HOW TO IDENTIFYAND ASSESS EVIDENCE FROM DIAGNOSTIC STUDIES OF IMAGING

INFORMATION PACK ON HOW TO DEVELOP GUIDELINES FOR 'MAKING THE BEST USE OF A DEPARTMENT OF CLINICAL RADIOLOGY - FIFTH EDITION (MBUR5)'

Miriam Brazzelli ♦

Margaret Astin ♦

Jeremy Grimshaw ♦

Gillian Needham ♠

♦ Health Services Research Unit, University of Aberdeen

♠ Postgraduate Medical Department. Medical School Foresterhill, Aberdeen

INDEX

LIST OF APPENDICES

FOREWORD

This information pack has been devised in response to specific issues raised by the Special Interest Group leaders during the MBUR5 Training Day held in London on 13th December 2000. It aims to clarify the selection process and interpretation of diagnostic studies of imaging and radiological investigations. In particular, the information pack aims to assist the searching and critical appraising processes. It is complementary to the existing templates developed by Chris Squire for the Royal College of Radiologists. We have included a revised grading system which is more appropriate for diagnostic studies. We believe that even for those who have already completed their tables the adoption of this new system will require minimal additional effort.

The information pack comprises the following:

- Guidance on which studies should be considered when developing recommendations for MBUR5.
- Search strategies for retrieval of relevant studies.
- Guidance on how to critically appraise diagnostic studies.
- Guidance on how to assess the technical performance of diagnostic tests and calculate sensitivity, specificity, predictive values and likelihood ratios.
- Examples of the current guideline, evidence and reference tables format.
- Factors to be considered when making recommendations.
- A revised grading system for diagnostic studies.

In the Appendices, we provide: examples of checklists for the appraisal of primary studies, meta-analyses and clinical decision rules; a description of the technical characteristics of diagnostic studies; and a glossary of terms.

WHICH STUDIES SHOULD BE CONSIDERED WHEN DEVELOPING RECOMMENDATIONS FOR MBUR5?

Studies of diagnostic tests must demonstrate that the new test is accurate in distinguishing patients with the target disease from patients without the target disease. All comparative studies in which a new test is compared with a reference ("gold") standard are eligible for inclusion.

The following types of study should be considered when developing recommendations for MBUR5:

- Comparative prospective studies (e.g. cross-sectional studies, cohort studies) in which all participants undergo the new test as well as the reference ("gold") standard.

- Comparative studies performed on a non-consecutive series of patients (e.g. retrospective case note studies, case-control studies).

- Randomised controlled trials to determine the magnitude of the combination of both diagnostic test and treatment effect on outcomes when a proper reference ("gold") standard is not available and follow-up of patients (after treatment) is required.

- Systematic reviews of primary studies, which employ explicit and reproducible methods to identify, appraise and synthesise evidence.

- Meta-analyses, which are systematic reviews which statistically combine the results of two or more primary studies.

- Clinical decision rules, which formally contribute to the accuracy of diagnostic and prognostic assessments and inform the decision making process.

The following studies should not be considered :

- Narrative reviews (although it may be useful to scan reference lists of narrative reviews for other potentially relevant studies).

- Case reports, pictorial essays.

- Therapeutic, prognostic, pilot, volunteer, phantom and animal studies.

- Studies about technical developments of instrumentation.

Precise definitions of study designs and technical terms used in diagnostic studies are detailed in the glossary in Appendix 9.

IDENTIFYING EVIDENCE

In this section, we provide examples of simple search strategies for identifying relevant primary studies, systematic reviews, and clinical decision rules. The aim of these searches is to maximise specificity rather than sensitivity to increase the relevance of retrieved articles. As a consequence, guideline developers should be aware that some relevant articles might not be retrieved.

MEDLINE is the most frequently used electronic database for retrieval of published medical articles. It is the electronic version of the *Index Medicus*, compiled by the National Library of Medicine of the United States, and indexes over 3,700 journals.

The Medline database is easily accessible on line or by CD-ROM by means of various interfaces. The most common interfaces are: Ovid and Silver Platter, which are almost universally available on networked university systems or in medical and science libraries. Medline can be also accessed over the Internet as PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi).

Articles can be retrieved from Medline in two ways:

- by any word or phrase contained in the citations.
- by MeSH terms (or Medical Subject Headings), a limited thesaurus of medical terms which has a tree-like structure. The further down a branch the more specific the terms become.

The 'explode' command for a MeSH term is very powerful and enables the search to retrieve all articles indexed with the term, and any of the narrower terms in the branch.

The syntax of a search strategy (the format of letters, numbers and symbols) is very important and differs according to the particular interface used (Ovid, Silver Platter, PubMed).

Some frequently asked questions about searching and a key to the symbols used in the search strategies are reported in Appendix 1.

Identifying Potentially Relevant Primary Studies

STEP 1: Identification of potentially relevant diagnostic articles (irrespective of study design). Enter the following search strategy in the command line of Medline.

Line no.	**OVID**	**SILVER PLATTER**	**PUBMED**
1	exp "sensitivity and specificity"/	exp sensitivity-and-specificity	Sensitivity and specificity[mh]
2	exp mass screening/	exp mass screening	mass screening[mh]
3	sensitivity.tw.	sensitivity in ti,ab	sensitivity[tw]
4	specificity.tw.	specificity in ti,ab	specificity[tw]
5	(predictive adj3 value$).tw.	predictive near3 value* in ti,ab	"predictive value*"[tw]
6	accuracy.tw.	accuracy in ti,ab	accuracy[tw]
7	screen$.tw.	screen* in ti,ab	screen*[tw]
8	or/1-7	#1 or #2 or #3 or #4 or #5 or #6 or #7	#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7

STEP 2: Elimination of papers with irrelevant study designs. The set can be further refined by excluding irrelevant articles such as case reports, editorials, letters and reviews. The follow-on set should mainly contain primary diagnostic studies (see lines 9-14), but might exclude some systematic reviews.

Line no.	**OVID**	**SILVER PLATTER**	**PUBMED**
9	case report/	Case-report in tg	case report [mh]
10	letter.pt.	letter in pt	letter [pt]
11	editorial.pt.	editorial in pt	editorial [pt]
12	comment.pt.	Comment in pt	comment [pt]
13	review.pt.	review in pt	review [pt]
14	8 not (9 or 10 or 11 or 12 or 13)	#8 not (#9 or #10 or #11 or #12 or #13)	#8 NOT (#9 OR #10 OR #11 OR #12 OR #13)

STEP 3: Identification of studies relating to specific investigation modalities. A very broad MeSH term covering all diagnostic imaging modalities is the term *diagnostic imaging* (see line 15). To narrow down the search to a particular type of investigation (e.g. angiography, mammography, ultrasonography, tomography) select the appropriate term in the sub-branch of the "diagnostic imaging" heading of the MeSH tree. For MeSH terms, USA rather than UK terminology and spelling are usually used (see Appendix 1 for more information).

