Variation of a test's sensitivity and specificity with disease prevalence

Mariska M.G. Leeflang PhD, Anne W.S. Rutjes PhD, Johannes B. Reitsma MD PhD, Lotty Hooft PhD, Patrick M.M. Bossuyt PhD

Abstract

Background: Anecdotal evidence suggests that the sensitivity and specificity of a diagnostic test may vary with disease prevalence. Our objective was to investigate the associations between disease prevalence and test sensitivity and specificity using studies of diagnostic accuracy.

Methods: We used data from 23 meta-analyses, each of which included 10–39 studies (416 total). The median prevalence per review ranged from 1% to 77%. We evaluated the effects of prevalence on sensitivity and specificity using a bivariate random-effects model for each meta-analysis, with prevalence as a covariate. We estimated the overall effect of prevalence by pooling the effects using the inverse variance method.

Results: Within a given review, a change in prevalence from the lowest to highest value resulted in a corresponding change in sensitivity or specificity from 0 to 40 percentage points. This effect was statistically significant ($p < 0.05$) for either sensitivity or specificity in 8 meta-analyses (35%). Overall, specificity tended to be lower with higher disease prevalence; there was no such systematic effect for sensitivity.

Interpretation: The sensitivity and specificity of a test often vary with disease prevalence; this effect is likely to be the result of mechanisms, such as patient spectrum, that affect prevalence, sensitivity and specificity. Because it may be difficult to identify such mechanisms, clinicians should use prevalence as a guide when selecting studies that most closely match their situation.

Competing interests: None declared.

This article has been peer reviewed.

Correspondence to:
Mariska Leeflang,
m.m.leeflang@amc.uva.nl

CMAJ **2013. DOI:10.1503/cmaj.121286**

Diagnostic accuracy plays a central role in the evaluation of medical diagnostic tests. Test accuracy may be expressed as sensitivity and specificity, as positive and negative predictive values or as positive and negative likelihood ratios.[1] Some feel that the positive and negative predictive values of a test are more clinically relevant measures than sensitivity and specificity. However, predictive values directly depend on disease prevalence and can therefore not directly be translated from one situation to another.[2] In contrast, a test's sensitivity and specificity are commonly believed not to vary with disease prevalence.[3–5]

Stability of sensitivity and specificity is an assumption that underlies the use of Bayes theorem in clinical diagnosis. Bayes theorem can be applied in clinical practice by using the likelihood ratio of a test and the probability of the disease before the test was done (pretest probability) to estimate the probability of disease after the test was done.[2] Because likelihood ratios are a function of sensitivity and specificity, it is assumed that the likelihood ratios also remain the same when prevalence varies.

A number of studies have shown that sensitivity and specificity may not be as stable as thought.[6–10] We previously summarized the possible mechanisms through which differences in disease prevalence may lead to changes in a test's sensitivity and specificity.[10] Prevalence affects diagnostic accuracy because of clinical variability or through artifactual differences, as described in the theoretical framework in Table 1. Clinical variability is usually associated with spectrum effects, referral filters or reader expectation. For example, using a test in a more severely diseased population may be associated with a higher prevalence, or with better performance of the test.[6,7] Artifactual differences can result from using additional exclusion criteria, verification bias or an imperfect reference standard. For example, using an imperfect reference standard may lead to an underestimate of diagnostic accuracy, but as prevalence increases, the extent to which this happens will vary.[8,9]

If these associations between prevalence and test accuracy are not just hypothetical, this may have immediate implications for the translation of research findings into clinical practice. It

would imply that sensitivity and specificity of a test, estimated in one setting, cannot unconditionally be translated to a setting with a different disease prevalence. To document the magnitude of these effects, we reanalyzed a series of previously published meta-analyses that included studies of diagnostic test accuracy.

Methods

We included 28 systematic reviews, containing 31 different meta-analyses. These reviews were selected and analyzed for a previously published report on bias and variation in diagnostic accuracy studies; the details of the search process, selection and data extraction are available.[11] In short, we searched several electronic databases for systematic reviews published between January 1999 and April 2002 that met the following criteria: estimation of a diagnostic test's accuracy as the review's objective; included at least 10 original studies of the same diagnostic test; did not exclude primary studies based on design features; and the ability to reproduce the 2×2 tables from the original studies. For the present study, we excluded case–control studies because it is impossible to estimate disease prevalence from such studies.

