ELSEVIER

Journal of Clinical Epidemiology 65 (2012) 1088–1097

Journal of Clinical Epidemiology

Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity

Mariska M.G. Leeflang[a,*], Jonathan J. Deeks[b], Anne W.S. Rutjes[c,d], Johannes B. Reitsma[e], Patrick M.M. Bossuyt[a]

[a]*Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, PO Box 22700, Amsterdam 1100 DE, The Netherlands*
[b]*Biostatistics, Evidence Synthesis and Test Evaluation Research Group, School of Health and Population Sciences, Public Health Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*
[c]*Division of Clinical Epidemiology & Biostatistics, Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland*
[d]*Clinical Center for Aging Sciences (Ce.S.I.), University 'Gabriela d'Annunzio' Foundation, Via Colle Dell'Ara, 66013 Chieti Scalo, Chieti, Italy*
[e]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands*

Accepted 26 March 2012; Published online 27 June 2012

Abstract

Objective: Meta-analysis of predictive values is usually discouraged because these values are directly affected by disease prevalence, but sensitivity and specificity sometimes show substantial heterogeneity as well. We propose a bivariate random-effects logitnormal model for the meta-analysis of the positive predictive value (PPV) and negative predictive value (NPV) of diagnostic tests.

Study Design and Setting: Twenty-three meta-analyses of diagnostic accuracy were reanalyzed. With separate models, we calculated summary estimates of the PPV and NPV and summary estimates of sensitivity and specificity. We compared these summary estimates, the goodness of fit of the two models, and the amount of heterogeneity of both approaches.

Results: There were no substantial differences in the goodness of fit or amount of heterogeneity between both models. The median absolute difference between the projected PPV and NPV from the summary estimates of sensitivity and specificity and the summary estimates of PPV and NPV was 1% point (interquartile range, 0–2% points).

Conclusion: A model for the meta-analysis of predictive values fitted the data from a range of systematic reviews equally well as meta- ıe primary

Keywords: Systematic reviews; Meta-analyses; Diagnostic test accuracy; Sensitivity and specificity; Positive predictive value; Negative predictive value

1. Background

Current guidance for meta-analyses of diagnostic test accuracy advocates the use of hierarchical methods to summarize estimates of sensitivity and specificity or the diagnostic odds ratio [1,2]. Sensitivity and specificity are not always the most intuitive measures for clinicians, as they express at the group level how many diseased and nondiseased were correctly identified as such by the test. Clinicians may be more familiar with predictive values [3,4]. The positive predictive value (PPV), for example, expresses the probability of disease in those testing positive. Similarly, the negative predictive value (NPV) expresses the probability of the absence of disease in those testing negative.

For clinicians interested in predictive values, meta-analyses of these statistics may be easier to understand and to apply in practice. In addition, predictive values may suffer less from work-up bias and problems associated with partial and differential verification bias than sensitivity and specificity [5]. In many studies, test positives are verified by a different clinical reference standard than test negatives. Examples are accuracy studies where biopsy is the reference standard; if no lesion is found, biopsy cannot be done. In these studies, test negatives are either not verified at all or clinically followed-up for verification of the negative test result. Different reference standards may also be applied for ethical reasons, where one hesitates to use the preferred but invasive reference standard in test negatives,

* Corresponding author. Tel.: +31-20-5666934; fax: +31-20-6912683.
E-mail address: m.m.leeflang@amc.uva.nl (M.M.G. Leeflang).

0895-4356 © 2012 Elsevier Inc. Open access under the Elsevier OA license.
doi: 10.1016/j.jclinepi.2012.03.006

What is new?

- Current guidance for meta-analyses of diagnostic test accuracy advocates the use of hierarchical methods to summarize estimates of sensitivity and specificity or the diagnostic odds ratio.
- We propose a model for direct meta-analysis of predictive values, using similar hierarchical methods.
- A model for the meta-analysis of predictive values fitted the data from a range of systematic reviews equally well as meta-analysis of sensitivity and specificity.
- The choice for either model could be guided by considerations about the designs used in the primary studies and likely sources of heterogeneity.

who have a lower probability of disease. Under these circumstances, summarizing predictive values may be more meaningful than meta-analysis of sensitivity and specificity.

Meta-analysis of predictive values in systematic reviews of test accuracy studies has been discouraged for several reasons. Predictive values are expected to be more heterogeneous than other accuracy measures because they would vary more directly with changes in disease prevalence. Sensitivity and specificity, in contrast, are assumed to be more stable characteristics of tests, which would make summary estimates of sensitivity and specificity more meaningful. The sources of variation and bias for sensitivity and specificity are better understood than those for predictive values. Yet, one could successfully argue that these arguments are too simple. Sensitivity and specificity are not fixed test properties either. They describe the behavior of a test under specific conditions, and they typically change across different segments of the disease spectrum and with varying disease prevalence [6–8].

We propose a model for the meta-analysis of predictive values, based on a previously published bivariate logitnormal random-effects model for the meta-analysis of sensitivity and specificity.

If it is true that predictive values are more heterogeneous than sensitivity and specificity, one can expect a lower goodness of fit with a model for meta-analysis based on predictive values. We therefore compared, across a range of published systematic reviews, to what extent a model for meta-analysis based on predictive values provides a better—or worse—summary of the data than an equivalent model based on sensitivity and specificity [9]. We also compared the summary estimates from the model for predictive values with projected estimates from a meta-analysis of sensitivity and specificity, using the median prevalence and Bayes' rule.