STEP 4: Identification of studies relevant to a specific clinical condition. The final stage is to add the clinical condition of interest preferably as a MeSH term (e.g. pulmonary embolism, gallbladder, digestive system - see line 16).

The relevant sets are then combined using the AND operator (see line 17).

At this stage the set can be limited to "human" studies by selecting the *Limit* option on the interface (see line 18).

	OVID	SILVER PLATTER	PUBMED
15	exp diagnostic imaging/	exp diagnostic imaging	diagnostic imaging[mh]
16	exp *insert condition of interest*	exp *insert condition of interest*	*Insert condition of interest* [mh]
17	14 and 15 and 16	#14 and #15 and #16	#14 AND #15 AND #16
18	Limit 17 to human	Limit 17 to human	Limit 17 to human

Identifying potentially relevant systematic reviews

To include systematic reviews and meta-analyses in a set, add on the following strategy to the end of the diagnostic strategy (Step 1 above):

	OVID	SILVER PLATTER	PUBMED
19	meta-analysis.pt.	meta-analysis in pt	meta-analysis[pt]
20	meta-analysis/	meta-analysis in mesh	meta-analysis[mh]
21	(data adj3 synthesis).tw.	"data near synthesis" in ti,ab	"data synthesis" [tw]
22	(published adj3 studies).ab.	"published near studies" in ab	"published studies" AND hasabstract
23	(data adj extraction).ab.	"data near extraction" in ab	"data extraction" {tw]
24	systematic$ adj3 (review$ or overview$).tw.	Systematic* near3 review*	"systematic* review*"[tw]
25	(meta?analy$ or meta analy$).tw.	Systematic* near3 overview*	"systematic-overview"
26	or/19-26	meta-analy* or meta?analys*	"meta-analy*"[tw]
27	8 and 15 and 16 and 26	#19 or #20 or #21 or #22 or #23 or #24 or #25 or #26	#19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR 26
28		#8 and #15 and #16 and #27	#8 AND #15 AND #16 AND #27

(Source for the systematic review strategy adapted from: NHS Centre for Reviews and Dissemination, University of York (http://www.york.ac.uk/inst/crd/search.htm - Nov 2001)

Identifying potentially relevant clinical decision rules

To identify articles about clinical decision rules add on the following strategy to the end of the diagnostic strategy (see Step 1 above):

	OVID	**SILVER PLATTER**	**PUBMED**
1	exp clinical protocols/	exp clinical protocols	clinical protocols[mh]
2	exp practice guidelines/	exp practice guidelines	practice guidelines[mh]
3	exp algorithm/	exp algorithm	algorithm[mh]
4	exp decision making/	exp decision making	decision making[mh]
5	rule$.tw.	rule* in ti, ab	rule*[tw]
6	or/1-5	#1 or #2 or #3 or #4 or #5	#1 OR #2 OR #3 OR #4 OR #5
7	*add in clinical and modality terms and combine with* AND	*add in clinical and modality terms and combine with* AND	*add in clinical and modality terms and combine with* AND

Saving search strategies

Ovid and Silver Platter search histories can be saved and re-run or modified afterwards. Look out for a "save search" key, click on it and follow directions. Precise instructions on how to save a search strategy are also available from your local medical library.

PubMed searches can be saved in a number of ways:

- As a text file
- On to the clipboard
- On to a floppy disk
- As a URL on your web browser

A comprehensive "Help" is available from the PubMed Help file.

In case of uncertainty about using the search strategies outlined above it is advisable to consult a medical librarian or an information scientist.

Relevant web sites for further information about search strategies (accessed Nov 2001)

http://wwwlib.jr2.ox.ac.uk/caspfew/filters
http://www.nthames-health.tpmde.ac.uk/evidence_strategies/index.htm
http://cebm.jr2.ox.ac.uk/docs/searching.html
http://www.york.ac.uk/inst/crd/search.htm

QUALITY APPRAISAL

Once potentially relevant studies have been identified by the searches, it is important to screen the Medline citation to assess the likely relevance of the identified study to the recommendation of interest. Following this, hard copies of relevant studies should be retrieved for further assessment. Ideally, guideline developers shold consider all relevant studies, however this may not be feasible within available resources. If a guideline developer needs to select a subset of studies to consider in greater details, they should initially select studies with more robust designs (see below).

When developing guideline recommendations, it is important to appraise the quality of studies to allow the guideline developer to give greater weight to studies that are likely to be unbiased. Some types of study design are less likely to be biased (Box 1), thus we can develop a hierarchy of evidence based upon design and give greater weight to studies with more robust designs. This is the rationale for excluding studies with certain designs (e.g. narrative reviews). Thus guideline developers should give greater weight to a well-conducted systematic review than individual cohort or case control studies.

Box 1 Hierarchy of evidence

1 Systematic reviews/meta-analyses/clinical decisions rules

2 Cross-sectional studies/randomised controlled trials

3 Cohort studies

4 Case-control studies

However a poorly conducted systematic review may be more biased than a well conducted cross-sectional study. Hence, it is important to appraise the quality of individual studies. Most published studies are likely to be flawed in some way. The purpose of critically appraising a paper is to identify methodological flaws and assess their likely impact on study results. Minor methodological flaws may be unlikely to influence study results; whereas a major flaw undermines the main findings of a study.

Various checklists have been published for the assessment of diagnostic studies. When assessing the quality and validity of a primary diagnostic study it is important to consider how the authors selected their patients and whether they applied both the test and the reference standard to the sample of patients. Ideally, the best diagnostic study is a comparative prospective study in which all participants undergo the new test as well as the reference ('gold') standard and the results are

independently and blindly interpreted by at least two assessors (Box 2). The checklist for evaluating the validity of diagnostic studies reported in Appendix 2 has been specifically modelled on existing checklists for the purposes of the MBUR5. Checklists for evaluating randomised controlled trials, meta-analyses of diagnostic tests and clinical decision rules are presented in Appendices 3, 4 and 5 respectively.

Box 2 Key points for diagnostic studies

- ***Selection of patients***
- ***Choice of gold standard***
- ***Both the target test and the reference standard performed on all patients***
- ***Blind assessment***

ASSESSING THE TECHNICAL PERFORMANCE OF A TEST

The technical performance of a diagnostic study is measured in terms of sensitivity and specificity, predictive values and likelihood ratios. Often these measures are not adequately reported, labelled or calculated in primary diagnostic studies and may need to be recalculated from raw data (where these are provided).

	Disease Present	*Disease Absent*
Test Positive	*a*	*b*
Test Negative	*c*	*d*

2x2 Table used to measure the technical performance of diagnostic studies

Sensitivity of a test (a/a+c) is the proportion of people with the target disorder who have a true positive test result while the **specificity** (d/b+d) is the proportion of people without the disease who have a true negative test result.

The **positive predictive value of a test** (or post-test probability of the target disease) is the proportion of people with true positive test results over all positive results (a/a+b).

The **negative predictive value of a test** (or post-test probability of not having the target disease) is the proportion of people with true negative results over all negative results (d/c+d)

A **likelihood ratio** measures the probability of having a given test result in a patient with the target disease compared with the corresponding probability in a patient without the disease.