Statistical analyses

Sensitivity and specificity of a test move in opposite directions when the test-positivity threshold varies. Methods for meta-analyses of diagnostic accuracy should take this threshold effect into account.[12] We used the bivariate logitnormal random-effects model, which allows for this correlation between sensitivity and specificity,[13] and models the logits of sensitivity and specificity. The logit is the natural logarithm of sensitivity (or specificity) divided by 1 minus sensitivity (or specificity). In our results, we back-transformed these estimates to the original 0 to 100 scale for sensitivity and specificity.

For each of the eligible meta-analyses, we fitted the model to generate summary estimates of sensitivity and specificity. To evaluate the association between prevalence and sensitivity and specificity, we included prevalence as a continuous covariate in the model.

We also calculated the summary effect of prevalence on sensitivity and specificity, by pooling the effects across meta-analyses using the inverse variance method.[14]

Based on the summary estimates for sensitivity and specificity in each model and the estimated effect of prevalence in that model, we calculated the range in sensitivity and specificity for the observed range in prevalence of that meta-analysis. We graphically plotted these ranges.

To investigate whether study characteristics, rather than prevalence, could explain the heterogeneity, we reanalyzed the data from the reviews in which a significant association for prevalence

Table 1: Theoretical framework of how disease prevalence and test accuracy may be related[10]

Factor	Effect on prevalence	Effect on accuracy
Clinical variability		
Patient spectrum	• Distribution of symptoms and severity may change with varying prevalence	• Differences in symptoms and severity influences sensitivity and specificity
Referral filter	• How and through what care pathway patients are referred may influence the spectrum of disease in the population	• A change in setting and patient spectrum may also alter a test's sensitivity and specificity
Reader expectations	• Prevalence influences reader expectations: if one knows that the prevalence should be high, then one's intrinsic threshold may be lowered	• Changing one's intrinsic threshold will influence accuracy
Artifactual variability		
Distorted inclusion of participants	• Excluding patients with difficult to diagnose conditions may influence the prevalence	• Excluding patients with difficult to diagnose conditions will overestimate the accuracy of a test
Verification bias	• If not all patients receive the (same) reference standard, this influences prevalence	• Verification bias has an effect on test accuracy
Imperfect reference standard	• Prevalence will be over- or underestimated	• Test accuracy may be underestimated; the extent of which varies with prevalence

was been found. In these reviews, we tested, one at a time, whether study characteristics were associated with accuracy.

We added the following characteristics to the model: setting; patient-referral pattern; consecutive enrolment; exclusion of patients with difficult-to-diagnose conditions; differential verification; and partial verification. Setting was scored as primary, secondary or tertiary care. Referral could be based on symptoms, results of an index test or another test. Differential verification was considered present if the results of the index test were verified by use of different reference standards for positive versus negative index test results. We considered partial verification to be present if not all patients underwent testing with the reference standard. To avoid problems in convergence and unstable estimates, we added covariates to the model only if there was a minimum of 3 studies at each level of the covariate.

We compared the goodness-of-fit of a model that had prevalence as a covariate with the goodness-of-fit of a model for the same data that had another study characteristic as a covariate. We expressed goodness of fit as the Akaike information criterion, with a lower Akaike information criterion value indicating a better fit of the model.

Analyses were done in SAS for Windows, version 9.2, using the PROC NLMIXED procedure. The syntaxes are available on request from the first author. We calculated the summary effect using Review Manager 5.2 (The Cochrane Collaboration, 2012). We considered p values less than 0.05 to be significant.

Results

Of the 31 meta-analyses in our dataset, we excluded 8 because they contained less than 10 eligible studies. The final dataset consisted of 23 meta-analyses of diagnostic test accuracy, all involving different medical tests; these analyses contained data from 416 individual studies.[15–38] More information about the included meta-analyses can be found in the Appendix 1 (available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.121286/-/DC1).

The disease prevalence ranged from 0.1% to 98% in the 416 included studies, with a median of 37%. Across the 23 meta-analyses, the median prevalence varied from 1% to 77% (Figure 1).