2. Methods

2.1. Study set

We used a set of 31 meta-analyses, selected and analyzed for a previously published report on bias and variation in diagnostic accuracy studies. The meta-analyses cover a wide range of clinical topics and diagnostic tests, such as imaging tests, laboratory tests, physical examination, and questionnaires. For more details on the search process, selection and data-extraction, we refer to the original report [10]. In short, a number of electronic databases were searched for systematic reviews published between January 1999 and April 2002 and fulfilling the following criteria: 1) assessment of diagnostic test accuracy; 2) including at least 10 original studies on the same diagnostic test; 3) no exclusion of primary studies based on design features; and 4) the ability to reproduce the 2×2 tables from the original studies. We excluded case–control studies to allow for realistic estimates of prevalence and predictive values.

2.2. Definitions

For every study, in each systematic review, we calculated the PPV and NPV, and the sensitivity and specificity from the reported numbers. The PPV was defined as the proportion of patients with the target condition in those testing positive on the test under evaluation. Similarly, the NPV was defined as the proportion of patients without the target condition in those testing negative on the test under evaluation. Conventionally, sensitivity was defined as the proportion of patients testing positive in those with the target condition. Specificity was defined as the proportion of patients testing negative in those without the target condition.

2.3. Descriptive statistics

We calculated the median, interquartile range (IQR), and minimum and maximum predictive values, sensitivity, and specificity of the studies in each review. We plotted these descriptive statistics for PPV and NPV side by side to those for sensitivity and specificity.

2.4. Meta-analysis

We developed a bivariate logitnormal random-effects model for meta-analysis of predictive values. This model has the same form as the previously proposed bivariate logitnormal model for meta-analysis of sensitivity and specificity [9]. That model was in itself based on an approach to meta-analysis introduced by Van Houwelingen et al. [11,12].

The original bivariate model was used to obtain summary estimates of sensitivity and specificity for each test in a review. In this model, pairs of sensitivity and specificity are jointly analyzed, incorporating any correlation that is to be expected between these two measures using a random-effects approach. The correlation between sensitivity and specificity will be mainly driven by threshold effects:

when a higher test result is more associated with disease, sensitivity will be lower when the positivity threshold increases, whereas specificity will be higher. Variability within each individual study is assumed to be binomial, affected by the number of patients and the study-specific sensitivity and specificity. Across studies, sensitivity and specificity for individual studies are assumed to follow a bivariate logitnormal distribution [9].

In the bivariate logitnormal random-effects model for predictive values, pairs of PPV and NPV are jointly analyzed, incorporating any correlation between these two measures using a random-effects approach. The correlation between PPV and NPV will be mainly driven by prevalence: when prevalence increases, PPV will be higher, whereas NPV will be lower. Here too, we assume that the true logit PPV and true logit NPV of the individual studies are normally distributed around a common mean (logitnormal distribution), but that the true values may vary between studies (random-effects assumption). The variability within each individual study is assumed to be binomial, affected by the number of patients and the study-specific PPV and NPV.

Data were analyzed in SAS for Windows, version 9.2 (Cary, NC), using the PROC NLMIXED procedure (see Appendix). For every meta-analysis, the starting values for the parameters were adapted, either based on graphs plotting sensitivity by 1 − specificity or PPV by 1 − NPV, or on separate univariate analyses of each outcome.

2.5. Goodness of fit

We calculated *Akaike's Information Criterion* (AIC) for both bivariate models, in each application, as a measure of goodness of fit of the models. We used the Wilcoxon signed-rank test to evaluate a systematic difference in goodness of fit between the two approaches.

2.6. Heterogeneity

As heterogeneity in diagnostic accuracy studies can be expressed in terms of the variance in logit sensitivity (or logit PPV), variance in logit specificity (or logit NPV), and the covariance between the two measures, we expressed heterogeneity as the area of the prediction ellipse around the summary point estimate of the mean in logit space. The prediction ellipse is a two-dimensional representation of the 95% prediction region around the logit sensitivity (or PPV) and logit specificity (or NPV). The area of any ellipse is calculated as:

$$\text{Area} = a * b * \pi$$

where a and b are the minor and major axis of the ellipse (see Fig. 1A). Derived from a bivariate logitnormal model, the length of both axes depends on the variances of logit sensitivity (or PPV) and logit specificity (or NPV), and the correlation between the two. This can be expressed as the eigen value, which corresponds to the variances adjusted for correlation and follows from the variance–covariance matrix of the model. The square roots of the two eigen values are the lengths of the minor and major axis a and b (see also Fig. 1B).The areas for the two bivariate approaches were compared by computing the relative area for each included meta-analysis. We used the Wilcoxon signed-rank test to evaluate whether the area derived from one approach was systematically smaller or larger than the area derived from the other approach.

2.7. Derivation of predictive values from sensitivity and specificity

From the summary estimates of sensitivity and specificity, the natural logarithms of the corresponding positive and negative likelihood ratios were estimated, their standard errors, and hence confidence intervals being estimated using the delta method. These estimates were then used to calculate projected predictive values at the median prevalence in each review, using Bayes' rule [3,4]:

$$\text{PPV} = \frac{\text{post-test odds}}{(1 + \text{post-test odds})}$$

$$\text{NPV} = 1 - \left(\frac{\text{post-test odds}}{(1 + \text{post-test odds})}\right)$$

$$\text{Post-test odds} = \left[\left(\frac{\text{prevalence}}{(1 - \text{prevalence})}\right) \times \text{likelihood ratio}\right]$$

Where the positive likelihood ratio is used to calculate the PPV and the negative likelihood to calculate the NPV. These results were compared with the summary estimates from the predictive values model.