Information on how to calculate and interpret data from diagnostic studies is provided in Appendix 6 and an example of a data abstraction form is given in Appendix 7.

Measures to consider in a diagnostic study

- ***Sensitivity and specificity of the test***
- ***Predictive values***
- ***Likelihood ratios***

EVIDENCE TABLES

Findings of individual diagnostic studies, systematic reviews and clinical decision rules must be tabulated using the templates developed for the 5th edition of the Making the Best Use of a Department of Clinical Radiology guidelines (MBUR) by the Royal College of Radiologists, London. An example of completed tables is reported in Appendix 8.

MAKING RECOMMENDATIONS

To draw recommendations for clinical practice the following factors should be taken into account:

- Technical performance of the test for the purpose of the diagnosis
- Impact of the diagnostic test on clinical decision making
- Potential harm
- Costs
- Practicality

Where several tests perform adequately, information about all these tests should be provided and tests available in the UK should preferably be highlighted.

GRADING RECOMMENDATIONS

Recommendations for clinical practice should be based on the best evidence available. The level of evidence and grade of recommendations for the diagnostic literature are formulated according to the quality and precision of primary diagnostic studies and secondary research publications. The stronger the evidence, the more confident one can be about the use of new diagnostic investigations. The recommended grading system (A, B, C, D) reported below is a revised version of that originally developed by Sackett and colleagues (2000).

Grade	Level of evidence	Diagnosis
A	1	▪ High quality diagnostic studies in which a new test is independently and blindly compared with a reference standard in an appropriate spectrum of patients. ▪ Systematic reviews and meta-analyses of such high quality studies. ▪ Diagnostic clinical practice guidelines/ clinical decision rules validated in a test set
B	2/3	Any one or two of the following: ▪ Studies with a blind and independent comparison of the new test and reference standard in a set of non-consecutive patients or confined to a narrow spectrum of subjects ▪ Studies in which the reference standard was not performed on all subjects ▪ Systematic reviews of such studies ▪ Diagnostic clinical practice guidelines/clinical decision rules not validated in a test set
C	4/5	Any one or two of the following: ▪ Studies in which reference standard was not objective ▪ Studies in which the comparison between the new test and the reference standard was not blind or independent ▪ Studies in which positive and negatives test results were verified using different reference standards ▪ Studies performed on an inappropriate set of patients
D	6	Experts' opinion

Modified from Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine. How to Practice and Teach EBM. 2nd ed. Edinburgh: Churchill Livingstone, 2000.

APPENDIX 1

FREQUENTLY ASKED QUESTIONS ON HOW TO SEARCH FOR EVIDENCE

How is it possible to reduce the number of irrelevant articles?

- Use MeSH terms instead of free text terms.
- Use narrower MeSH terms for a more specific search.
- Use appropriate subheadings with the MeSH terms.
- Limit the search to some of the available options such as human studies, English language papers, review articles, etc. (the *Limit* command should appear on your interface). Note the Limit "human" option should exclude articles purely about animal research, but will still include articles about animal and human research. Articles can also be limited by year of publication and publication type.
- Use the Boolean AND operator to combine different aspects of the search question.
- The Boolean operator NOT can be used to narrow the search results by excluding irrelevant articles (e.g. to exclude the Publication Types case reports and editorials from your diagnostic search). Use with caution.

How is it possible to increase the number of articles retrieved?

- Use the MeSH term and a textword search, and combine using the Boolean operator OR.
- The explode command of the MeSH term will retrieve all the articles in that branch of the index tree. Note that some interfaces like PubMed will automatically do this by mapping the term to MeSH and searching all fields for the term.
- Use more search terms.
- Try different combinations of terms with related meaning/synonyms.
- Find more suitable terms to use from retrieved relevant articles.
- Use truncation or wildcard (* or $) symbols at the end of a word stem.
- Search back in time

Where is the list of MeSH terms?

The list of Medical Subject Headings can be found in most interfaces either under the "toolbox" icon (a tools icon or tools pull down menu) or in the thesaurus. This will show the full "tree" structure of MeSH. Suggestions for each interface are outlined below:

OVID:
Appropriate MeSH terms can be found in one of two ways:

1. From the Tools icon or pull down menu, select 'Permuted Index'. Enter a single term in the subject box to find all MeSH headings which contain this term. By clicking on a heading, the full tree structure can be viewed. Terms can be included in the search by clicking the relevant select checkbox.
2. Enter keyword or phrase in the search box and ensure that 'Map term to Subject Heading' box is ticked. The most appropriate MeSH terms will be displayed (up to a maximum of ten). Again, hotlinks are provided to the full tree structure or terms can be selected by clicking the select checkbox.

Command language can be used if a Mesh term is known. For example, entering *tree ultrasonography* in the search field will show the related part of the branch. Each branch of the tree can then be expanded to reveal the narrower terms within it. The *explode* command will retrieve all the articles containing the term, as well as all the narrower terms within it.

Silverplatter
Switch to the Thesaurus option and enter a single term. The appropriate section of the permuted index will be displayed containing all MeSH headings with this term. Options to browse the full tree structure or select terms to be included in the search are provided.

PubMed
Select the MeSH browser under Pubmed Services. Enter the term and click 'Go'. PubMed will map the terms to possible MeSH headings and display this list. By selecting a term to browse, the tree structure for that term will be displayed. PubMed will automatically explode any selected MeSH heading but clicking on 'Detailed Display' allows the option not to explode or to select subheadings.

What are subheadings used for?
Subheadings are used in Medline to add further precision to a MeSH term and to reduce the number of articles retrieved. Some of the most useful subheadings are: adverse effects (ae); complications (co); diagnosis (di); diagnostic use (du); drug therapy (dt); epidemiology (ep); pathology (pa); radiography (ra); radionuclide imaging (ri); ultrasonography (us) and therapy (th).

Subheadings are very useful for performing a quick search on a specific topic (e.g. **breast neoplasms/di** retrieves diagnostic articles on breast neoplasms) but will not retrieve all potentially relevant articles on the topic.

Subheadings should be used with caution as not all MeSH terms are indexed accurately with subheadings in Medline.

Table of Search Symbols

TERM OR SYMBOL	OVID	SILVER PLATTER	PUBMED
MeSH	/ or map to Mesh	.mh	[mh]
Explode	Exp	exp	(automatic)
Truncation	$	*	*
Abstract	.ab.	in ab	AND hasabstract
Wild card	?	?	Not available
Free text	.tw.	in ti,ab	[tw]
Publication type	.pt.	in pt	[pt]
Adjacency	Adj	near	Not available

APPENDIX 2

CHECKLIST FOR REVIEWING PAPERS EVALUATING DIAGNOSTIC TESTS

Author(s)

Title Journal

Type of investigation

Main findings

Comments

Were the patients selected consecutively?

Yes No Unclear

Was the diagnostic test compared with a valid reference ("gold") standard?