Figure 2 summarizes the associations between prevalence and logit sensitivity or logit specificity. In 8 of the 23 meta-analyses (35%) we observed a significant association between prevalence and either logit sensitivity or logit specificity, or both. In all cases, a higher prevalence accompanied a lower specificity. In the 2 meta-analyses with a significant association between prevalence and sensitivity,[27,33] sensitivity was higher with higher prevalence.

Overall, there was a significant association between specificity and prevalence. Based on the pooled estimate, logit specificity decreased on average by 0.02 units (95% confidence interval –0.03 to –0.01) for every 1 percentage point increase in prevalence. This corresponds with a decrease in specificity of between 0.1 and 0.5 percentage points; the effect is larger for a specificity around 50% and smaller for specificity around 95%. There was no significant overall association between prevalence and sensitivity.

We compared the meta-analyses in which prevalence had a significant effect with those in which prevalence had no effect. There were no systematic differences between these 2 sets of meta-analyses in terms of the type of test used, the range of prevalences or sample size.

Figure 3 shows the differences in sensitivity and specificity when moving from the lowest reported prevalence in each review to the highest prevalence in the same review. For example, in the review by Safriel and colleagues,[33] the lowest prevalence was 25% and the highest prevalence was 82%. The estimated sensitivity in the study with the lowest prevalence was 69%; in the study with the highest prevalence in the same review, the sensitivity estimate was 98%. Estimated specificity varied between 97% and 58% over the same prevalence range.

We reanalyzed the 8 reviews in which prevalence was significantly associated with logit specificity by including the study characteristics, one by one, as covariates in the model. The results of these analyses are summarized in Appendix 2 (available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.121286/-/DC1). For one review, it was not possible to achieve model convergence, probably because of an almost perfect correlation between sensitivity and specificity. In 6 of the remaining 7 reviews, prevalence explained the variation in logit specificity better (i.e., a smaller Akaike information criterion, indicating a better goodness-of-fit) than any other study characteristic analyzed. For only the review of nuchal translucency in the diagnosis of Down syndrome,[31] models that included study characteristics as covariates fit marginally better than did a model with prevalence as a covariate.

Interpretation

In this reanalysis of test accuracy reviews, we found significant associations between prevalence and sensitivity or specificity in 1 out of

every 3 reviews. Overall, specificity was lower in studies with higher prevalence. We found an association more often with specificity than with sensitivity, implying that differences in prevalence mainly represent changes in the spectrum of people without the disease of interest.

We do not and cannot claim that changes in prevalence cause differences in sensitivity and specificity. Because sensitivity is estimated in people with the disease of interest and specificity in people without the disease of interest, changing the relative number of people with and without the disease of interest should not introduce systematic differences. Therefore, the effects that we found may be generated by other mechanisms that affect both prevalence and accuracy, as we described earlier.[10] In practice, it will be difficult to identify these mechanisms. Poor