3. Results

3.1. Descriptive statistics

Twenty-three meta-analyses fulfilled our criteria. They contained 346 original studies in total, varying in sample size from 10 to 20,381 patients (Table 1) [13–36]. The variability in sensitivity, specificity, PPV, and NPV is shown in Fig. 2. Mean prevalence varied between the reviews from 1% to 52%; the range in prevalence varied for each review between 5% points (narrow prevalence range) to 78% points (broad prevalence range) (see Fig. 3).

3.2. Goodness of fit

The random-effects model for predictive values converged in all 23 cases. In 11 of the 23 reviews, the sensitivity/specificity model showed a better fit (lower AIC), whereas in the 12 reviews, the predictive values model showed a better fit ($P = 0.83$).

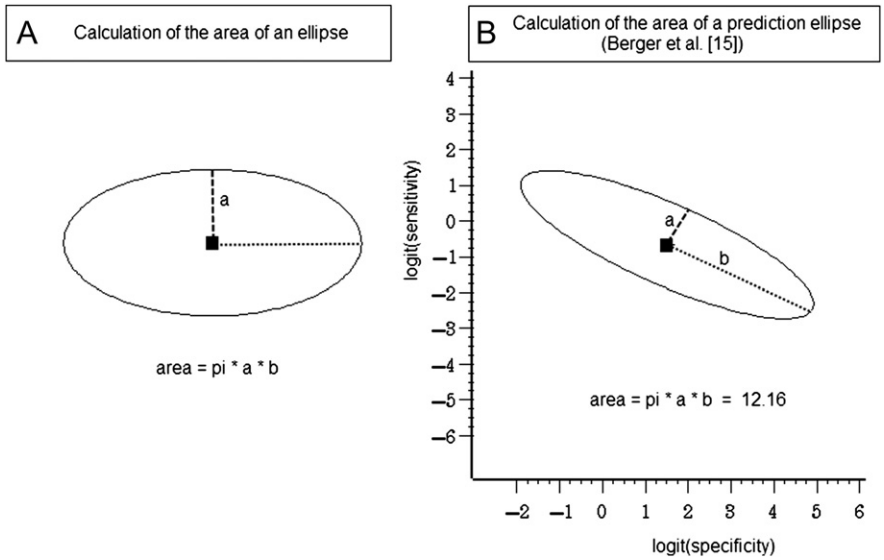

Fig. 1. Calculation of the surface area of an ellipse. A. General calculation. B. Example of 95% prediction ellipse for logit sensitivity and logit specificity. The dashed line is the shorter radius of the ellipse; the dotted line is the major radius of the ellipse.

3.3. Heterogeneity

Heterogeneity was assessed by calculating the area of the prediction ellipse around the point estimates (see Fig. 1). The mean ratio of the areas for the models of sensitivity and specificity vs. the predictive values approach varied from 0.5 (i.e., area around predictive values was about two times larger than around sensitivity and specificity) to 3.5 (i.e., area around sensitivity and specificity was 3.5 times larger than around predictive values). In seven meta-analyses, this area was larger for sensitivity and specificity than for PPV and NPV. In 15 meta-analyses, it was the other way around (see Fig. 4; $P = 0.09$). In one meta-analysis, the

Table 1. Characteristics of included meta-analysis

Author	Topic	N[a]	Prevalence (median + range)
Mol [29]	Ultrasound for Down Syndrome	22 (99–20,381)	0.01 (0.00–0.05)
Hoogendam [21]	Digital rectal examination for prostate cancer	13 (309–6,630)	0.03 (0.01–0.07)
Lederle [27]	Screening for abdominal aortic aneurysm	10 (21–424)	0.07 (0.01–0.17)
Huicho [22]	Urine marker: dipstick nitrate	16 (23–3,251)	0.15 (0.01–0.30)
Sloan [32]	Diagnosis of gonorrhea and chlamydial infections	14 (172–1,222)	0.13 (0.03–0.37)
Berger [15]	Upper abdominal pain for gallstones	11 (83–1,896)	0.13 (0.05–0.41)
Lau [14] + Balk [13]	Creatine kinase-myoglobin for acute myocardial infarction in the emergency department	19 (59–2,093)	0.22 (0.06–0.42)
Wiese [36]	Wet mount slide for vaginal trichomoniasis	29 (68–1,199)	0.26 (0.07–0.55)
Fiellin [17]	Questionnaires for lifetime alcohol dependence	12 (84–1,333)	0.28 (0.05–0.45)
Hurley [23]	Limulus amebocyte lysate assay for diagnosis of Gram-negative infections	22 (10–218)	0.25 (0.10–0.68)
Visser [35]	Ultrasound for peripheral arterial stenosis	17 (12–167)	0.29 (0.05–0.72)
Nelemans [30]	Magnetic resonance angiography for peripheral arterial disease	13 (12–45)	0.41 (0.15–0.73)
Kwok [26]	Exercise electrocardiography for coronary artery disease in women	19 (33–613)	0.37 (0.18–0.62)
Hoffman [20]	Workup of prostate cancer	10 (29–673)	0.41 (0.22–0.76)
Sonnad [33]	Magnetic resonance imaging for staging of prostate cancer	21 (18–235)	0.48 (0.23–0.82)
Mitchell [28]	Pap smear for squamous intraepithelial lesions of the cervix	17 (18–3,534)	0.43 (0.17–0.95)
Hobby [19]	Diagnosis of tears of the triangular fibrocartilage complex in the wrist	11 (13–102)	0.50 (0.39–0.94)
Safriel [31]	Spiral computed tomography for pulmonary emboli	10 (20–149)	0.46 (0.27–0.82)
Vasquez [34]	Workup of acute cholecystitis	15 (20–163)	0.47 (0.10–0.72)
Kelly [24]	Workup of staging in gastroesophageal carcinoma	13 (13–328)	0.60 (0.23–0.85)
Kim [25]	Dobutamine stress echocardiography for coronary artery disease	39 (27–288)	0.72 (0.29–0.87)
Gould [18]	Positron emission tomography in the workup of pulmonary nodules	29 (17–109)	0.70 (0.50–0.96)
Deville [16]	Workup of herniated discs in patients selected for surgery	11 (52–2,504)	0.77 (0.56–0.98)