Yes No

Were the diagnostic test and the reference ("gold") standard performed on all participants or on randomly allocated patients (i.e. avoidance of verification bias)?

Yes No

Were the test and the reference ("gold") standard measured independently (i.e. were the assessors of the diagnostic test blind to the results of the gold standard and vice versa)?

Yes No

Did the patients sample include an appropriate spectrum of subjects (mild and severe; treated and untreated cases)?

Yes No

Were the methods for performing the diagnostic test described in sufficient detail to permit replication?

Yes No

Was the interpretation of test results consistent both within and between observers? (i.e. intra and inter-observer reliability)

Yes No

Were the characteristics of the diagnostic test adequately described? (sensitivity and specificity; predictive values; likelihood ratios) or were the data necessary for these calculations provided?

Yes No

RATIONALE FOR THE CRITERIA USED FOR THE QUALITY ASSESSMENT OF DIAGNOSTIC STUDIES

Were the patients selected consecutively?
As the best design for diagnostic studies is a comparative prospective study in which all participants undergo the new test as well as the reference ('gold') standard, it is important to determine whether the patients were adequately selected consecutively or whether selection biases existed.

Was the diagnostic test compared with a valid reference ("gold") standard?
The reference ("gold") standard must be clearly defined and must be the best available method to assess the presence or absence of the target disease. Usually pathological/histological findings, biopsy results and surgical outcomes are used as reference standards. If the reference standard is not adequate, the diagnostic test under investigation cannot appear better than the reference standard (e.g. when a new imaging test is compared to an old one) even though it might be so. Sometimes a perfect or adequate reference standard is not available however and the choice of a particular reference standard by the study authors requires critical consideration. In particular, the results of the diagnostic test under evaluation should not be incorporated into the reference standard ('incorporation' bias). Occasionally, when only clinical data are available as a reference standard, clinicians have the wrong tendency to include the diagnostic test result among the set of information they use to make the diagnosis.

When a proper reference standard is not available follow-up is sometimes used to assess the true condition of patients under scrutiny. However, if treatment intervenes during follow-up, the "gold standard status" of the patients is hampered and the specificity of the diagnostic procedure cannot be established (one cannot decide if 'non-diseased' patients at follow-up were initially false positives or were diseased people subsequently cured by means of the treatment).

Were the diagnostic test and the reference ("gold") standard performed on all participants or on randomly allocated patients (i.e. avoidance of verification bias)?
It is important to check whether all patients undergo both the diagnostic test and the reference standard. In some instances the results of the diagnostic test may have an impact on the decision to perform the reference standard. This is the case, for example, when the reference standard is performed only on individuals who have a positive result at the test under evaluation leading to a bias known as 'work-up' or 'verification' bias. Nevertheless, in certain specific circumstances it is not desirable or ethical to perform invasive procedures, which carry a morbidity and mortality risk (such as angiography), on patients with a negative test result. In this case a randomised controlled trial might be required to assign patients to each test or alternatively patients must be followed up for an adequate period of time.

Were the test and the reference ("gold") standard measured independently (i.e. were the assessors of the diagnostic test blind to the results of the gold standard and vice versa)?
One should ensure that the investigators who judged and interpreted the features of the diagnostic test being evaluated were not aware of the results of the reference standard and vice-versa. This is because knowledge of one test result can indirectly influence the interpretation of the other, leading to 'expectation' or 'ascertainment' biases.

Did the patients sample include an appropriate spectrum of subjects (mild and severe; treated and untreated cases)?
To produce useful information the test should be applied in the study to patients at different stages of the target disease; treated and untreated cases; and patients with common and less common presentations of the target disorder. This is because the selection of patients can affect the results of the diagnostic test and in particular the distribution of disease stage may affect the sensitivity and specificity of the test. The same test, for example, can generate different sensitivity figures depending on whether it has been performed only on patients with severe symptoms (more likely to be diagnosed – high sensitivity) or only on patients with early symptoms of the disease (more difficult to be diagnosed – low sensitivity).

It is necessary to ensure that patients with a variety of presentations of the target disease, as well as a variety of symptoms, have been included in the study sample.

Were the methods for performing the diagnostic test described in sufficient detail to permit replication?
The procedures to conduct the diagnostic test should be described in sufficient detail to permit replication of the study. This implies description of issues related to the preparation of patients and to technical aspects of the procedure used (e.g. dose of radiation, number of films obtained, etc.)

Was the interpretation of test results consistent both within and between observers? (i.e. intra and inter-observer variability)
Different observers must ideally agree upon the interpretation of the same test result and the same observer judging the same test on two different occasions should reach the same conclusions. However, it is possible to have different results within and between observers in a certain proportion of cases. Observer variability should be investigated and explained by the authors of the diagnostic study. Attempts to measure observer variability should be made in the study.

Were the characteristics of the diagnostic test adequately described? (sensitivity and specificity; predictive values; likelihood ratios) or were the data necessary for these calculations provided?

A diagnostic test must perform well technically to be worth using. The technical precision of a test is measured in terms of sensitivity and specificity; positive and negative predictive values; and likelihood ratios. These features of the test should be clearly reported in the study or calculated from raw data when not reported by the authors.

APPENDIX 3

CHECKLIST FOR APPRAISING THE QUALITY OF RANDOMISED CONTROLLED TRIALS

(from: Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine. How to Practice and Teach EBM. 2nd ed. Edinburgh: Churchill Livingstone, 2000)

1. Was the assignment of patients to treatment randomised and was the randomisation list concealed?

2. Was follow-up of patients sufficiently long and complete?

3. Were all patients analysed in the groups to which they were randomised?

4. Were patients and clinicians kept blind to treatment?

5. Were groups treated equally, apart from the experimental intervention?

6. Were the groups similar at the start of the trial?

APPENDIX 4

CHECKLIST FOR EVALUATING META-ANALYSES OF DIAGNOSTIC TESTS

(from: Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnsotic tests. Annals of Internal Medicine 1994; 120(8): 667-676)

- Is there a clear statement about:
  - the test of interest?
  - the disease of interest and the reference standard by which it is measured?
  - the clinical question and context?

- Is the objective to evaluate a single test or to compare the accuracy of different tests?

- Is the literature retrieval procedure described with search and link terms given?

- Are inclusion and exclusion criteria stated?

- Are studies assessed by two or more readers?
  - do the authors explain how disagreements between readers were resolved?

- Is a full listing of diagnostic accuracy and study characteristics given for each primary study?

- Does the method of pooling sensitivity and specificity take account of their interdependence?

- When multiple test categories are available, are they used in the summary?

- Is the relation examined between estimates of diagnostic accuracy and study validity of the primary studies for each of the following design characteristics:
  - appropriate reference standard?
  - independent assessment of the test or tests and reference standard?

- In comparative studies, were all of the tests of interest applied to each patient or were patients randomly allocated to tests?

- Are analytic methods used that estimate whether study design flaws affect diagnostic accuracy rather than just test threshold?

- Is the relation examined between estimates of diagnostic accuracy and characteristics of the patients and test?