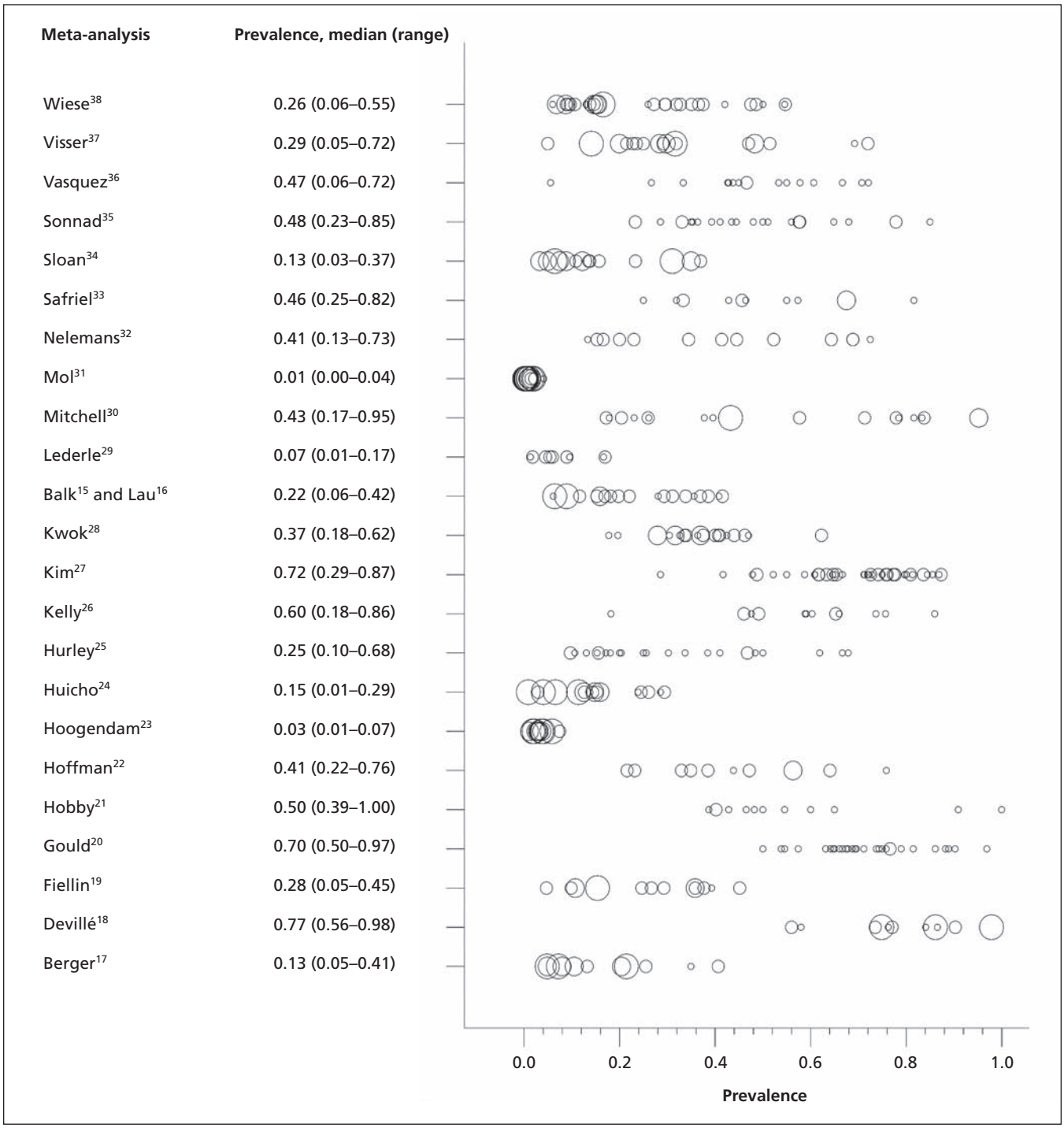

Figure 1: **Prevalence estimates for each primary study in the 23 included meta-analyses. The size of the circle reflects the study size: < 100 participants; 100–500 participants; 500–1000 participants; and > 1000 participants. Prevalence is shown as a proportion.**

reporting of the design and patient characteristics in studies of test accuracy is a common problem, and more recent primary studies are only slightly better reported.[39] Furthermore, there may be intricate relations between patient features, study setting and prevalence, making it difficult to disentangle the separate contributing factors.

For clinicians using Bayes theorem in evidence -based medicine to translate patient-based pretest probabilities to posttest probabilities, our results may foster caution. For example, the results from a study with a prevalence of 45% may not necessarily justify the calculation of posttest probabilities for patients with low pretest probabilities in a setting where the prevalence of disease is only 5%. The consequences of these differences for practice will vary, depending on the extent of differences in accuracy across setting.

As an illustration, imagine a clinician who would like to know how a positive result of a spiral computed tomography scan would increase the probability of a patient having pulmonary embolism. In the review by Safriel and colleagues,[33] the prevalence of pulmonary embolism ranged from 25% to 82%, and the positive likelihood ratio varied from 2.35 to 25.4, with the highest likelihood ratio coming from the study with the lowest prevalence. If the prevalence in the clinician's situation was at the lower end of the spectrum (e.g., around 30%),