First author and reference number, topic of the review, number of studies and range of number of participants per study, and median and range of prevalence. The reviews are sorted by mean prevalence.

[a] Studies (range of included patients per study).

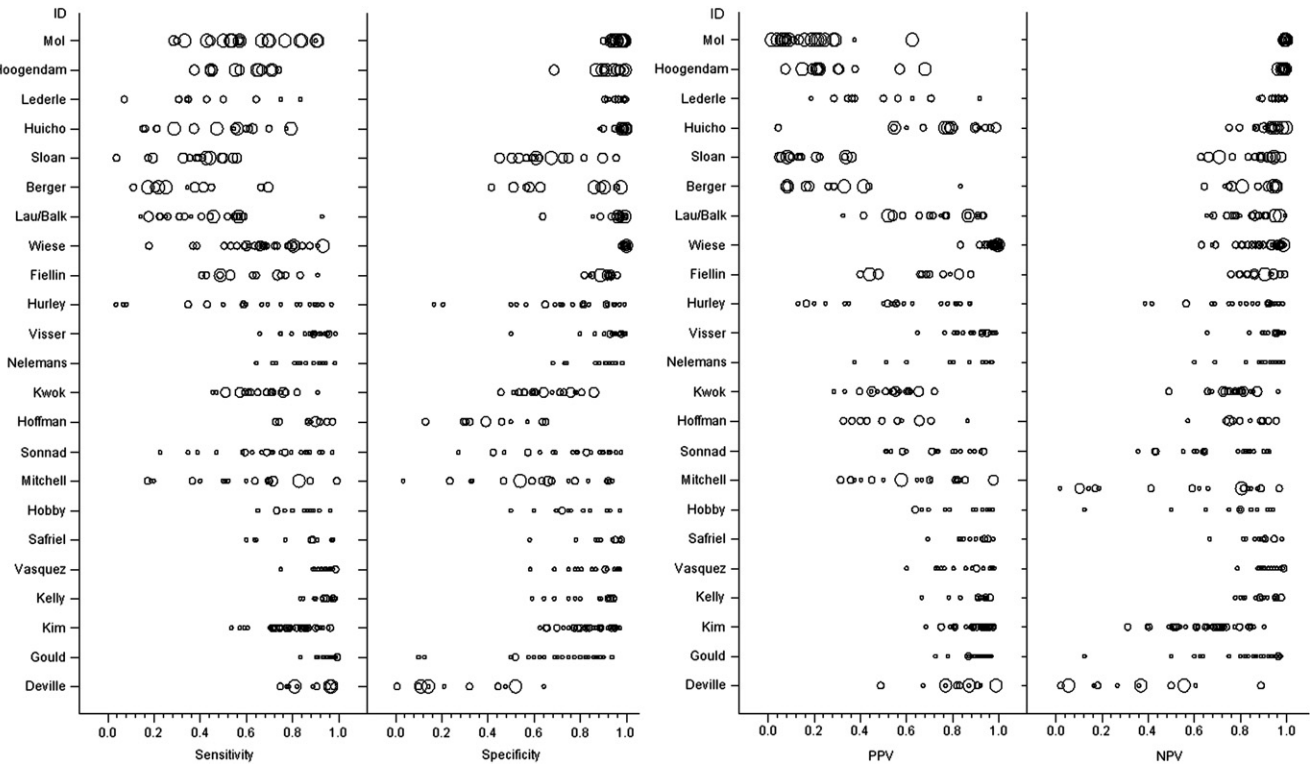

Fig. 2. Raw results for sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Reviews are in alphabetical order. Each bubble represents a single primary study. The bubbles have four different sizes, representing four different study size categories: <100 participants; 100−500 participants; 500−1,000 participants; and >1,000 participants. The reviews are sorted by mean prevalence.

two regions were equal in size, indicating that heterogeneity for both approaches was similar.

Predictive values are assumed to be directly influenced by variation in prevalence. Fig. 3 shows the summary estimates of the predictive values alongside the median and range of prevalence within each review. Across reviews, there is a trend toward a higher PPV and a lower NPV as median prevalence increases. We tested for a relationship between prevalence and predictive values within reviews by adding covariate terms for prevalence in the predictive-value bivariate model, with separate terms for its impact on PPV and NPV. Significant relationships ($P < 0.05$) were detected within most reviews, with only six reviews not demonstrating significant relationships for either PPV or NPV (results not shown). For example, the meta-analysis of 17 studies of ultrasound for peripheral arterial stenosis [35] reported a broad range in prevalence (5−72%) but a narrow range in predictive values. PPV ranged from 0.65 to 0.99 and showed no significant relationship with prevalence ($P = 0.40$); NPV ranged from 0.66 to 0.98 and the effect of prevalence was significant ($P < 0.001$).