- Are analytic methods used which differentiate whether characteristics affect diagnostic accuracy or test threshold?

APPENDIX 5

METHODOLOGICAL STANDARDS FOR VALIDATION OF A CLINICAL DECISION RULE

(from: McGinn TG, Guyatt GH, Wyer PC, David Naylor C, Stiell IG, Scott Richardson W, for the Evidence-Based Medicine Working Group. Users' Guide to the Medical Literature. XXII: How to use articles about clinical decision rules. JAMA 2000; 284(1): 70-84)

- Were the patients chosen in an unbiased fashion and do they represent a wide spectrum of severity of disease?
- Was there a blind assessment of the criterion standard for all patients?
- Was there an explicit and accurate interpretation of the predictor variables and the actual rule without knowledge of the outcome?
- Was there 100% follow-up of those enrolled?

APPENDIX 6

HOW TO INTERPRET DIAGNOSTIC DATA

(adapted from Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine. 2nd edition. Boston: Little, Brown and Company, 1991: 69-152)

In diagnostic studies results of an investigation (e.g. X-ray, CT scan, MRI) are used to single out patients with the target clinical condition from patients without the condition. Patients are therefore classified into two categories according to the absence or presence of the target disease.
The number of patients with positive test results and the number of people with negative test results are normally tabulated using a 2x2 table.

	Disease Present	*Disease Absent*
Test Positive	***a***	***b***
Test Negative	***c***	***d***

The two most common indices of the performance of a test are the sensitivity and the specificity. **Sensitivity** of a test is the proportion of people with the target disorder who have a true positive test result (a/a+c) while the **specificity** is the proportion of people without the disease who have a true negative test result (d/b+d).

Example of a study assessing the accuracy of x-ray in detecting tumours of the spine in 1000 patients with back pain.

	Tumour of the spine		
		Present	***Absent***
x-ray test result	***Positive***	**True positives 300** **a**	**False positives 44** **b**
	Negative	**c** **False negatives 200**	**d** **True negatives 456**
		(a + c) 500 **SENSITIVITY a/ (a+c) = 60%**	**(b + d) 500** **SPECIFICITY d/ (b+d) = 92%**

Sensitivity and specificity are calculated vertically in the table and they are constant within the population undergoing the test. In other words they do not change according to the characteristics of the population on which the test is carried out. Therefore they are useful in assessing the general impact of a test on a population of patients but not to guide clinical practice. To make decisions, clinicians need to know what is the probability for a patient with a positive test result of having the target disorder. In this case, sensitivity and specificity are of no use and predictive values of the test and likelihood ratios should be calculated instead.

Predictive values are calculated horizontally in the 2x2 table.

	Tumour of the spine		
		Present	**Absent**
x-ray test result	**Positive**	**True positives 300** **a**	**False positives 44** **b**
	Negative	**c** **False negatives 200**	**d** **True negatives 456**
		(a + c) 500 **Sen 60%**	**(b + d) 500** **Spe 92%**

Total positives=344 **(a + b)**	**Positive predictive value =** **Post-test probability of disease** $\frac{a}{(a+b)} = \frac{300}{344} = 87\%$
Total negatives= 656 **(c + d)**	**Negative predictive value =** **Post-test probability of no disease =** $\frac{d}{(c+d)} = \frac{456}{656} = 69\%$
Total = 1000 (a+b+c+d)	

Method to calculate positive and negative predictive values.

The **positive predictive value of a test** (or post-test probability of the target disease) is the proportion of people with true positive test results over all positive results. The **negative predictive value of a test** (or post-test probability of not having the target disease) is the proportion of people with true negative results over all negative results.

Predictive values are not constant but change according to the pre-test probability of the target disorder (Bayes' theorem) in the studied population. The pre-test probability (or prevalence) of the target disorder is the proportion of patients with the disease among the patient population. When the pre-test probability of a certain disease cannot be estimated from existing epidemiological studies, it is possible to extrapolate it from the data provided in the primary study (pre-test probability

observed in the study = a+c/a+b+c+d). The pre-test probability has a major influence on the diagnostic process.

In the following example predictive values of an x-ray investigation for the detection of bone tumour are calculated for patients who hypothetically have a high, low or intermediate pre-test probability of the disease.

Consider the following three clinical scenarios:

1. *Patients with back pain who have a high probability (at least 90%) of bone tumour (weight loss, persistent pain not relieved by analgesics);*

2. *Patients with back pain whose symptoms do not fit with a diagnosis of bone tumour but need to be reassured (no more than 1% of pre-probability);*

3. *Patients with back pain whose symptoms may indicate a tumour but a certain level of uncertainty is still present (intermediate probability, about 50%).*

SCENARIO 1

Pre-test Probability 90%	**Tumour of the spine**				
		Present	**Absent**		
x-ray test result	**Positive**	**True positives 545** **a**	**False positives 8** **b**	**Total positives=568** **(a + b)**	**Positive predictive value =** **Post-test probability of disease** $\frac{a}{(a+b)} = \frac{545}{568} = 96\%$
	Negative	**c** **False negatives 355**	**d** **True negatives 92**	**Total negatives= 447** **(c + d)**	**Negative predictive value =** **Post-test probability of no disease =** $\frac{d}{(c+d)} = \frac{92}{447} = 20\%$
		(a + c) 900 **Sen 60%**	**(b + d) 100** **Spe 92%**	**Total = 1000 (a+b+c+d)**	

Scenario 2

Pre-test probability 1%	Tumour of the spine				
		Present	Absent		
x-ray test result	Positive	True positives 6 a	False positives 80 b	Total positives= 86 (a + b)	Positive predictive value = Post-test probability of disease $\frac{a}{(a+b)} = \frac{6}{86} = 7\%$
	Negative	c False negatives 4	d True negatives 910	Total negatives= 914 (c + d)	Negative predictive value = Post-test probability of no disease $\frac{d}{(c+d)} = \frac{910}{914} = 99\%$
		(a + c) 10 Sen 60%	(b + d) 990 Spe 92%	Total = 1000 (a+b+c+d)	

SCENARIO 3

Pre-test probability 50%	Tumour of the spine				
		Present	Absent		
x-ray test result	Positive	True positives 300 a	False positives 44 b	Total positives = 344 (a + b)	Positive predictive value = Post-test probability of disease = $\frac{a}{(a+b)} = \frac{300}{344} = 87\%$
	Negative	c False negatives 200	d True negatives 456	Total negatives= 656 (c + d)	Negative predictive value = Post-test probability of no disease = $\frac{d}{(c+d)} = \frac{456}{656} = 69\%$
		(a + c) 500 Sen 60%	(b + d) 500 Spe 92%	Total = 1000 (a+b+c+d)	

It is likely that patients in the scenarios 1 and 2 do not need an x-ray. Patients in scenario 1 already have a high pre-test probability which is not significantly affected by the result of the x-ray (post-test probability of positive disease 96%). Even in the occurrence of a negative test result there is still an 80% probability (100% - 20% = 80%) of having a tumour. These patients will probably benefit more from CT or MRI investigations.