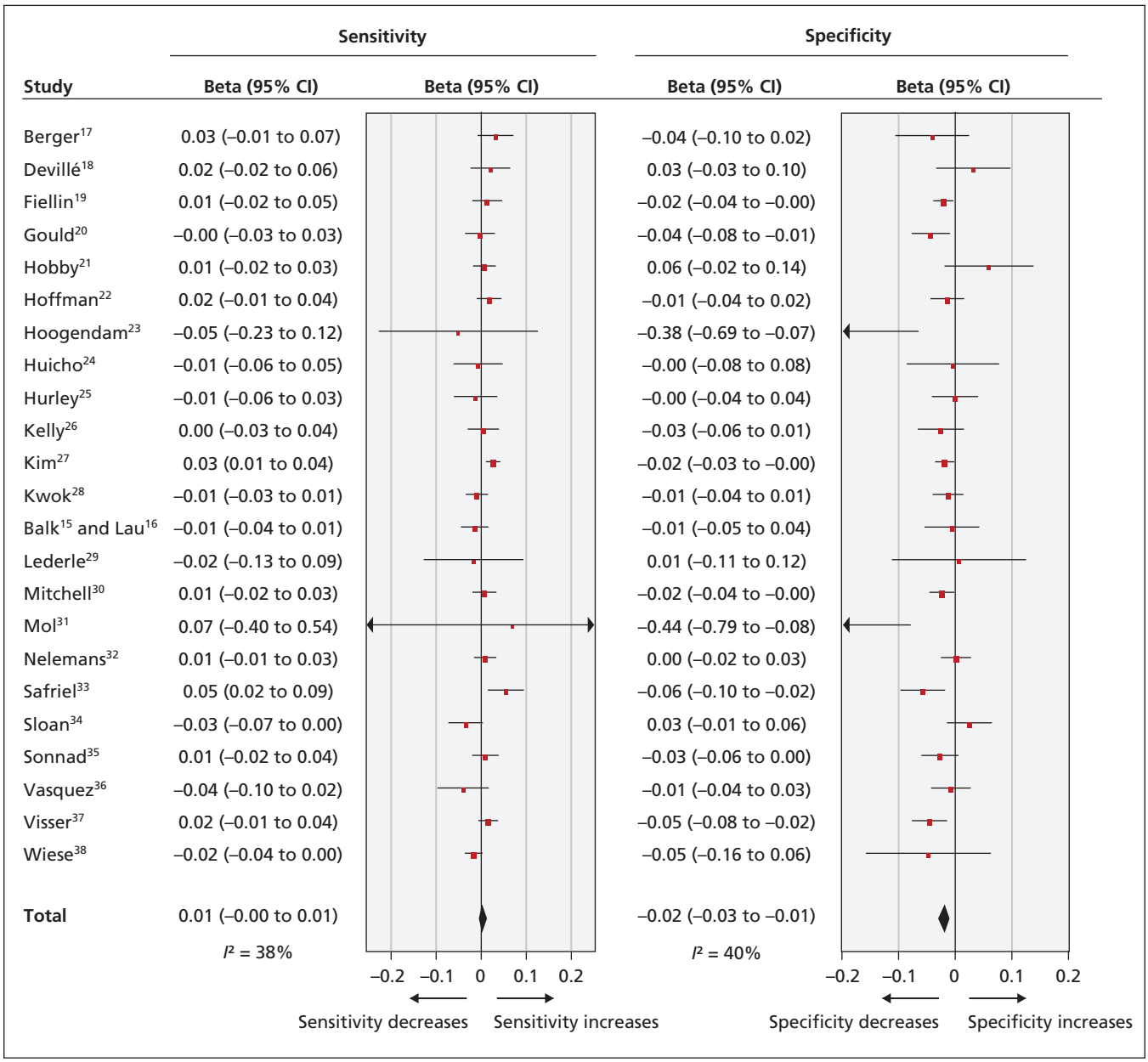

Figure 2: The effect of prevalence on logit sensitivity and specificity. Prevalence effects on logit sensitivity and specificity are shown per 1%. Beta reflects the effect size. CI = confidence interval.

the positive likelihood ratio would be 18 and the posttest probability of pulmonary embolism with a positive result would be 89%. However, if the clinician had used evidence from a paper in which the prevalence was closer to the higher end (e.g., 57%), the results would have been different. A prevalence of 57% corresponds with a likelihood ratio of 6.2. Applying the likelihood ratio of 6.2 to a prevalence of 30% would lead to a posttest probability of 73%, which in some situations may lead to a different decision than a posttest probability of 89%. Applying the Bayes rule requires that clinicians have an idea of the disease prevalence among patients suspected of having the condition of interest.