3.4. Comparison of meta-analysis of predictive values vs. estimates derived from meta-analysis of sensitivity and specificity

The absolute difference between directly estimated predictive values and those projected after pooling sensitivity and specificity and using the median prevalence varied between 0% points and 10% points (median 1% point, IQR 0−2% points) (see Fig. 3). One review showed a difference of 10% points for both PPV and for NPV: the directly estimated PPV was 10% higher and the directly estimated NPV was 10% points lower than the predictive value estimated after estimating sensitivity and specificity [28]. This review also showed the broadest range in prevalence among the included studies (range 17−95%). In one other review, these differences were 6% points and 9% points, respectively [16]. This review showed a moderate range in prevalence (from 56% to 98%).

4. Discussion

In this article, we have proposed an approach for the meta-analysis of NPV and PPV. Our strategy relies on a logitnormal bivariate random-effects model, similar to the one previously proposed for meta-analysis of sensitivity and specificity. In a series of existing systematic reviews, the model converged in all included meta-analyses. There was no systematic difference in the goodness of fit between the random-effects models based on predictive values and the conventional one, based on sensitivity and specificity, for the same data. In most meta-analyses, sensitivity and specificity did show less variation than predictive values, but this difference was not significant. There were no

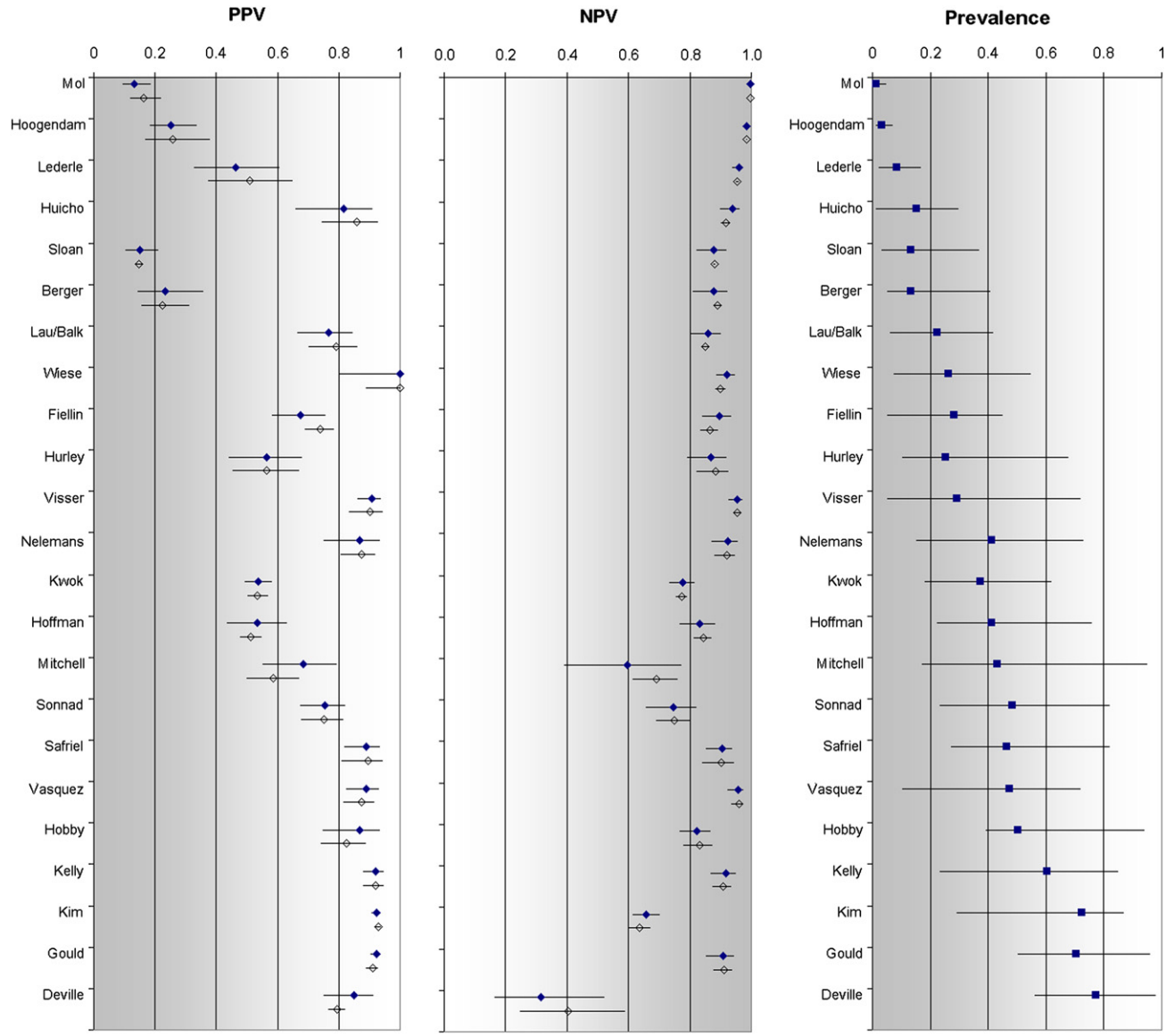

Fig. 3. Positive predictive value (PPV), negative predictive value (NPV), and prevalence together with their corresponding 95% confidence intervals. PPV and NPV are either directly calculated (blue closed diamonds) or calculated from the estimates for sensitivity and specificity (open diamonds). The right-hand graph shows the accompanying median (closed squares) and range for prevalence in each meta-analysis. The reviews are sorted by mean prevalence.

systematic differences between directly estimated summary estimates of predictive values and those calculated through Bayes' rule from summary estimates of sensitivity and specificity.

