BMC Medical Research Methodology

Research article

Conducting systematic reviews of diagnostic studies: didactic guidelines

Walter L Devillé*[1,2], Frank Buntinx[3], Lex M Bouter[1], Victor M Montori[4], Henrica CW de Vet[1], Danielle AWM van der Windt[1] and P Dick Bezemer[1,5]

Address: [1]Institute for Research in Extramural Medicine (EMGO Institute), VU University Medical Centre, Amsterdam, The Netherlands, [2]Programme Migrant Health, Netherlands Institute for Health Services Research (Nivel), Utrecht, The Netherlands, [3]Department of General Practice, Catholic University Leuven, Leuven, Belgium & Department of General Practice, University of Maastricht, Maastricht, The Netherlands, [4]Division of Endocrinology, Metabolism, Nutrition and Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA and [5]Department of Clinical Epidemiology and Biostatistics, VU University Medical Centre, Amsterdam, The Netherlands

E-mail: Walter L Devillé* - w.deville@nivel.nl; Frank Buntinx - frank.buntinx@med.kuleuven.ac.be; Lex M Bouter - lm.bouter.emgo@med.vu.nl; Victor M Montori - montori.victor@mayo.edu; Henrica CW de Vet - hcw.devet.emgo@med.vu.nl; Danielle AWM van der Windt - dawm.van_der_windt.emgo@med.vu.nl; P Bezemer - pd.bezemer.biostat@med.vu.nl

*Corresponding author

Published: 3 July 2002

BMC Medical Research Methodology 2002, **2**:9

Received: 21 December 2001
Accepted: 3 July 2002

This article is available from: http://www.biomedcentral.com/1471-2288/2/9

Abstract

Background: Although guidelines for critical appraisal of diagnostic research and meta-analyses have already been published, these may be difficult to understand for clinical researchers or do not provide enough detailed information.

Methods: Development of guidelines based on a systematic review of the evidence in reports of systematic searches of the literature for diagnostic research, of methodological criteria to evaluate diagnostic research, of methods for statistical pooling of data on diagnostic accuracy, and of methods for exploring heterogeneity.

Results: Guidelines for conducting diagnostic systematic reviews are presented in a stepwise fashion and are followed by comments providing further information. Examples are given using the results of two systematic reviews on the accuracy of the urine dipstick in the diagnosis of urinary tract infections, and on the accuracy of the straight-leg-raising test in the diagnosis of intervertebral disc hernia.

Background

Systematic reviews and meta-analyses of studies evaluating the accuracy of diagnostic tests (we will refer to them generically as diagnostic systematic reviews) are appearing more often in the medical literature [1,2]. Of the 26 reviews on diagnostic tests published between 1996 and 1997, 19 were systematic reviews or meta-analyses[2]. In the field of clinical chemistry and haematology, 23 of 45 reviews published between 1985 and 1998 were systematic reviews [3]. Although guidelines for critical appraisal of diagnostic research and meta-analyses have already been published [1,4–9], these may be difficult to understand for clinical researchers or do not provide enough detailed information.

We want to present a set of practical guidelines, based on evidence and the expertise of the Cochrane Collaboration, to facilitate understanding of and appropriate adherence

to methodological principles when conducting diagnostic systematic reviews.

We reviewed reports of systematic searches of the literature for diagnostic research [10–13], methodological criteria to evaluate diagnostic research [1,4–9], methods for statistical pooling of data on diagnostic accuracy [9,14–22], and methods for exploring heterogeneity [23–27].

Guidelines for conducting diagnostic systematic reviews are presented in a stepwise fashion and are followed by comments providing further information. Examples are given using the results of two systematic reviews on the accuracy of the urine dipstick in the diagnosis of urinary tract infections[28], and on the accuracy of the straight-leg-raising test in the diagnosis of intervertebral disc hernia[29]. Still, clinical readers are advised to look for assistance of a statistician when it comes to pooling.

Table 1: Search strategy in PubMed (MEDLINE) for publications about the evaluation of diagnostic accuracy.

((((((((((("sensitivity and specificity"[All Fields] OR "sensitivity and specificity/standards"[All Fields]) OR "specificity"[All Fields]) OR "screening"[All Fields]) OR "false positive"[All Fields]) OR "false negative"[All Fields]) OR "accuracy"[All Fields]) OR ((((("predictive value"[All Fields] OR "predictive value of tests"[All Fields]) OR "predictive value of tests/standards"[All Fields]) OR "predictive values"[All Fields]) OR "predictive values of tests"[All Fields])) OR (("reference value"[All Fields] OR "reference values"[All Fields]) OR"reference values/standards"[All Fields])) OR ((((((((((("roc"[All Fields] OR "roc analyses"[All Fields]) OR "roc analysis"[All Fields]) OR "roc and"[All Fields]) OR "roc area"[All Fields]) OR "roc auc"[All Fields]) OR "roc characteristics"[All Fields]) OR "roc curve"[All Fields]) OR "roc curve method"[All Fields]) OR "roc curves"[All Fields]) OR "roc estimated"[All Fields]) OR "roc evaluation"[All Fields])) OR "likelihood ratio"[All Fields]) AND notpubref [sb]) AND "human"[MeSH Terms])

The guidelines

• *How to search the literature for studies evaluating the accuracy of diagnostic tests*

A systematic review should include all available evidence, so a systematic and comprehensive search of the literature is needed. The reviewer has to design a search strategy based on a clear and explicit description of the subjects receiving the test of interest, the diagnostic test and its accuracy estimates, the target disease, and the study design. These elements are usually specified in the criteria for inclusion of primary studies in the review. The search will include electronic literature databases. However, because computerised databases only index a subset of all the available literature, the search should be extended using other resources (including reference checking and consultation of experts, as explained below) [11].

The search to identify primary studies may follow the next basic but labour-intensive steps:

1) A computer-aided search of MEDLINE (PUBMED website [http://www.ncbi.nlm.nih.gov/entrez], EMBASE and other databases. A search strategy begins by creating a list of database-specific keywords and text words that describe the diagnostic test and target disease of interest (subject-specific strategy). Because the number of diagnostic accuracy studies is often small, the subject-specific strategy usually yields a limited number of publications to be screened[12]. An accurate search strategy for diagnostic publications (generic strategy) was recently published [13] and can be combined with the subject-specific strategy if the number of publications resulting from the latter is large. We found a combination of two published generic strategies adapted for use in PubMed (MEDLINE) to be more sensitive and precise than previously published strategies[10,12] (Table 1). Each electronic database will need to be searched using a specially designed search strategy.