Comparison with other studies

In the late 1970s, it was already noted that sensitivity and specificity may not be as stable as thought.[40] Most of the publications addressing variability of sensitivity and specificity focus on spectrum differences as an explanation.[6–8] Another explanation often referred to is the variation in reference standards, or in disease definition. Variation in disease definition may cause variation in patient spectrum, in accuracy and in

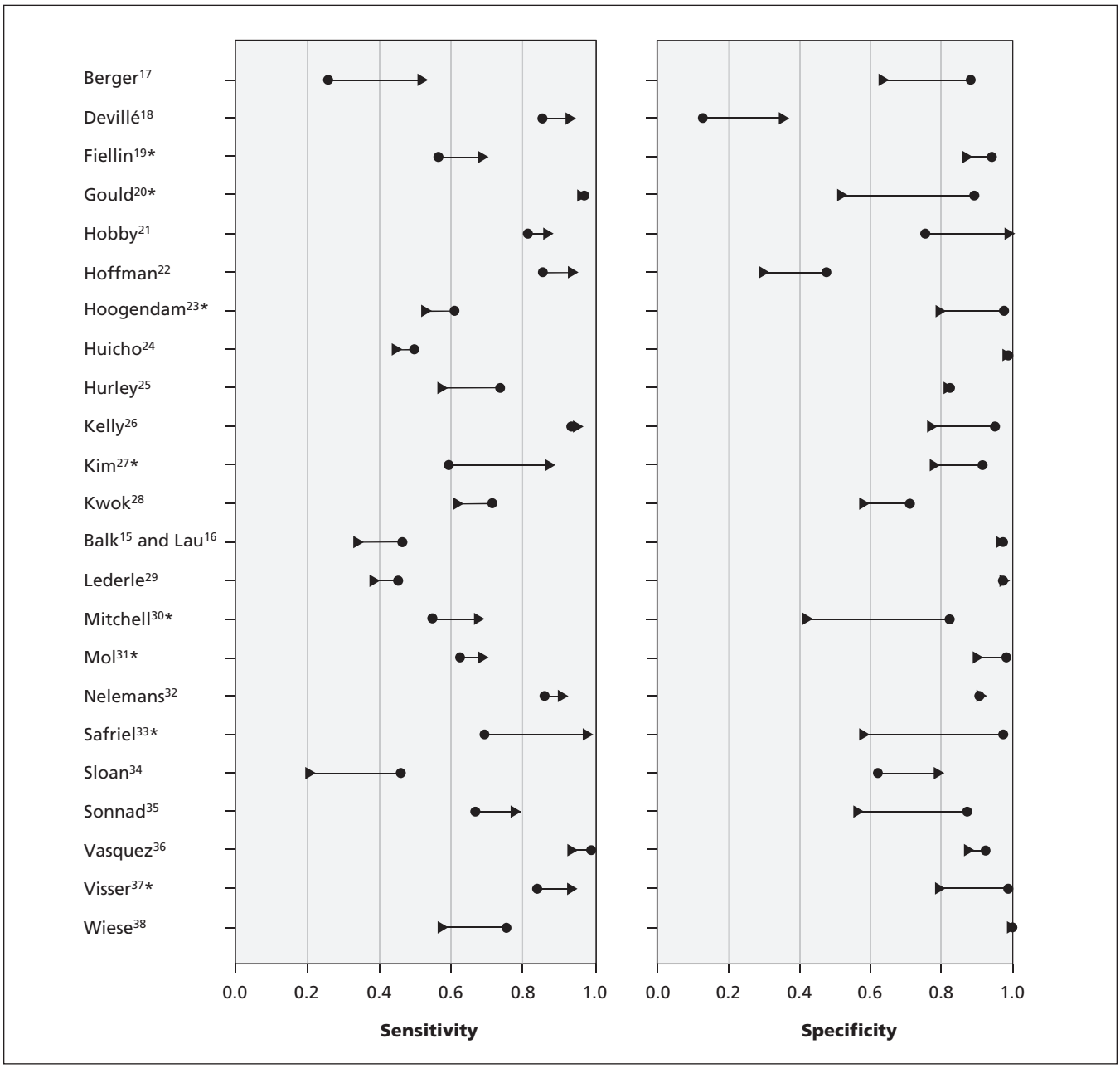

Figure 3: Change in sensitivity and specificity with increasing prevalence. The lines represent sensitivity and specificity at the minimum and maximum prevalence in each meta-analysis. Sensitivity and specificity are shown as proportions. The circles reflect sensitivity and specificity at the lowest prevalence, and the arrowheads reflect sensitivity and specificity at the highest prevalence.*p value < 0.05

prevalence.[8,9] Most of these publications are based on theoretical reasoning and a few examples. This study adds empirical evidence from a range of diseases and tests. Regarding the intricate relations between all contributing factors and the poor reporting of these factors in accuracy studies, we focused on prevalence, being a relatively well-reported factor.