Patients in scenario 2 do not need an x-ray as their pre-test probability of having a tumour is very low (1%) and does not change significantly with the test (7%). On the other hand, patients in scenario 3 are more likely to be correctly diagnosed by an x-ray as their post-test probability increases from 50% to 87% (37%).

It is also possible to calculate the post-test probability of a target condition from its pre-test probability using a measure called likelihood ratio.

The likelihood ratio expresses the odds that a given test result is more likely to be obtained in a subject with the target disease than in a subject without the disease.

Pre-test odds for the target disorder (based on prevalence) x Likelihood ratio for the diagnostic test

= Post-test odds for the target disorder

Likelihood ratio for a positive test result = sensitivity / 1-specificity

Likelihood ratio for a negative test result = 1-sensitivity / specificity

The pre-test odds are calculated as pre-test probability/(1-pre-test probability) and the post-test odds can then be converted back to post-test probability:

Post-test probability = post-test odds/(post-test odds + 1)

With a likelihood ratio above 1 the probability of the disease being present increases; with a likelihood ratio below 1 the probability decreases and when the likelihood ratio is 1, the probability is unchanged (the test result does not bear diagnostic information).

In the above example the likelihood ratio for a positive test result is 7.5 (0.60/1-0.92) (that particular x-ray result is 7.5 times as likely to come from patients with tumour than from patients without a tumour). If we assume a pre-test probability of the disease of 50% the pre-test odds is 1 (0.50/1-0.50) and therefore:
post-test odds = 1 x 7.5 = 7.5 that means:

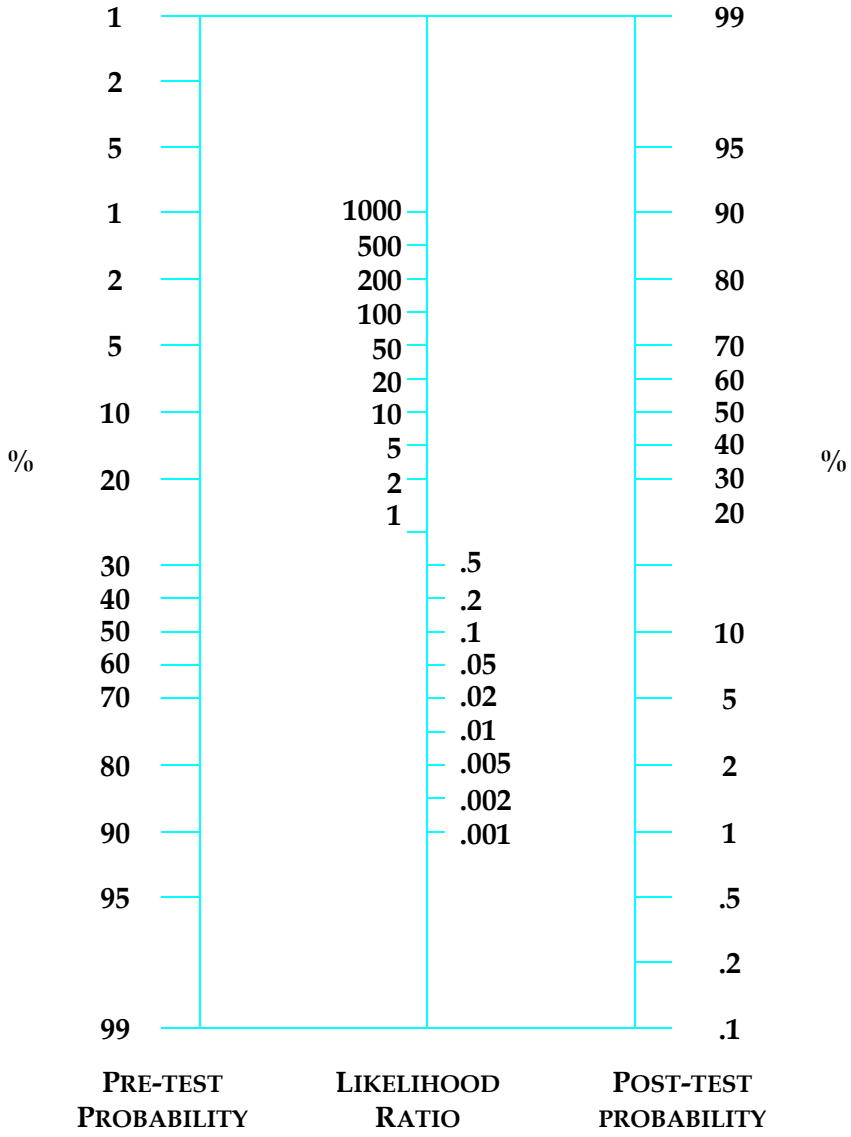
post-test probability = 7.5/7.5 +1 = 88%

Pre-test Probability → Probability/(1- Probability) → Pre-test Odds

Post-test Odds → Post-test Odds/(Post-test Odds + 1) → Post-test Probability

A quick method to compare pre-test with post-test probabilities is to use a nomogram (see *Fagan TJ. Nomogram for Bayes' theorem. New England Journal of Medicine 1975, 293: 257*) in which pre-test and post-test odds have been already converted to their corresponding probabilities. The post-test probability on the right side of the nomogram can be found by drawing a straight line from the pre-test probability (50%) on the left-hand side of the nomogram passing through the likelihood ratio (7.5) in the middle.

NOMOGRAM



Likelihood ratios are more powerful measures for the assessment of the clinical value of a diagnostic test. Contrary to sensitivity, specificity and predictive values, likelihood ratios can be determined for more than two categories. For continuous variables when different cut-off points are available, it is possible to calculate the probability of having the disease compared to the probability of not having the disease for each individual cut-off.

APPENDIX 7

DATA ABSTRACTION FORM FOR DIAGNOSTIC STUDIES

Author(s)

Journal Year of publication

Reviewer Date of data abstraction

Study characteristics

Design of the study: *Cross-sectional Randomised trial Cohort Case Control Systematic review Other……..*

Data collection: *Prospective Retrospective Unclear Unreported*

Clinical problem:

Type of comparison:

Blind assessment of tests: *Yes No Unclear*

Number of patients enrolled:

Number of patients who underwent the diagnostic test:

Number of patients who underwent the reference standard:
- number of test positive subjects who underwent the reference standard
- number of test negative subjects who underwent the reference standard

Description of the diagnostic test: *Adequately reported Unclearly reported Inadequately reported Unreported*

Type of test
Model *Manufacturer*
Scan Time
Slice Thickness
Contrast medium

Description of the reference standard: *Adequately reported* *Unclearly reported* *Inadequately reported* *Unreported*

Type of test
Model *Manufacturer*
Scan Time
Slice Thickness
Contrast medium

Results

	Disease Present +	Disease Absent -	Totals
Test Positive +	a	b	a +b
Test Negative –	c	d	c+d
Totals	a+c	b+d	a+b+c+d