2) The reference section of primary studies, narrative reviews, and systematic reviews should be reviewed to search for additional primary studies that could have been missed by the electronic search. Identification methods for systematic reviews have also been published [30]. The MEDION database available at the University of Maastricht, the Netherlands, collects some 250 published reviews of diagnostic and screening studies. It is available through berna.schouten@hag.unimaas.nl and will shortly be published on the Internet.

3) Consultation of experts in the disease of interest to identify further published and unpublished primary studies. As diagnostic accuracy studies are often based on routinely collected data, publication bias may be more prevalent in diagnostic than in therapeutic research [19].

Comments

The first step in a literature search is the identification of relevant publications. Diagnostic research reports, older publications in particular, are often poorly indexed in the electronic databases. It is often fruitful to conduct pilot searches using the subject-specific strategy. This process is iterated after identifying and incorporating additional keywords and text words used to describe and index the retrieved reports. Studies found only on the reference sections of the retrieved reports but missed by the search strategy should be searched in the database using the articles' title or first author's name. If a study is found in the database, its keywords should be noted and added to the strategy. Citation tracking may provide additional studies. The Science Citation Index could be searched forward in time to identify articles citing relevant publications[31].

Once the search is completed, two independent reviewers should screen the titles and abstracts of the identified citations by using specific pre-specified inclusion criteria. The inclusion criteria can be pilot tested on a sample of articles. If disagreements cannot be resolved by consensus or if insufficient information is available, a third reviewer and/or the full papers should be consulted.

• *Criteria for inclusion of studies*

Reference test

The accuracy of a diagnostic or screening test should be evaluated by comparing its results with a 'gold standard', criterion standard or reference test accepted as the best available by content experts. The reference test may be a single test, a combination of different tests, or the clinical follow-up of patients [22]. The publication should describe the reference test since it is a *conditio sine qua non* for the evaluation of a diagnostic test.

Population

Detailed information about the participants in diagnostic research is often lacking. Participants should be defined explicitly in terms of age, gender, complaints, signs and symptoms, and their duration. At least, a definition of participants with and without the disease, as determined by the reference test, should be available. The minimal number of participants needed with and without the disease depends of the type of study, the estimates of diagnostic accuracy, and the precision used to estimate these parameters [32].

Outcome data

Information should be available to allow the construction of the diagnostic 2 by 2 table with its four cells: true positives, false negatives, false positives and true negatives.

Language

If a review is limited to publications in certain languages, it should be reported.

Comments

As the patient mix (spectrum of disease severity) is different at different levels of care, a diagnostic review may focus on a specific setting (primary care, etc.) or include all levels. This information may be important for subgroup analyses in case of heterogeneity. If test characteristics have changed over the years as a result of changing methods or technological evolution, one may consider the inclusion of only studies that used the new version of the testing [33]. However, as advocated for systematic reviews of trials by the Cochrane Collaboration, all published research may be included: in the analysis the association with time may be studied and explained. All evidence available should be reviewed regardless of their language of publication. It is not easy to identify non-English publications, as they are often not indexed in computerised databases. In the field of intervention research, there is some evidence of bias when excluding non-English publications [34]. Our research on the accuracy of the urine dipstick revealed differences in the methodological validity between European and American studies, but these differences had no effect on accuracy. Although large samples are no guarantee against selective patient sampling, small samples seldom result from a consecutive series of patients or a random sample, but often constitute a convenience sample. Small samples are, therefore, very vulnerable to selection bias.

• *Methodological quality*

The methodological quality of each selected paper should be assessed independently by at least two reviewers. Chance-adjusted agreement should be reported and disagreements solved by consensus or arbitration. To improve agreement, reviewers should pilot their quality assessment tools in a subset of included studies or studies evaluating a different diagnostic test.

Validity criteria for diagnostic research have been published by the Cochrane Methods Group on Screening and Diagnostic Tests [35] [http://www.cochrane.org/cochrane/sadt.htm], and by other authors [4–6], [8,9]. Criteria, assessing internal and external validity, should be coded and described explicitly in the review (Table 2). The internal validity criteria refer to study characteristics that safeguard against the intrusion of systematic error or bias. External validity criteria provide insight into the generalisibility of the study and judge if the test under evaluation was performed according to accepted standards. Internal and external validity criteria, describing participants, diagnostic test and target disease of interest, and study methods may be used in meta-analysis to assess the overall 'level of evidence' and in sensitivity and subgroup analyses (see Data Extraction and Data Analysis sections).

It is important to remember that studies may appear to be of poor methodological quality either because they were poorly conducted or poorly reported. Methodological appraisal of the primary studies is frequently hindered by lack of information. In these instances, reviewers may choose to contact the studies' authors or score items as 'don't know or unclear'.

Example

A urine dipstick is usually read before the material is cultured. So, it can be interpreted that the dipstick was read without awareness of the results of the urine culture. However, the culture (reference test) may be interpreted with full awareness of the results of the dipstick. If blinding is not explicitly mentioned, reviewers may choose to score this item as 'don't know' or 'diagnostic test blinded

Table 2: List of validity criteria operationalised for papers reporting on the accuracy of urine dipsticks in the diagnosis of Urinary Tract Infections (UTI) or Bacteriuria

Criteria of internal validity (IV)	***Positive score***
Valid reference standard	(semi-)quantitative (2 points) ot dipslide culture (1 point)
Definition of cut-off point for reference standard	definition of Urinary Tract Infection/Bacteriuria by colony forming units per ml (1 point)
Blind measurement of index test and reference test	in both directions (2 points) or only index or reference test
Avoidance of verification bias	assessment by reference standard independent from index test results (1 point)
Index test interpreted independently of all clinical information	explicitly mentioned in the publication or urine samples from mixed out-patient populations examined in a general laboratory (1 point)
Design	prospective (consecutive series) (1 point) or retrospective collection of data (0 points)
Criteria of external validity (EV)	
1 Spectrum of disease	in- and/or exclusion criteria mentioned (1 point)
2 Setting	enough information to identify setting (1 point) (community through tertiary care)
3 Previous tests/referral filter	details give about clinical and other diagnostic information as to which the index test is being evaluated (symptomatic or asymptomatic patients (1 point)
4 Duration of illness before diagnosis	duration mentioned (1 point)
5 Co-morbid conditions	details given (type of population) (1 point)
6 Demographic information	age (1 point) and/ or gender (1 point) data provided
7 Execution of index test	information about standard procedure directly or indirectly available, urine collection procedure, first voided urine, distribution of micro-organisms, procedure of contaminated urine samples, time of transportation of urine sample, way of reading index test, persons reading index test (1 point each)
8 Explicitation of cut-off point of index test	trace, 2 or more + (1 point if applicable)
9 Percentage missing	if appropriate: missings mentioned (1 point)
10 Reproducibility of index test	reproducibility studied or reference mentioned (1 point)

Blinding (IV3): When information about *blinding* of measurements was not given and the dipstick was performed in another setting than the culture, we assumed blind assessment of the index test versus the reference test but not vice versa. *Explicitation of the cut-off point* (EV8) was only necessary for the leukocyte-esterase measurement.

for reference test' (implicitly scoring the reference test as not blinded). Or, the authors may be contacted for clarification.