Limitations

Although the 23 included systematic reviews, published from 1999 to April 2002, cover 416 studies of test accuracy, this still constitutes a moderate sample size, relative to the full body of diagnostic accuracy research. These reviews cover a wide range of medical tests and conditions and had not been selected based on suspected prevalence effects. Still, we must acknowledge that the prevalence observed in the included studies may not always be a good reflection of the disease prevalence in the actual patient population from which study participants were sampled. This will especially be the case if the study was not based on consecutive enrolment of eligible patients. The most extreme example would be a case–control study, in which the prevalence of the target condition in the study group is artificially set by design. For this reason, we excluded these case–control designs from our analyses.

Conclusion

Sensitivity and specificity of a test often vary with prevalence, likely due to mechanisms that affect both prevalence and sensitivity and specificity, such as patient spectrum. Therefore, investigators are invited to think of the intended use of the test when designing a study of test accuracy, and specify the inclusion criteria that define the study population accordingly.[41] If the accuracy study recruited patients from different settings, a separate sensitivity–specificity pair for each setting could be reported, for example. Once the study has been completed, the Standards for the Reporting of Diagnostic Accuracy Studies (STARD) checklist can be of help in achieving complete and informative reporting.[42]

Our results have implications for clinicians who turn to the medical literature for estimates of the accuracy of a test. Physicians should try to identify the study that most closely matches their setting. In doing so, they should rely on the definition of the target condition, inclusion criteria and prior testing, but they should also use the reported prevalence in the study as a guide when evaluating the applicability of the study, providing that the disease prevalence in the intended population is known.

References

1. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res* 2002;2:4.
2. Straus SE, Richardson WS, Glasziou P, et al. Diagnosis and screening. In: *Evidence-based medicine: how to practice and teach EBM*. 3rd ed. Oxford (UK): Elsevier; 2005:89-90.
3. Guyatt G, Sackett DL, Haynes RB. Evaluating diagnostic tests. In: *Clinical epidemiology: how to do clinical practice research.* Haynes RB, Sackett DL, Guyatt GH, et al. editors. New York (NY): Lippincott William and Wilkins; 2006:294-5.
4. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *J Eval Clin Pract* 2006;12:132-9.
5. Scales CD Jr, Dahm P, Sultan S, et al. How to use an article about a diagnostic test. *J Urol* 2008;180:469-76.
6. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med* 2002;137:598-602.
7. Feinstein AR. Misguided efforts and future challenges for research on "diagnostic tests." *J Epidemiol Community Health* 2002;56:330-2.
8. Szklo M, Nieto J. Quality assurance and control. In: *Epidemiology: beyond the basics.* Burlington (MA): Jones and Bartlett Learning; 2007:315-7.
9. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* 1997;16:981-91.
10. 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.
11. Rutjes AW, Reitsma JB, Di Nisio M, et al. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76.
12. Macaskill P, Gatsonis C, Deeks JJ, et al. Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. *Cochrane handbook for systematic reviews of diagnostic test accuracy.* Version 1.0. Oxford (UK): The Cochrane Collaboration; 2010. Available: http://srdta.cochrane.org
13. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982-90.
14. Higgins JPT, Green S, editors. *Cochrane handbook for systematic reviews of interventions*. 5.1.0 ed. Oxford (UK): The Cochrane Collaboration; 2011. Available: www.cochrane-handbook.org (accessed 2012 Mar. 27).
15. Balk EM, Ioannidis JP, Salem D, et al. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. *Ann Emerg Med* 2001;37:478-94.
16. Lau J, Ioannidis JP, Balk EM, 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.
17. Berger MY, van der Velden JJ, Lijmer JG, et al. Abdominal symptoms: Do they predict gallstones? A systematic review. *Scand J Gastroenterol* 2000;35:70-6.
18. Devillé WL, van der Windt DA, Dzaferagic A, et al. The test of Lasègue: systematic review of the accuracy in diagnosing herniated discs. *Spine* 2000;25:1140-7.
19. Fiellin DA, Reid MC, O'Connor PG. Screening for alcohol problems in primary care: a systematic review. *Arch Intern Med* 2000;160:1977-89.
20. Gould MK, Maclean CC, Kuschner WG, et al. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *JAMA* 2001;285:914-24.
21. Hobby JL, Tom BD, Bearcroft PW, et al. Magnetic resonance imaging of the wrist: diagnostic performance statistics. *Clin Radiol* 2001;56:50-7.
22. Hoffman RM, Clanon DL, Littenberg B, et al. 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.
23. 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.
24. 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.
25. 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.
26. Kelly S, Harris KM, Berry E, et al. A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma. *Gut* 2001;49:534-9.
27. Kim C, Kwok YS, Heagerty P, et al. Pharmacologic stress test-