True positives = (a) = False positives = (b) =

False negatives = (c) = True negatives = (d) =

Sensitivity = a/(a+c) =

Specificity = d/(b+d) =

Positive predictive value = a/(a+b) =

Negative predictive value = d/(c+d) =

Likelihood ratio for a positive test result = (sensitivity/1-specificity) =

Likelihood ratio for a negative test result = (1-sensitivity/specificity) =

Observed pre-test probability = (a+c)/(a+b+c+d)

Pre-test odds = prevalence/1-prevalence

Post-test odds = pre-test odds x likelihood ratio

Post-test probability = post-test odds/(post-test odds +1)

Accuracy = (a+d)/(a+b+c+d) =

APPENDIX 8

EXAMPLE OF GUIDELINE, REFERENCE AND EVIDENCE TABLES FOR MBUR5

A Guideline Table:

A	**Clinical Problem**	Neck pain, brachalgia, degenerative change
A1	**Ref No. in MBUR4 (if any)**	C04
A2	**Section in MBUR4**	C. The Spine: Cervical
B	**Search strategy: databases used; period; MeSH headings; other key words**	MEDLINE, EMBASE 1966 to present. Cervical vertebrae, intervertebral disc displacement, Magnetic resonance imaging, tomography xray computed. Also Internet sites AHCPR, Canadian Medical Association Clinical Practice Guidelines Database, Center for Disease Control and Prevention
C	**Search results: no. found; no. used.**	295 found, 13 used
D	**Reference numbers of cited references in attached master list**	
E	**Summarised results from each of the cited references**	[3] [4] and [5] are reviews. Hedberg et al [6] looked retrospectively at 130 patients who had MR for suspected radiculopathy 20 had myelography and 13 went to surgery. They found a good correlation between MRI and surgical findings but this study does not assess MR false negatives. Brown et al [7] retrospectively looked at the MRIs of 34 surgical patients. 28 of these had CT and/or myelography. The surgical finding were correctly predicted in 88% (MRI), 81% (CT myelo), 58% (myelo) and 50% (CT). Neuhold et al [8] prospectively performed MR and Myelography in a large group of patient with neck signs 30 of whom went to surgery. MR detected the clinically relevant segment in 29/30, myelo in 28/30. Wilson et al [9] looked retrospectively at the MRIs of 40 surgical patients 13 of whom also had CT myelography. The CT myelography did not add any new information. Perneczky et al [10] prospectively performed MRI and CT myelography on 63 surgical patients. Both investigations had a diagnostic accuracy of 95%. MRI tended to miss small laterally protruding disk fragments and myelography to underestimate severity of central discs. Bartlet et al [11] prospectively performed MRI and CT myelography on 23 patients. Although they found diagnostic accuracies comparable to Perneczky et al [10], they argued that apparently high accuracy can be seen with an unacceptably high number of inappropriate surgery/inaction. Specifically, MRI with 4mm slices can miss lateral disc protrusions. Schellhas et al [12] performed MR and discography in 10 patients and 10 controls and found 17/20 and 10/11 MR 'normal ' discs to have tears. 2 of these were painful on discography.

F	**Statement (= the conclusion drawn from E)**	Consider MRI and specialist referral when pain affecting lifestyle or when there are neurological signs. Myelography (with CT) may occasionally be required to provide further delineation or when MRI is unavailable or impossible.
G	**Investigation**	MRI
H	**Recommendation ***	5
I	**Grade of Recommendation A - C**	C
J	**Comment (if any) to go in booklet version of *MBUR5***	
K	**Any other comments on this problem (e.g. caveats; suggestions for research or systematic review; cost or opportunity cost; users' views.**	Many reports but few prospective studies. Most studies have been done on a selected population or surgical patients. Gold standard is surgery but this is only available in a minority of (presumably severe) cases. For similar reasons, CT myelography (and discography) are not always part of the diagnostic work up. MRI has been adopted as 1 stop test because it is non invasive.

* 1 Indicated
2 Not indicated initially
3 Not indicated routinely
4 Not indicated
5 Specialised investigation

B1 Reference table of papers (for all problems studied):

Ref ID	Authors	Title	Journal	Date	Vol	Pages
1	Irvine DH, et al	Prevalence of cervical spondylosis in a general practice	Lancet	1965	1	1089-1091
2	Gore DR, et al	Neck pain: a long term follow-up study of 205 patients	Spine	1987	12	1-5
3	Russell EJ	Cervical disk disease	Radiology	1990	177	313-325
4	Ruggieri PM	Cervical radiculopathy	Neuroimaging Clinics of N America	1995	5	349-366
5	Russell EJ	Computed Tomography and myelography in the evaluation of cervical degenerative disease	Neuroimaging Clinics of N America	1995	5	329-348
6	Hedberg MC et al	Gradient Echo (GRASS) MR Imaging in cervical radiculopathy	AJR	1988	150	683-689
7	Brown BM et al	Preoperative evaluation of cervical radiculopathy and myelopathy by surface-coil MR imaging	AJR	1988	151	1205-1212
8	Neuhold A et al	Combined use of spin-echo and gradient-echo MRI-imaging in cervical disk disease: comparison with myelography and intraoperative findings	Neuroradiology	1991	33	422-426

9	Wilson DW et al	Magnetic Resonance Imaging in the preoperative evaluation of cervical radiculopathy	Neurosurgery	1991	28	175-179
10	Perneczky G et al	Diagnosis of cervical disk disease	Acta Neurochirurgica	1992	116	44-48
11	Bartlett RJV et al	Two-dimensional MRI at 1.5 and 0.5 T versus CT myelography in the diagnosis of cervical radiculopathy	Neuroradiology	1996	38	142-147
12	Schellhas KP et al	Cervical discogenic pain: prospective correlation of magnetic resonance imaging and discography in asymptomatic subjects and pain sufferers	Spine	1996	21	300-312
13	Van de Kelft E et al	Diagnostic imaging algorithm for cervical soft disc herniation	JNNP	1994	57	724-728

C Evidence table

Ref ID	**Author; Year; Country; Level (I-V).**	**Aims**	**Patient Population**	**Study design*;**	**Results**	**Comments**
1	Irvine; 1965 III	Prevalence study	General practice	Observational	In 3rd decade 13% men show cervical spondylosis on plain neck x-rays. At 70 years, 100% men show cervical spondylosis	Cervical spondylosis is very common in the general population, so finding it is often not helpful for clinical management
2	Gore; 1987 III	Monitoring outcomes of patients with neck pain	205 patients with neck pain		79% improved, 43% resolved, 32% persistent pain	
3	Russell 1990 USA III-IV	Review				
4	Ruggieri 1995 USA III-IV	Review				
5	Russell 1995 USA III-IV	Review				