A survey of the diagnostic literature from 1990 through 1993 in a number of peer-reviewed journals, showed that only a minority of the studies satisfied methodological standards [7]. There is some evidence that inadequate methods may have an impact on the reported accuracy of a diagnostic test: Lijmer [2] screened diagnostic meta-analyses published in 1996 and 1997, and showed that the diagnostic accuracy of a test was overestimated in studies 1) with a case-control design; 2) using different reference tests for positive and negative results of the index test; 3) accepting results of observers that were unblinded to the index test results when performing the reference test; 4) that did not describe diagnostic criteria for the index test; and 5) where participants were inadequately described.

Comments

Ideally, all participants should be submitted to the same reference test. Sometimes different groups of patients are submitted to different reference tests, but details are not given. In this case, it is important to assess if the different reference tests are recognised by experts as being adequate. In some conditions results of the index test may be incorporated into the diagnostic criteria, what may lead to incorporation bias and overestimation of accuracy [36]. Verification or work-up bias may be present if not all participants who received the index test, are referred to the reference test(s). Verification bias is present if the participants are referred according to the index test results. That is usually the case in screening studies where only subjects with positive index test results receive the reference test, so that only a positive predictive value can be calculated. Estimation of accuracy will not be possible in these studies unless complete follow-up registries are available. In case-control studies the contrast between diseased and non-diseased may be artificially sharpened by sampling only persons who are clearly diseased and persons who are completely healthy, resulting overestimated sensitivity and specificity[36].

• Data extraction

Two reviewers should independently extract the required information from the primary studies. Detailed information has to be extracted about the participants included in

the study, time of data collection and the testing procedures. The cut-off point used in dichotomous testing and the reasons and the number of participants excluded because of indeterminate results or unfeasibility is always required.

Example
Detailed information extracted in the case of the dipstick meta-analysis: mean age, male/female ratio, different cut-off points for leukocyte-esterase (trace, 2+, 3+), time needed for transportation, if indeterminate results were excluded, included as negative or repeated.

As the information extracted may be used in subgroup analyses and statistical pooling of the validity, possible sources of heterogeneity should be defined based on a priori existing evidence or hypotheses.

Example
In the dipstick meta-analysis we hypothesised that the following factors may explain heterogeneity if present: procedures of collection of material for the test (method of urine collection, delay between urine collection and culture), who was executing the test and how (manually or automatic), and different brands of commercial products.

Accuracy may be presented in different ways. For the meta-analysis of dichotomous tests (see below) it is necessary to construct the diagnostic 2×2 table: absolute numbers in the four cells are needed. Totals of 'diseased' and 'non-diseased' participants are needed to calculate prior probability (pre-test probability), and to reconstruct the 2×2 table from sensitivity, specificity, likelihood ratios, predictive values or receiver-operator characteristic (ROC) curves. If possible, the 2×2 table should be generated for all relevant subgroups. Further information to extract includes year of publication, language of publication, and country or region of the world where the study was performed.

Comments
A standardised data extraction form may be used simultaneously with but separately from the quality assessment form. This approach facilitates data extraction and comparison between reviewers. The form has to be piloted to ensure that all reviewers interpret data in the same way. Like in other steps of the review were judgements are made, disagreements should be recorded and resolved by consensus or arbitration. Lack of details about test results or cut-off points, inconsequential rounding off of percentages, and data errors require common sense and careful data handling when reconstructing 2×2 tables. If predictive values are presented with sensitivity and specificity in 'diseased' and 'non-diseased' individuals, the calculation of the four cells from sensitivity and specificity can be confirmed by using the predictive values. Details can be requested from the authors of the studies, but these attempts are often unsuccessful, as the raw data may no longer be available.

Example
In a review on the accuracy of the CAGE questionnaire for the diagnosis of alcohol abuse, sufficient data were made available of 9 only out of 22 studies selected, although the authors of the review tried to contact the original authors by all means [37].

• Data analysis
Whether or not meta-analysis – statistical analysis and calculation of a summary diagnostic accuracy estimate – can be conducted depends on the number and methodological quality of primary studies included and the degree of heterogeneity of their estimates of diagnostic accuracy. Because diagnostic accuracy studies are often heterogeneous and present limited information it is typically difficult to complete a meta-analysis. If heterogeneity is identified, important information is obtained from attempts to explain it. For instance, the effect that each validity criterion has on the estimates of diagnostic accuracy and the influence of *a priori* defined study characteristics should be explored as potential explanations of the observed study-to-study variation [23–27]. If meta-analysis is not possible or advisable, the review can be limited to a qualitative descriptive analysis of the diagnostic research available (best evidence synthesis) [38].

Several meta-analytic methods for diagnostic research have been published in the last decennium [14–22]. For the analysis we recommend the following steps: 1-presentation of the results of individual studies, 2-searching for the presence of heterogeneity, 3-testing of the presence of an (implicit) cut-point effect, 4-dealing with heterogeneity, 5-deciding which model should be used if statistical pooling is appropriate and 6-statistical pooling.