A number of potential limitations to our comparison have to be taken into account. One of the reasons that we did not find statistically significant differences between pooling sensitivity and specificity and pooling predictive values may be the relatively low number of studies per meta-analyses. The number of studies varied between 10 and 39, so analyses may have been underpowered to detect such differences. We also followed the judgments of the review authors that the studies they included in each meta-analysis were homogeneous enough to combine. Had stricter criteria been used in terms of only combining studies with similar diagnostic thresholds and similar disease prevalence heterogeneity may have been reduced in both sets of analyses. Another limitation is related to the difficulty in expressing heterogeneity in two dimensions. There is no clear guidance on how heterogeneity in meta-analyses of diagnostic accuracy should be measured and expressed. Here, we proposed measuring the area of the prediction ellipse as an overall measure of heterogeneity in two directions.

There is no a priori statistical justification for selecting a model for predictive values over a model for sensitivity and specificity. Yet, in itself, meta-analysis of predictive values has some advantages. Clinicians often tend to think

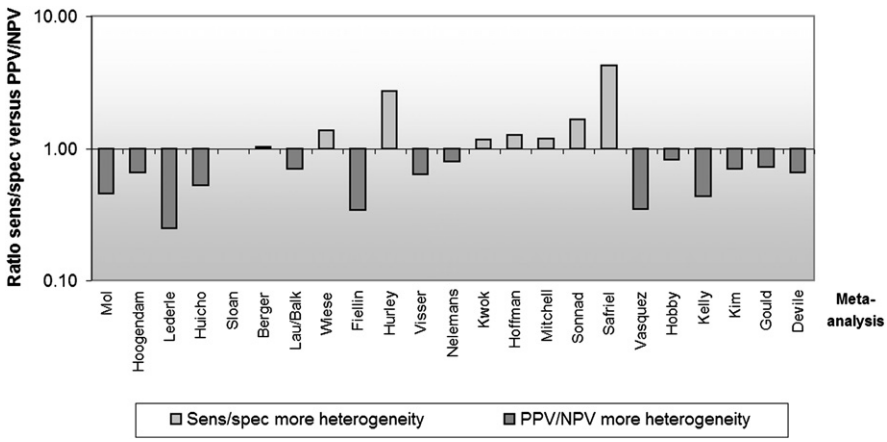

Fig. 4. Ratio between surface area of the prediction ellipses for the sens/spec approach and the positive predictive value (PPV)/negative predictive value (NPV) approach. Sensitivity and specificity showed more heterogeneity in eight meta-analyses (light gray columns); PPV and NPV showed more heterogeneity in 14 meta-analyses (dark gray columns); and in one meta-analysis both prediction ellipses had an equal area. The reviews are sorted by mean prevalence.

in predictive values or posttest probabilities. Anecdotal and theoretical evidence exists that sensitivity and specificity vary frequently from setting to setting, so it might be just as practical and reasonable to perform meta-analysis of predictive values rather than sensitivity and specificity. Another advantage of meta-analysis of predictive values is that it provides researchers with valid estimates even when those for sensitivity and specificity may be biased because of differential verification, or when meta-analysis of these measures is simply impossible. This may be the case, for example, in colonoscopy studies where negative findings cannot be verified, or in cancer studies in which biopsy is the reference standard: if the tumor is not found, biopsy cannot be done. So none or only a small part of the index test negatives will be verified. When estimating sensitivity and specificity in such studies, there will be less false negatives and true negatives (partial verification bias) and the sensitivity will be inflated while the specificity will be underestimated [5,9,37]. A solution for this may be to use a different reference standard to verify the test negatives, but this may lead to differential verification bias (of which the effect is difficult to predict). Estimates of PPV will not suffer from verification bias, as the test positives are correctly verified. Only the NPV will be affected, but not in the same way as sensitivity and specificity.

Meta-analysis of predictive values also has some potential disadvantages. The effects of sources of heterogeneity—in terms of bias and variation—are well documented for sensitivity and specificity but have not been investigated as well for their effect on predictive values. For example, the major source of variation in sensitivity and specificity is related to threshold differences for test positivity. When a higher test value corresponds to a higher probability of disease, then increasing the threshold will result in a lower sensitivity and a higher specificity. Although this effect can also be seen in predictive values, the effect of prevalence can be much larger. Small changes in prevalence then overtake the changes caused by threshold effects. Another disadvantage is a minor but very practical one: in our data set the models for predictive values took a bit more effort to converge than the models for sensitivity and specificity. It was more often needed to adjust starting values for parameters (although this can be automated).

The main disadvantage of meta-analysis of predictive values could be the interpretation of the results and the translation into practice. Although sensitivity and specificity are correlated with changes in positivity threshold, PPV and NPV correlate, depending on changes in prevalence. Furthermore, these models are all random-effects models and assume that the true parameters of the outcome are not constant but follow a normal distribution. When a review contains studies using a broad range of different thresholds, current guidance indicates that these studies should not be analyzed using the bivariate method, but rather with a hierarchical summary ROC method [2]. The latter provides an overall estimate of the receiver operating characteristic (ROC) curve, expressing dependence of accuracy on positivity threshold. This may be more appropriate in such a case than providing pooled estimates of sensitivity and specificity at an ''average'' threshold. A similar problem can occur when analyzing predictive values. When a review contains studies with a broad range of prevalence, there will be difficulties in working out how the average predictive values (estimated for the average included study with an average prevalence) can be applied in clinical practice. However, a broad range in prevalence should urge researchers to investigate whether these difference may be caused by differences in populations and settings, disregard whether the outcome measure is sensitivity and specificity or PPV and NPV.

