quasi-randomised trial
De Dombal, 1974,[168] UK	Not a randomised or quasi-randomised trial
De Dombal, 1993,[164] Nineteen European countries	Not a randomised or quasi-randomised trial
Fenyo, 1987,[19] Sweden	Not a randomised or quasi-randomised trial
Fenyo, 1987,[166] Sweden	Not a randomised or quasi-randomised trial
Fenyo, 1997,[173] Sweden	Not a randomised or quasi-randomised trial
Gough, 1988,[174] Australia	Not a randomised or quasi-randomised trial
Gruer, 1977,[169] UK	Not a randomised or quasi-randomised trial
Lawrence, 1987,[167] UK	Not a randomised or quasi-randomised trial
Scarlett, 1986,[170] UK	Not a randomised or quasi-randomised trial
Sutton, 1989,[135] UK	Not an impact study (inappropriate study question)
Wilson, 1975,[171] UK	Not a randomised or quasi-randomised trial
Wilson, 1977,[172] UK	Not a randomised or quasi-randomised trial

Health Technology Assessment Programme

Prioritisation Strategy Group

HTA Commissioning Board

Current and past membership details of all HTA ‘committees’ are available from the HTA website (www.hta.ac.uk)

Therapeutic Procedures Panel

Current and past membership details of all HTA 'committees' are available from the HTA website (www.hta.ac.uk)

Expert Advisory Network

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.hta.ac.uk) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@hta.ac.uk
http://www.hta.ac.uk

ISSN 1366-5278