• Describing the results of individual studies
Reporting the main results of all included studies is an essential part of each review. It provides the reader the outcome measures and gives at a first glance insight in their heterogeneity. Each study is presented with some background information (year of publication, geographical region, number of diseased and non-diseased patients, selection of the patients, methodological characteristics) and a summary of the results. In view of the asymmetric nature of most diagnostic tests (some tests are good to exclude a disease, others to confirm), it is important to report pairs of complimentary outcome measures, i.e. at least both sensitivity and specificity, as this is necessary information for readers who would like to reproduce the systematic review. The diagnostic odds ratio (DOR) can

be added, but better not alone, as a same odds ratio can relate to different combinations of sensitivity and specificity. The DOR is a measure for the discriminative power of a diagnostic test: the ratio of the odds of a positive test result among diseased to the odds of a positive test result among the non-diseased.

$$DOR = \frac{sensitivity\ /\ (1 - sensitivity)}{(1 - specificity)\ /\ specificity}$$

For more details regarding the calculation of the DOR based on the study-specific sensitivity and specificity we refer to Littenberg [15], Midgette [16] or Moses [17]. The potential problems associated with sensitivities and specificities of 100% are solved by adding 0,5 to all cells of the diagnostic 2×2 table. Main outcome measures should always be reported with their 95% confidence intervals (CI).

• *Searching for heterogeneity*

When setting inclusion criteria, most reviewers will try to define a more or less homogeneous set of studies. Reality is however, that even then most diagnostic reviews show considerable heterogeneity in the results of included studies. When different studies have largely different results, this may result from either random error or heterogeneity due to differences in clinical or methodological characteristics of studies. A chi-square test or an extension of the Fisher's exact test for small studies [39] can be used to statistically test the presence of heterogeneity in study results. This may offer some guidance, but the power of this test tends to be low. A basic, but very informative method for assessing heterogeneity is to produce a graph in which the individual study-outcomes are plotted, together with their 95% confidence intervals (Figure 1), and subjectively evaluate the variation in study results.

• *Searching for the presence of an (implicit) cut-off point effect*

Estimates of diagnostic accuracy differ if not all studies use the same cut-off point for a positive test result or for the reference standard. The interpretation of test results often depends on human factors (e.g. radiology, pathology, etc) or on the process of testing (e.g. clinical examination). In such cases different studies may use a different implicit cut-off point. Variation in the parameters of accuracy may be partly due to variation in cut-off points. In case of diagnostic tests with a continuous or ordinal outcome the ROC curve presents pairs of sensitivity and specificity for different values of the cut-off point of a test.

One can test for the presence of a cut-off point effect between studies by calculating a Spearman correlation coefficient between sensitivity and specificity of all included studies. If strongly negatively correlated, pairs of parameters represent the same DOR [17]. A strong correlation between both parameters will usually result in a homogeneous logarithmic transformed DOR (lnDOR).

Example

In a systematic review of the urine dipstick studying the accuracy of nitrites for the diagnosis of urinary tract infections [28], sensitivity and specificity were poorly correlated (Spearman $\rho = -0.377$) and highly heterogeneous in 58 studies. So was the lnDOR. Subgroup analysis of the factor 'setting of care' gave the results in Table 3.

The Spearman ρ indicates a strong cut-off effect in the family medicine studies and to a lesser degree in the emergency department studies. Despite heterogeneity of sensitivity and specificity, the pairs of sensitivity and specificity in the 6 family practice studies presented a homogeneous DOR.

Moses et al. [17] mention a Spearman correlation of $\rho < -0.6$, but evidence is still limited (see example above: a ρ of -0.4 results in a homogeneous lnDOR). The test for homogeneity of the lnDOR is described in Fleiss [39]. If the lnDOR of the included studies are homogeneous, a Summary ROC curve (SROC) can be fitted based on the pairs of sensitivity and specificity of the individual studies (see further). If sufficient information is available also the pooling of ROC curves of individual studies will be possible.

4. Dealing with heterogeneity

• *Dealing with heterogeneity*

In many cases the interpretation of observed heterogeneity is the most fascinating and productive part of a meta-analysis. The inspection of the plot of sensitivity, specificity and DOR with their 95% CI may indicate the presence of outliers. In that case the reason for this situation should carefully be examined.

Example

In the straight-leg-raising test review the plots of sensitivity and specificity showed clear heterogeneity, confirmed by statistical testing. The plot of the DOR revealed one outlier study (figure 1).

In such cases an outlier can be excluded and the analysis continued with the homogeneous group of remaining studies. Deviant results should be explored and explained. The decision to exclude outliers is complex and should be handled in the same way as in other fields of research.

Outliers can also be searched by using a Galbraith plot [40]. To construct this plot, the standardised lnDOR = lnDOR/se is plotted (y-axis) against the inverse of the se