Recently, a trivariate model has been proposed for meta-analysis of predictive values and prevalence jointly, resulting in summary estimates of PPV, NPV, and prevalence while taking into account the correlation between the three

measures [38]. This may, however, not solve the problems outlined above, as it does not provide a structure for PPV and NPV to vary with prevalence. A solution may be to always include prevalence as a covariate to the models when analyzing predictive values and to report summary estimates for, for example, the first quartile prevalence, the median prevalence, and the third quartile prevalence, as presented in the included studies. This way, clinicians can choose the prevalence, which they think corresponds best with their particular situation, which would be a major strength of this approach.

In this article, we did not explore using predictive value models for comparing tests, nor did we investigate including covariates for clinical subgroups in any of the models. Such covariates are often used in explorations of heterogeneity, when meta-regression is applied to explore the sources of variability related to differences in study design and execution, or to differences in patient groups and testing. These covariates will have a very different meaning when applied in a model for the meta-analysis of predictive values than when applied to sensitivity and specificity. In models comparing multiple tests, differences in disease prevalence between tests or subgroups will hamper the interpretation of observed differences in PPV and NPV. An example would be if all studies of Test A are at a higher prevalence than those of Test B. Evaluations of tests at the same prevalence and clinical setting will be needed to draw clinically robust conclusions as to which test performs best in which situations [39]. An improved understanding of the different effects of sources of heterogeneity and the pros and cons of both approaches is required. Awaiting this, researchers may consider reporting the results from both models, including appropriate covariates.

In nearly all systematic reviews of medical tests, there is a complex interplay between sensitivity, specificity, prevalence, and predictive values within and between studies, something which review authors have to address carefully. We have proposed a model for meta-analysis of predictive values and applied it successfully to a range of previously published systematic reviews of test accuracy studies. Across the included systematic reviews, we found no signs of a lower goodness of fit for this model compared with a similar model for meta-analysis of sensitivity and specificity from the same data, in the absence of covariates. In our view, the choice between the two models should rely on the design of the original accuracy studies, the purpose of the review and the questions guiding it, and on the heterogeneity between studies.

Acknowledgments

The authors thank Dr M. Di Nisio, Dr J.C. van Rijn, and Dr N. Smidt for their contribution to the collection of data. The authors thank Professor Dr Aeilko Zwinderman for his contribution to the calculation of the area of the ellipse. M.M.G.L. is supported by The Netherlands Organization for Scientific Research (NWO); project 916.10.034. A.W.S.R. was supported by a research grant from the NWO (registration no. 945-10-012). J.J.D. is partially supported by the Medical Research Council Midland Hub for Trials Methodology Research, University of Birmingham (grant G0800808). No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the article.

Competing interests: The authors declare that they have no competing interests.

Authors' contributions: M.M.G.L. initiated the research, analyzed the data, and wrote draft versions of the article and the final version of the article. J.J.D. designed the calculations for the surface areas of the ellipses and commented on draft and final versions of the article. A.W.S.R. collected the data from the included reviews, developed and maintained the databases, and commented on final versions of the article. J.B.R. assisted with the analyses, and commented on draft and final versions of the article. P.M.M.B. commented on the draft and final versions of the article.

Appendix

SAS syntax

Bivariate meta-analysis of sensitivity and specificity

```
data nl_test;
 set _meta;
 _rec+1;
 _dis=1; _nondis=0; _pos=tpor; _n=tpor+fnor; output;
 _dis=0; _nondis=1; _pos=tnor; _n=fpor+tnor; output;
 label _pos='no. correct classified';
 label _n='Total';
 run;

ods output parameterestimates=est covmatparmest=
covmean;
proc nlmixed data=nl_test cov; *corr;
title ''Bivariate analysis of sens and spec using
NLMIXED; &meta'';
parms _sens=1 _spec=1 s2usens=0.1 s2uspec=0.5
covsesp=0;
logitp = (_sens+usens)*_dis + (_spec+uspec)*_nondis;
p=exp(logitp)/(1+exp(logitp));
model _pos~binomial(_n,p);
random usens uspec~normal([0, 0], [s2usens,covsesp,
s2uspec]) subject=IDOR;
run;
ods output close;
```

Bivariate meta-analysis of predictive values

```
data nl_pvtest;
 set _meta;
 _rec+1
```

```
_testpos=1; _testneg=0; _pos=tpor; _n=tpor+fpor;
output;
_testpos=0; _testneg=1; _pos=tnor; _n=fnor+tnor;
output;
label _pos='no. correct classified';
label _n='Total';
run;
ods output parameterestimates=pvest covmatparmest
=pvcovmean;
proc nlmixed data=nl_pvtest df=1000 cov; *corr;
title ''Bivariate analysis of ppv and npv using
NLMIXED &meta'';
 parms _ppv=1 _npv=2 s2uppv=1 s2unpv=0.5
 covppvnpv=-0.5;
  logitp = (_ppv+uppv)*_testpos+(_npv+unpv)
  *_testneg;
  p=exp(logitp)/(1+exp(logitp));
  model _pos~binomial(_n,p);
  random uppv unpv~normal([0, 0], [s2up-
  pv,covppvnpv,s2unpv]) subject=IDOR;
run;
ods output close;
```

References

[1] Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Cochrane diagnostic test accuracy working group: systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889–97.