ing for coronary disease diagnosis: a meta-analysis. *Am Heart J* 2001;142:934-44.
28. Kwok Y, Kim C, Grady D, et al. Meta-analysis of exercise testing to detect coronary artery disease in women. *Am J Cardiol* 1999;83:660-6.
29. Lederle FA, Simel DL. The rational clinical examination. Does this patient have abdominal aortic aneurysm? *JAMA* 1999;281:77-82.
30. Mitchell MF, Cantor SB, Brookner C, et al. Screening for squamous intraepithelial lesions with fluorescence spectroscopy. *Obstet Gynecol* 1999;94:889-96.
31. Mol BW, Lijmer JG, van der Meulen J, et al. Effect of study design on the association between nuchal translucency measurement and Down syndrome. *Obstet Gynecol* 1999;94:864-9.
32. Nelemans PJ, Leiner T, de Vet HC, et al. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. *Radiology* 2000;217:105-14.
33. 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.
34. Sloan NL, Winikoff B, Haberland N, et al. Screening and syndromic approaches to identify gonorrhea and chlamydial infection among women. *Stud Fam Plann* 2000;31:55-68.
35. 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.
36. Vasquez TE, Rimkus DS, Hass MG, et al. Efficacy of morphine sulfate-augmented hepatobiliary imaging in acute cholecystitis. *J Nucl Med Technol* 2000;28:153-5.
37. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US — a meta-analysis. *Radiology* 2000;216:67-77.
38. Wiese W, Patel SR, Patel SC, et al. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. *Am J Med* 2000;108:301-8.
39. Smidt N, Rutjes AW, van der Windt DA, et al. The quality of diagnostic accuracy studies since the STARD statement: Has it improved? *Neurology* 2006;67:792-7.
40. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978; 299:926-30.
41. Irwig L, Bossuyt P, Glasziou P, et al. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002; 324:669-71.
42. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative: Standards for Reporting of Diagnostic Accuracy. *Ann Intern Med* 2003;138:40-4.

Affiliations: Department of Clinical Epidemiology (Leeflang, Bossuyt), Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; the Institute for Social and Preventive Medicine (Rutjes), University of Bern, Bern, Switzerland; the Clinical Center for Aging Sciences (Rutjes), University G. d'Annunzio Foundation, Chieti, Italy; the Julius Center for Health Sciences and Primary Care (Reitsma), University Medical Center Utrecht, The Netherlands; and the Dutch Cochrane Centre (Hooft), Amsterdam, The Netherlands.

Contributors: Mariska Leeflang and Lotty Hooft conceived the study design. Anne Rutjes developed and maintained the study databases and data-extraction forms. Anne Rutjes, Johannes Reitsma and Patrick Bossuyt performed data collection and quality assessment. Mariska Leeflang and Patrick Bossuyt interpreted the data. Mariska Leeflang drafted the manuscript, which was revised for important intellectual content by all of the authors. All of the authors approved the final version submitted for publication.

Funding: Mariska Leeflang is supported by The Netherlands Organization for Scientific Research (NWO); project 916 .10.034). Anne Rutjes was supported by a research grant from the NWO (registration no. 945–10–012). The funding bodies had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Acknowledgements: The authors thank Dr. Marcello Di Nisio, Dr. Jeroen C. van Rijn and Dr. Nynke Smidt for their contribution to the collection of data. The authors also thank Marga Goris, Evelien Roekevisch and Carlinde de Ruiter for their comments on earlier versions of the manuscript.