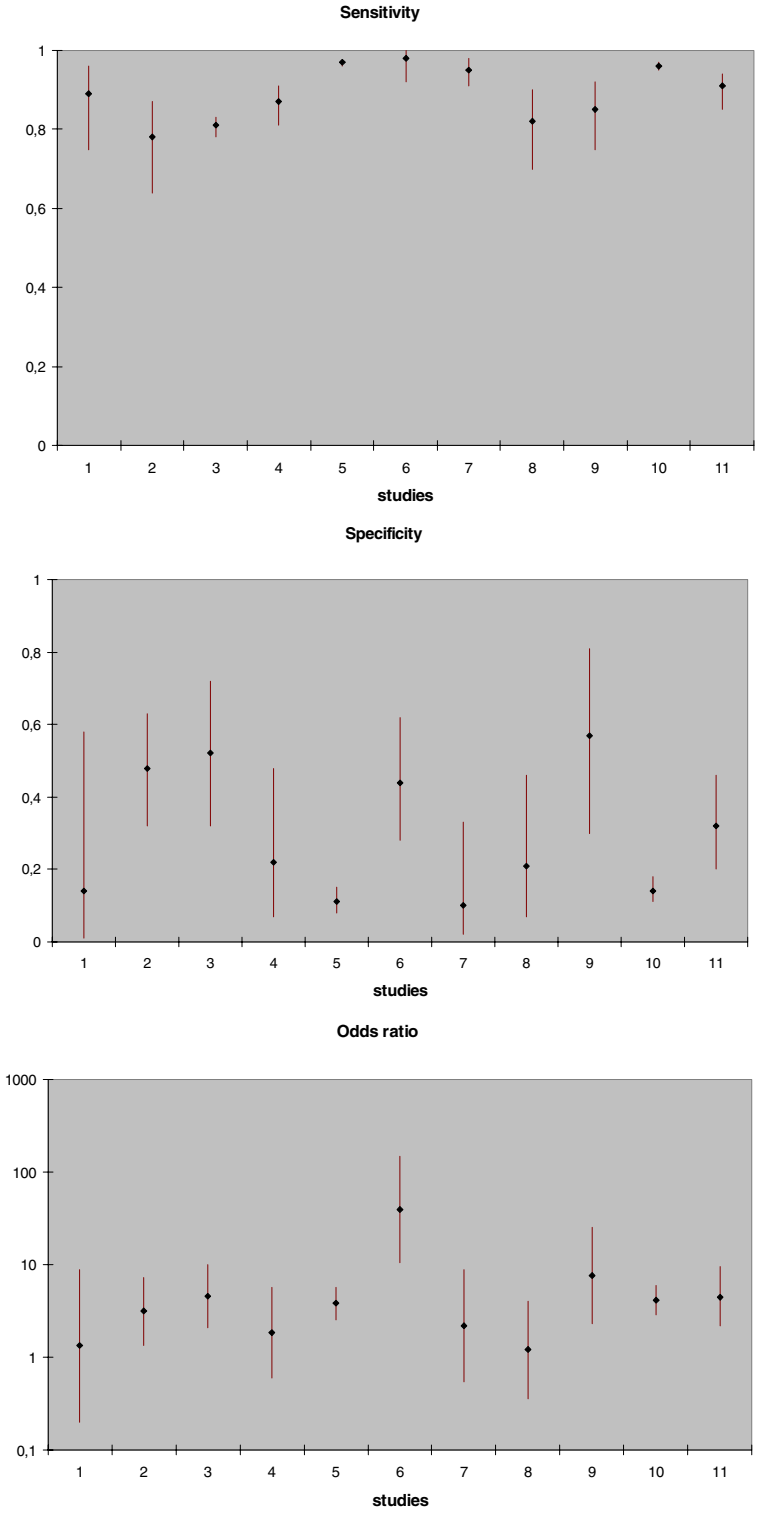

Figure 1
Point estimates (with confidence limits) of respectively sensitivity, specificity, and diagnostic odds ratio of 11 studies on the validity of the test of Lasègue for the diagnosis of disc hernia in low back pain. Study 6 is an outlier

Table 3:

	Family Medicine (n = 6)	Outpatients clinics (n = 17)	Hospital (n = 16)	Emergency (n = 4)
Spearman ρ	-0.714	-0.287	-0.228	-0.400
Sensitivity	heterogeneous	heterogeneous	heterogeneous	homogeneous
Specificity	heterogeneous	heterogeneous	heterogeneous	heterogeneous
lnDOR	homogeneous	heterogeneous	heterogeneous	homogeneous

(1/se) (x-axis). A regression line that goes through the origin is calculated, together with 95% boundaries (starting at +2 and -2 on the y-axis). Studies outside these 95% boundaries may be considered as outliers (Figure 2).

Subgroup analyses defined *a priori* in the protocol could be conducted to detect homogeneous subgroups. Analysis of variance with the lnDOR as dependent variable and categorical variables for subgroups as factors can be used to look for differences among subgroups.

Example

In the dipstick review, sensitivity and specificity were weakly associated (ρ = -0.337) and very heterogeneous. Subgroup analysis showed significant differences of the lnDOR between 6 different 'populations of participants'. In three populations (general population, pregnant women and surgery patients) there was a strong negative association between sensitivity and specificity (ρ = -0.539, -0.559, and -1.00 respectively), yielding homogeneous lnDOR in the three subgroups. Different SROC curves for each subgroup could be fitted (see 6.1.2) (Figure 3).

If many studies are available, a more complex multivariate model can be built in which a number of study characteristics are entered as possible co-variates. Multivariate models search for the independent effect of study characteristics, adjusted for the influence of other, more powerful ones.

• *Deciding on the model to be used for statistical pooling*

Models

There are two underlying models that can be used when pooling the results of individual studies.

A *fixed effect model* assumes that all studies are a certain random sample of one large common study, and that differences between study outcomes only result from random error. Pooling is simple. It essentially consists of calculating a weighted average of the individual study results. Studies are weighted by the inverse of the variance of the outcome parameter of test accuracy.

A *random effect model* assumes that in addition to the presence of random error, differences between studies can also result from real differences between study populations and procedures. The weighting factor is mathematically more complex, and is based on the work of Der Simonian and Laird, initially performed and published for the meta-analysis of trials [41]. It includes both within-study and between-study variation.

A more detailed description about these models can be found in Rutter and Gatsonis [20,42].

Homogeneous studies

If sensitivity and specificity are homogeneous, and if they show no (implicit) cut-off effect (see above), they can be pooled and a fixed effect model can be used. If there is evidence of a cut-off effect, SROC curves can be constructed (see below) or ROC curves can be pooled.

Heterogeneous studies

If heterogeneity is present, the reviewer has the following options:

1. Refrain from pooling and restrict the analysis to a qualitative overview.

2. Sub-group analysis if possible on prior factors and pooling within homogeneous sub-groups.

3. As a last resort, pooling can be performed, using methods that are based on a random effect model.

In view of the low methodological quality of most of the diagnostic studies that have been carried out, there is a tendency to advise using random effect models for the pooling of all diagnostic studies, even if there is no apparent heterogeneity.

• ***Statistical pooling*** (please see additional file 1: appendix statistical formulae for further information)

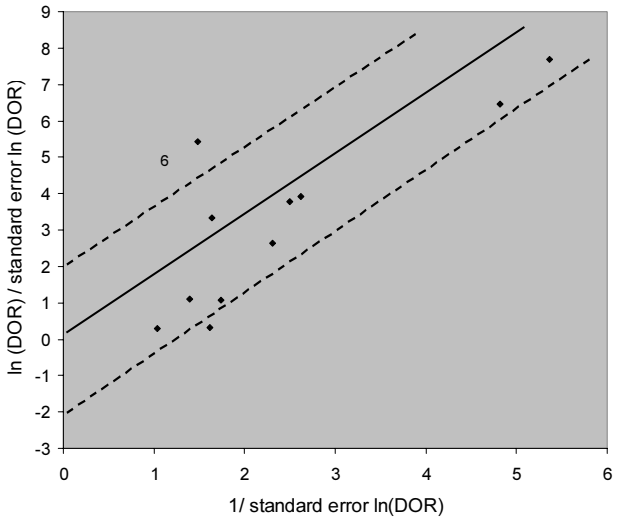

Figure 2
Galbraith plot of 11 studies on the validity of the test of Lasègue for the diagnosis of disc hernia in low back pain. Study 6 is an outlier.

1. Pooling of proportions

1.1. Homogeneous sensitivity and/or specificity

If *fixed effect* pooling can be used, pooled proportions are the average of all individual study results, weighted for the sample sizes. This is easily done by adding together all numerators and dividing the total by the sum of all denominators [16].