6	Hedberg 1988 USA III	Assess usefulness of MR for suspected cervical radiculopathy	130 pts suspected cervical radiculopathy.	Retrospective. All had MRI, 20 myelograms, 13 surgery	Good correlation MRI and surgical findings	No assessment of MRI false negatives because none of these pts had CTM
7	Brown 1988 USA III	MRI for myelopathy and radiculopathy	34 patients prior to surgery	Retrospective. 34 MRI, 28 CT and/or myelography	Correct predictions of surgical findings MR88%, CT myelo 81%, myelo 58%, CT 50%.	Selected population. MR replaced invasive evaluations in 32% of preoperative patients.
8	Neuhold 1991 Austria III	MR, myelography and surgery	30 patients with neck signs who subsequently underwent surgery	Prospective. Patients had MR, myelography	Clinically relevant segment detected by MR (29/30) and myelo (28/30)	
9	Wilson 1991 USA III	MR for preoperative patients	40 patients before surgery	Retrospective all had surgery, 27 MRI only, 13 had MRI and CT myelography	CTM didn't show anything not seen on MRI	
10	Perneczky 1992 Austria III	MRI vs myelography	63 surgical patients	Prospective. All had both Ix	Both had diagnostic accuracy of 95%. MRI tended to miss small laterally protruding disk fragments, myelo to underestimate severity of central discs	Myelography still has a place where symptoms and signs do not agree with MR data
11	Bartlett 1996 UK III	MRI vs myelography	23 patients with cervical spondylosis	Prospective. All had both Ix	MRI with 4mm slices inadequate for presurgical assessment of root lesions	Myelography still has a place where symptoms and signs do not agree with MR data
12	Schellhas 1996 USA Iia	MR vs discography	10 chronic neck pain patients vs 10 asymptomatic controls	Prospective	17/20 MR 'normal' disks had painless tears on discography (asymptomatic group) In symptomatic group, 10/11 MR 'normal' discs had tears - 2 of these were painful	Significant annular tears may not be apparent on MRI

13	Van de Kelft 1994 Belgium	Justify an algorithm	100 patients with radiculopathy	All had plain XR. Those with no osteophytes or instability had MRI, others had CT myelography	MRI (n=59) showed disk herniation in 55. CT myelo performed in 4 pts with normal MRI found one further disc. Results of other CT myelos not formally presented	Not sure the results fully support the conclusions

* e.g. RCT
Comparison (pro/retro-spective)
Series (pro/retro-spective)
Audit

Source: Templates for MBUR5 by Chris Squire, Royal College of Radiologists, London.

The information reported in the above tables is merely used as an example and does not necessarily correspond to that of the original published papers.

GLOSSARY

Accuracy
The degree to which indices of test performance measure the precision and validity of the diagnostic test.

Blinding
Blinding or masking is the process used in epidemiological studies and clinical trials by which study participants, investigators and/or outcome assessors are unaware of the intervention participants are receiving.

Bayes' theorem*
A probability theorem used to obtain the probability of a condition in a group of people with some characteristic (e.g. exposed to an intervention of interest, or with a specified result on a diagnostic test) on the basis of the overall rate of that condition (the prior probability) and the likelihood of that characteristic in people with and without the condition.

Case-control study*
A study that starts with identification of people with the disease or outcome of interest (cases) and a suitable control group without the disease or outcome. The relationship of an attribute (intervention, exposure or risk factor) to the outcome of interest is examined by comparing the frequency or level of the attribute in the cases and controls. For example, to determine whether thalidomide caused birth defects, a group of children with birth defects (cases) could be compared to a group of children without birth defects (controls). The groups would then be compared with respect to the proportion exposed to thalidomide through their mothers taking the tablets. Case-control studies are sometimes described as being retrospective as they are always performed looking back in time.

Case series*
An uncontrolled observational study involving an intervention and outcome for more than one person.

Clinical decision rules
A set of validated clinical rules used to increase the accuracy of clinicians' diagnostic and prognostic assessments. Different aspects of the history, medical examination, laboratory and imaging investigations of patients contribute to the choice of the clinical decision rules.

Cohort study*
An observational study in which a defined group of people (the cohort) is followed over time. The outcomes of people in subsets of this cohort are compared, to examine for example people who were exposed or not exposed (or exposed at different levels) to a particular intervention or other factor of interest. A cohort can be assembled in the present and followed into the future (this would be a prospective study or a "concurrent cohort study"), or the cohort could be identified from past records and

followed from the time of those records to the present (this would be a retrospective study or a "historical cohort study").

Cross-sectional study*
A study that examines the relationship between diseases (or other health related characteristics) and other variables of interest as they exist in a defined population at one particular time. The temporal sequence of cause and effect cannot necessarily be determined in a cross-sectional study.

Gold standard/Reference standard
The method, procedure or measurement which is commonly considered to be the best available.

Hierarchy of evidence
A way of grouping study designs according to their validity. Ideally the hierarchy indicates which studies should be assigned more weight. In the diagnostic literature, well-designed comparative prospective studies in which all participants undergo the new test as well as the reference ('gold') standard and the results are independently and blindly assessed by at least two investigators, are seen as being at the top of the hierarchy.

Levels of evidence
Method used to grade the quality and strength of evidence.

Likelihood ratios
The likelihood that a positive or negative test result would be foreseen in a patient with, versus a patient without, the target disorder.

MEDLINE* (electronic version of the *Index Medicus*)
An electronic database produced by the United States National Library of Medicine. It indexes millions of articles in selected (about 3,700) journals. It is available through most medical libraries, and can be accessed on CD-ROM, the Internet and by other means. Years of coverage: 1966 - present.

MESH headings (Medical Subject Headings)*
Terms used by the United States National Library of Medicine to index articles in *Index Medicus* and MEDLINE. Designed to reduce problems that arise from, for example, differences in British and American spelling. The MeSH system has a tree structure in which broad subject terms branch into a series of progressively narrower subject terms.

Meta-analysis
A systematic review which includes a mathematical synthesis of the results.

Negative predictive value (or post-test probability of not having the target disease)
The proportion of people with true negative results over all negative test results.

Positive predictive value (or post-test probability of the target disease)
The proportion of people with true positive test results over all positive test results.

Prevalence*
The number of existing cases of a particular disease or condition in a given population at a designated time.

Randomisation*
Method used to generate a random allocation sequence, such as using tables of random numbers or computer-generated random sequences. The method of randomisation should be distinguished from concealment of allocation because there is a risk of selection bias despite the use of randomisation, if the allocation concealment is inadequate. For instance, a list of random numbers may be used to randomise participants, but if the list is open to the individuals responsible for recruiting and allocating participants, those individuals can influence the allocation process, either knowingly or unknowingly.

Randomised controlled clinical trial (RCCT)*
An experiment in which investigators randomly allocate eligible people into intervention groups to receive or not to receive one or more interventions that are being compared. The results are assessed by comparing outcomes in the treatment and control groups.

Search strategy
Method used to identify studies in the literature.

Sensitivity of a test is the proportion of people with the target disorder who have a true positive test result.

Specificity of a test is the proportion of people without the disease who have a true negative test result.

Systematic review/systematic overview*
A review of a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the included studies.

** from: The Cochrane Library, Issue 4, 2000. Oxford: Update Software.*