[2] Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane handbook for systematic reviews of diagnostic test accuracy. Version 1.0. The Cochrane Collaboration; 2010. Available at http://srdta.cochrane.org/. Accessed August 8, 2011.

[3] Guyatt G, Sackett DL, Haynes RB. Evaluating diagnostic tests. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P, editors. Clinical epidemiology. How to do clinical practice research. Philadelphia, PA: Lippincott William & Wilkins; 2006:294–5.

[4] Straus SE, Richardson WS, Glasziou P, Haynes RB. Diagnosis and screening. In: Straus SE, Richardson WS, Glasziou P, Haynes RB, editors. Evidence-based medicine. How to practice and teach EBM. 3rd ed. Oxford, UK: Elsevier; 2005:89–90.

[5] Guggenmoos-Holzmann I, van Houwelingen HC. The (in)validity of sensitivity and specificity. Stat Med 2000;19:1783–92.

[6] Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5–12.

[7] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926–30.

[8] Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002;137:598–602.

[9] Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58:982–90.

[10] Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469–76.

[11] Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002;21:589–624.

[12] Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Stat Med 1993;12:2273–84.

[13] Balk EM, Ioannidis JP, Salem D, Chew PW, Lau J. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. Ann Emerg Med 2001;37:478–94.

[14] Lau J, Ioannidis JP, Balk EM, Milch C, Terrin N, Chew PW, et al. Diagnosing acute cardiac ischemia in the emergency department: a systematic review of the accuracy and clinical effect of current technologies. Ann Emerg Med 2001;37:453–60.

[15] Berger MY, van der Velden JJ, Lijmer JG, de Kort H, Prins A, Bohnen AM. Abdominal symptoms: do they predict gallstones? A systematic review. Scand J Gastroenterol 2000;35:70–6.

[16] Deville WL, van der Windt DA, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. Spine 2000;25:1140–7.

[17] Fiellin DA, Reid MC, O'Connor PG. Screening for alcohol problems in primary care: a systematic review. Arch Intern Med 2000;160:1977–89.

[18] Gould MK, Maclean CC, Kuschner WG, Rydzak CE, Owens DK. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. JAMA 2001;285:914–24.

[19] Hobby JL, Tom BD, Bearcroft PW, Dixon AK. Magnetic resonance imaging of the wrist: diagnostic performance statistics. Clin Radiol 2001;56:50–7.

[20] Hoffman RM, Clanon DL, Littenberg B, Frank JJ, Peirce JC. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. J Gen Intern Med 2000;15:739–48.

[21] Hoogendam A, Buntinx F, de Vet HC. The diagnostic value of digital rectal examination in primary care screening for prostate cancer: a meta-analysis. Fam Pract 1999;16:621–6.

[22] Huicho L, Campos-Sanchez M, Alamo C. Metaanalysis of urine screening tests for determining the risk of urinary tract infection in children. Pediatr Infect Dis J 2002;21:1–11. 88.

[23] Hurley JC. Concordance of endotoxemia with gram-negative bacteremia. A meta-analysis using receiver operating characteristic curves. Arch Pathol Lab Med 2000;124:1157–64.

[24] Kelly S, Harris KM, Berry E, Hutton J, Roderick P, Cullingworth J, et al. A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma. Gut 2001;49:534–9.

[25] Kim C, Kwok YS, Heagerty P, Redberg R. Pharmacologic stress testing for coronary disease diagnosis: a meta-analysis. Am Heart J 2001;142:934–44.

[26] Kwok Y, Kim C, Grady D, Segal M, Redberg R. Meta-analysis of exercise testing to detect coronary artery disease in women. Am J Cardiol 1999;83:660–6.

[27] Lederle FA, Simel DL. The rational clinical examination. Does this patient have abdominal aortic aneurysm? JAMA 1999;281:77–82.

[28] Mitchell MF, Cantor SB, Brookner C, Utzinger U, Schottenfeld D, Richards-Kortum R. Screening for squamous intraepithelial lesions with fluorescence spectroscopy. Obstet Gynecol 1999;94:889–96.

[29] Mol BW, Lijmer JG, van der Meulen J, Pajkrt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. Obstet Gynecol 1999;94:864–9.

[30] Nelemans PJ, Leiner T, de Vet HC, van Engelshoven JM. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. Radiology 2000;217:105–14.

[31] Safriel Y, Zinn H. CT pulmonary angiography in the detection of pulmonary emboli: a meta-analysis of sensitivities and specificities. Clin Imaging 2002;26:101–5.

[32] Sloan NL, Winikoff B, Haberland N, Coggins C, Elias C. Screening and syndromic approaches to identify gonorrhea and chlamydial infection among women. Stud Fam Plann 2000;31:55–68.

[33] Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. Acad Radiol 2001;8:149–57.

[34] Vasquez TE, Rimkus DS, Hass MG, Larosa DI. Efficacy of morphine sulfate-augmented hepatobiliary imaging in acute cholecystitis. J Nucl Med Technol 2000;28:153−5.

[35] Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US—a meta-analysis. Radiology 2000;216:67−77.

[36] Wiese W, Patel SR, Patel SC, Ohl CA, Estrada CA. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. Am J Med 2000;108:301−8.

[37] Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189−202.

[38] Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. Stat Med 2009;28:2384−99.

[39] Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol 1992;45: 1143−54.