1.2. Cut-off point effect: SROC curve

The SROC curve is presented with sensitivity on the y-axis and 1-specificity on the x-axis (ROC plot). The SROC curve differs from the ROC curve in primary diagnostic research, as each study provides one value of sensitivity and one value of specificity (Figure 3). If a SROC curve can be fitted, a regression model is used, with the natural logarithm of the DOR (lnDOR) of the primary research as dependent variable and two parameters as independent variables: one for the intercept (to be interpreted as the mean lnDOR) and one for the slope of the curve (as an estimate of the variation of the lnDOR across the studies due to threshold differences). Details and formula for fitting the curve can be found in the paper presented by Littenberg and Moses [15]. Co-variates representing different study characteristics or pre-test probabilities can be added to the model to examine any possible association of the diagnostic odds ratio with these variables [43]. This type of analysis is usually referred to meta-regression, and refers to (multivariate) regression of the summary estimate DOR of primary research as the dependent variable, and characteristics of the included studies as independent variables. The pooled lnDOR and confidence bounds have to be back-transformed into a diagnostic odds ratio and its confidence intervals. This meta-regression model can be unweighted or weighted, using the inverse of the variance as the weighting factor. The often-negative association of the weighting factor with the lnDOR gives studies with lower discriminative diagnostic odds ratios – because of lower sensitivity and/or specificity – a larger weight. This may be a problem when comparing the pooled accuracy of different tests, and has not yet been solved [19,21].

2. Pooling of likelihood ratios

Continuous test results can be transformed into likelihood ratios[9] obtained by using different cut-off points. Individual data-points from the selected studies can be used to calculate result-specific likelihood ratios[44], which can be obtained by logistic modelling. The natural log posterior odds is converted into a log likelihood ratio by adding a constant to the regression equation. The constant adjusts for the ratio of the number of 'non-diseased' to 'diseased' participants in the respective studies [19].

If primary studies present ordinal test results and they use the same number of categories, ROC curves can be constructed for each study and pooled as below or by using ordinal regression methods [19,45]

3. Pooling of the ROC curves

Results of diagnostic studies with a dichotomous gold standard outcome, and a test result that is reported on a continuous scale, are generally presented as a ROC curve with or without the related area under the curve (AUC) and its 95% CI. To pool such results, the reviewer has three options: to pool sensitivities and specificities for all

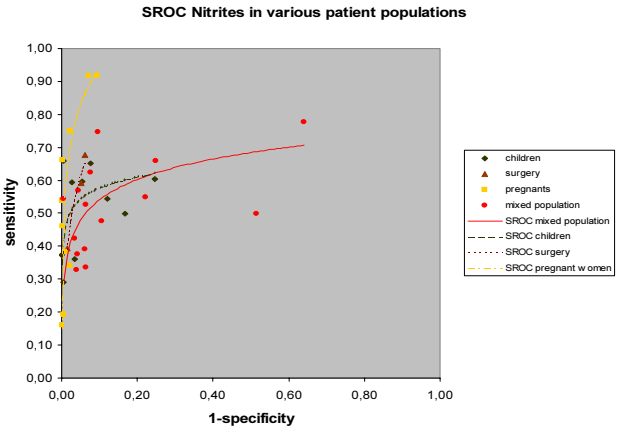

Figure 3
Summary ROC curves of nitrites in urine dipsticks for the diagnosis of bacteriuria and urinary tract infections in various homogeneous subgroups of patient populations.

relevant cut-off points, to pool the AUCs, or to model and pool the ROC curves themselves.

A. A pooled ROC curve and its confidence interval can be constructed on the basis of the pooled sensitivity / specificity values per cut-off point. To make this possible, sufficient raw data have to be available, which is seldom the case.

B. The AUC, like all one-dimensional measures, provides no information about the asymmetric nature of a diagnostic test. It cannot distinguish between curves with a high sensitivity at moderate values of the specificity and curves with a high specificity at moderate values of the sensitivity.

C. As ROC curves are based on ranking, they are robust with respect to inter-study shifts in the value or the meaning of cut-off points. They also provide information about the asymmetrical nature of the test information. To enable direct pooling of ROC curves, a method has been developed that requires only the published curves and the number of positive and negative participants on the gold standard test as input [46]. The ROC curve is scanned into a graphic computer file and then converted into a series of sensitivity versus specificity data, using appropriate software or, ultimately, by hand. Subsequently, a model is fitted for each study, similar to the model that is used for producing SROC curves.

For continuous scale tests, weighted linear regression is used to estimate the parameters for each curve, including a bootstrap method to estimate the standard errors. For ordinal tests, maximum likelihood estimation yields the parameters and their standard errors [46].

The resulting estimates are pooled separately, using a random effect model, and the resulting model is back-transformed into a new-pooled curve with its 95% confidence band.

In addition to causing calculation problems in specific situations, pooling published ROC curves also hides the test values from the picture. Although this is not a problem when evaluating a test method, or when comparing different methods, it limits the possible use of the pooled curve for evaluating the diagnostic value of each specific test result. Moreover, a published curve can be a fitted estimate of the real curve based on the initial values, and any bias resulting from this estimation will be included in the pooled estimates.

Data presentation

A DOR is difficult to interpret because it is a combination of sensitivity and specificity [9]. However, it is useful to present pooled sensitivity and specificity estimates, together with the relevant diagnostic odds ratios for different study characteristics or sub-groups (all estimates with their respective confidence intervals). To make this information accessible to clinicians, the predictive values could be obtained by using the mean prior (pre-test) probabilities of each sub-group. Alternatively, likelihood ratios could be reported so that users can calculate post-test probabilities based on the pre-test probabilities applicable to their patients.

Example taken from the dipstick review: see Table 4.

Pooled DOR (and confidence intervals) of different subgroups can also be presented graphically on a logarithmic scale to end up with symmetric confidence intervals and to reduce the width.

Example taken from the straight-leg-raising test review. In Figure 4 the DOR and confidence boundaries are plotted on the y-axis on a logarithmic scale. Relevant study characteristics (i.e., double-blind versus single-blind studies, studies with or without verification bias) are plotted on the x-axis.

Discussion

While the methodology to conduct a systematic review and meta-analysis of diagnostic research is developed up to a certain extent, at least for dichotomised tests, the exercise itself remains quite a challenge. Systematic reviews have to meet high methodological standards and the results should always be interpreted with caution. Several complicating issues need careful consideration: 1) it is dif-

Table 4:

Factor	DOR (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Prior probability	PPV	NPV
Mixed Population	11 (6–21)	0.50 (0.44–0.58)	0.82 (0.71–0.95)	0.32	0.57	0.78
Surgery	34 (25–47)	0.54 (0.39–0.74)	0.96 (0.93–0.99)	0.20	0.76	0.89

ficult to discover all published evidence, as diagnostic research is often inadequately indexed in electronic databases; 2) the research methods and characteristics of study population and test procedures are often poorly reported in primary research [1,7]; a set of minimal reporting standards for diagnostic research has only recently been discussed: Standards for Reporting of Diagnostic Accuracy-statement (STARD) [http://www.consort-statement.org/] Further Initiatives); 3) the methodological quality and validity of diagnostic research reports is often limited (i.e., no clear definition of 'diseased' participants, no blinding, no independent interpretation test results, insufficient description of participants) [2,7]; 4) accuracy estimates are often very heterogeneous, yet examining heterogeneity is cumbersome, and the process is full of pitfalls; 5) results have to be translated into information that is clinically relevant, taking into account the clinical reality at different levels of health care (prevalence of disease, spectrum of disease, available clinical and other diagnostic information). Even in a state-of-the-art systematic review, the reviewers have to make many subjective decisions when deciding on inclusion or exclusion of studies, on quality assessment and interpretation of limited information, on the exclusion of outliers, and on choosing and conducting subgroup analyses. Subjective aspects have to be assessed independently by more than one reviewer with tracking of disagreements and resolution by consensus or arbitration. These subjective decisions should be explicitly acknowledged in the report to allow the readers some insight into the possible consequences of these decisions on the outcomes of the review and the strength of inference derived from it.

While some researchers question the usefulness of pooling the results of poorly designed research or meta-analysis based on limited information [47,48], we think that examining the effects of validity criteria on the diagnostic accuracy measures and the analysis of subgroups adds valuable evidence to the field of diagnostic accuracy studies. The generation of a pooled estimate, the most likely estimate of the test's accuracy, provides clinicians with useful information until better-conducted studies are published. The reader should remember that evidence about the effect of different aspects of internal or external validity on the results of diagnostic accuracy is still limited [2,3,6]. Consequently, it is difficult to recommend a strict set of methodological criteria at this moment, recognizing that there is, as yet, insufficient evidence to support the use of any minimum set of criteria. Although we have discussed some practical approaches to statistical pooling, other methods are available in the literature [19–21]. Experience with these methods however is yet limited. The development of guidelines for systematic reviews of tests with continuous or ordinal outcomes, reviews of diagnostic strategies of more than one test, and reviews of reproducibility of diagnostic tests remain another challenge, as the methodology is still limited[1] or even non-existing.

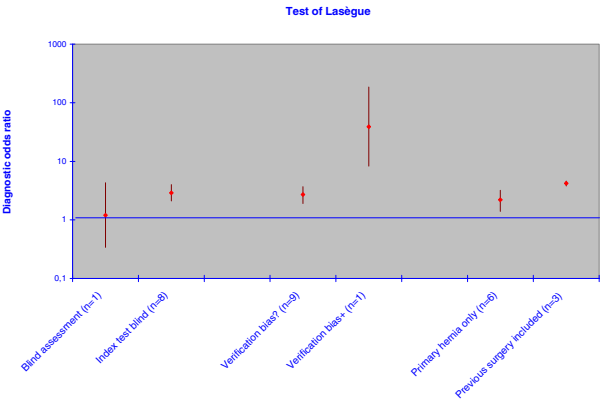

Figure 4
Sub-group analyses of the accuracy of the Lasègue's test for the diagnosis of disc hernia in low back pain. Odds ratios are pooled per sub-group.

Authors' contributions

WL Devillé, developed the guidelines based on his PhD research 'Evidence in diagnostic research' (2001) and studied the search strategy and both systematic reviews used as case studies.

F Buntinx, developed the guidelines with WLD, studied the pooling of ROC curves and provided illustrative examples based on own research.

LM Bouter, participated in the didactical development of the guidelines, the whole content and the structure of the manuscript.

VM Montori, participated in the didactical development of the guidelines, the whole content and commented on the English phrasing.

HCW de Vet, participated in the didactical development of the guidelines, the statistical part, the whole content and the structure of the manuscript.

DAWM van der Windt, participated in the didactical development of the guidelines, the whole content and commented on the English phrasing.

PD Bezemer, participated in the content of the manuscript and the statistical part.

All authors read and approved the final manuscript

Competing interests
None declared.

Note
This article was first produced as a chapter in the following book:

Knottnerus, J. A. ed., The Evidence Base of Clinical Diagnosis London: BMJ Books 2002, and has been reproduced with permission from BMJ Books.

Additional material

Additional file 1
Appendix statistical formulae
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-2-9-S1.doc]

Acknowledgments
We like to thank JG Lijmer, MP Zeegers, and RA de Bie for critically reviewing the manuscript.

References
1. Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F: **Guidelines for meta-analyses evaluating diagnostic tests.** *Ann Intern Med* 1994, **120:**667-676
2. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, PMM Bossuyt: **Empirical evidence of design-related bias in studies of diagnostic tests.** *JAMA* 1999, **282:**1061-1066
3. Oosterhuis WP, Niessen RWLM, Bossuyt PMM: **The science of systematic reviewing studies of diagnostic tests.** *Clin Chem Lab Med* 2000, **38:**577-588
4. Jeaschke R, Guyatt GH, Sackett DL: **User's guidelines to the medical literature, III: how to use an article about a diagnostic test, A: are the results of the study valid?** *JAMA* 1994, **271:**389-391
5. Jeaschke R, Guyatt GH, Sackett DL: **User's guidelines to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients?** *JAMA* 1994, **271:**703-707
6. Greenhalgh T: **How to read a paper:papers that report diagnostic or screening tests.** *BMJ* 1997, **315:**540-543
7. Reid MC, Lachs MS, Feinstein AR: **Use of methodological standards in diagnostic test research: getting better but still not good.** *JAMA* 1995, **274:**645-651
8. Devillé WL, Buntinx F: **Didactic Guidelines for Conducting Systematic Reviews of Studies Evaluating the Accuracy of Diagnostic Tests** *in: The Evidence Base of Clinical Diagnosis. (Edited by: A. Knottnerus) BMJ Books, London* 2002, 145-65
9. Deeks JJ: **Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests** *BMJ* 2001, **323:**157-162
10. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC: **Developing optimal search strategies for detecting clinically sound studies in Medline.** *J Am Med Informatics Assoc* 1994, **1:**447-458
11. Dickersin K, Scherer R, Lefebvre C: **Identifying relevant studies for systematic reviews.** *BMJ* 1994, **309:**1286-91
12. van der Weijden T, Yzermans CJ, Dinant GJ, van Duijn NP, de Vet R, Buntinx F: **Identifying relevant diagnostic studies in MEDLINE. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example.** *Fam Pract* 1997, **14:**204-208
13. Devillé WLJM, Bezemer PD, Bouter LM: **Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy.** *J Clin Epidemiol,* 2000, **53:**65-69
14. McClish DK: **Combining and comparing area estimates across studies or strata.** *Med Dec Making* 1992, **12:**274-279
15. Littenberg B, Moses LE: **Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method.** *Med Dec Making* 1993, **13:**313-321
16. Midgette AS, Stukel TA, Littenberg B: **A meta-analytic method for summarising diagnostic test performances: Receiver-operating-characteristic-summary point estimates.** *Med Dec Making* 1993, **13:**253-257
17. Moses LE, Shapiro D, Littenberg B: **Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations.** *Stat Med* 1993, **12:**1293-1316
18. Hasselblad V, Hedges LV: **Meta-analysis of screening and diagnostic tests.** *Psychol Bull* 1995, **117:**167-178
19. Irwig L, Macaskill P, Glasziou P, Fahey M: **Meta-analytic methods for diagnostic accuracy.** *J Clin Epidemiol* 1995, **48:**119-130
20. Rutter CM, Gatsonis CA: **Regression methods for meta-analysis of diagnostic test data.** *Acad Radiol* 1995, **2:**S48-S56
21. Shapiro DE: **Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test.** *Acad Radiol* 1995, **2:**S37-S47
22. Walter SD, Irwig L, Glasziou PP: **Meta-analysis of diagnostic tests with imperfect reference standards.** *J Clin Epidemiol* 1999, **52:**943-951
23. Yusuf S, Wittes J, Probsfield J, Tyroler HA: **Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials.** *JAMA* 1991, **266:**93-98
24. Oxman A, Guyatt G: **A consumer's guide to subgroup analysis.** *Ann Intern Med* 1992, **116:**78-84
25. Thompson SG: **Why sources of heterogeneity in meta-analysis should be investigated.** *BMJ* 1994, **309:**1351-1355
26. Colditz GA, Burdick E, Mosteller F: **Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary.** *Am J Epidemiol* 1995, **142:**371-382
27. Mulrow C, Langhorne P, Grimshaw J: **Integrating heterogeneous pieces of evidence in systematic reviews.** *Ann Int Med* 1997, **127:**989-995
28. Devillé WLJM, Yzermans JC, van Duin NP, van der Windt DAWM, Bezemer PD, LM Bouter: **Which factors affect the accuracy of the urine dipstick for the detection of bacteriuria or urinary tract infections? A meta-analysis.** *In: Evidence in diagnostic research. Reviewing diagnostic accuracy: from search to guidelines. In PhD thesis Vrije Universiteit Amsterdam, the Netherlands, Chapter 4* 2001, 39-73
29. Devillé WLJM, van der Windt DAWM, Dzaferagic A, Bezemer PD, Bouter LM: **The test of Lasègue: systematic review of the accuracy in diagnosing herniated discs.** *Spine* 2000, **25:**1140-1147
30. Hunt DL, McKibbon KA: **Locating and appraising systematic reviews.** *Ann Int Med* 1997, **126:**532-538
31. van Tulder MW, Assendelft WJJ, Koes BW, Bouter LM, and the Editorial Board of the Cochrane Collaboration Back Review Group: **Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for spinal disorders.** *Spine* 1997, **22:**2323-2330
32. Obuchowski NA: **Sample size calculations in studies of test accuracy.** *Stat Meth in Med Research* 1998, **7:**371-392
33. Campens D, Buntinx F: **Selecting the best renal function tests: a meta-analysis of diagnostic studies.** *Int J Techn Ass Health Care* 1997, **13:**343-356
34. Grégoire G, Derderian F, Le Lorier J: **Selecting the language of the publications included in a meta-analysis: is there a tower of Babel bias?** *J Clin Epidemiol* 1995, **48:**159-163
35. **Cochrane Methods Group on Screening and Diagnostic Tests.** *Recommended methods* [http://www.cochrane.org/cochrane/sadt.htm]
36. Knottnerus JA, Muris JW: **Assessment of the accuracy of diagnostic tests: the cros-sectional study.** *In: The evidence base of clinical diagnosis. (Edited by: JA Knottnerus) BMJ Books, London* 2002, 39-59
37. Aertgeerts B, Buntinx F, Kester A, Fevery J: **Diagnostic value of the CAGE questionnaire in screening for alcohol abuse and alcohol dependence: a meta-analysis.** *In: Screening for alcohol abuse or*

dependence. In PhD-thesis, Katholieke Universiteit Leuven, Belgium, 2000, **Appendix 3:**
38. **Centre for Evidence Based Medicine. Levels of Evidence and Grades of Recommendations.** [http://cebm.jr2.ox.ac.uk/docs/levels.html]
39. Fleiss JI: **The statistical basis of meta-analysis.** *Stat Methods Med Res* 1993, **2:**121-145
40. Galbraith RF: **A note on graphical presentation of estimated odds ratios from several clinical trials.** *Stat Med* 1988, **7:**889-894
41. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Controlled Clin Trials* 1986, **7:**177-188
42. Rutter CM, Gatsonis CA: **A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations.** *Stat Med* 2001, **20:**2865-84
43. Greenland S: **Invited commentary: a critical look at some popular meta-analytic methods.** *Am J Epidemiol* 1994, **140:**290-296
44. Irwig L: **Modelling result-specific likelihood ratios.** *J Clin Epidemiol* 1989, **42:**1021-1024
45. Tosteson ANA, Begg CB: **A general regression methodology for ROC curve estimation.** *Med Decis Making* 1988, 204-215
46. Kester A, Buntinx F: **Meta-analysis of ROC-curves.** *Med Decis Making* 2000, **20:**430-439
47. Greenland S: **A critical look at some popular meta-analytic methods.** *Am J Epidemiol* 1994, **140:**290-301
48. Shapiro S: **Meta-analysis/Shmeta-analysis.** *Am J Epidemiol* 1994, **140:**771-777

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/2/9/prepub