Systematic reviews of diagnostic test accuracy
Leeflang, M.M.G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Diagnostic accuracy may vary with prevalence: Implications for evidence-based diagnosis

Mariska M.G. Leeflang, Patrick M.M. Bossuyt, Les Irwig

Accepted by J Clin Epidemiol
Abstract

**Background:** Sensitivity and specificity of diagnostic tests are often assumed to be independent of prevalence. Yet several studies and systematic reviews have reported results that indicate otherwise.

**Methods:** We identify and explore mechanisms that may be responsible for sensitivity and specificity varying with prevalence and illustrate them with examples from the literature.

**Results:** Clinical and artefactual variability may be responsible for changes in prevalence and accompanying changes in sensitivity and specificity. Clinical variability refers to differences in the clinical situation that may cause sensitivity and specificity to vary with prevalence. For example, a patient population with a higher disease prevalence may include more severely diseased patients, in which the test performs better. Artefactual variability refers to effects on prevalence and accuracy associated with study design, for example the verification of index test results by a reference standard. Changes in prevalence influence the extent of overestimation due to imperfect reference standard classification.

**Conclusions:** Sensitivity and specificity may vary in different clinical populations, and prevalence is a marker for such differences. Clinicians are advised to base their decisions on studies that most closely match their own clinical situation, using prevalence to guide the detection of differences in study population or study design.
5.1 Introduction

Diagnostic test accuracy refers to the ability of a test to discriminate between those who have and those who do not have the target condition. Accuracy is assessed by comparing the results of the index test, the test under evaluation, with the results of the reference standard, which aims to classify patients as having or not having the target condition. Test accuracy is most often expressed as the test’s sensitivity (the proportion of those with the target condition who have a positive index test result) and specificity (the proportion of those without the target condition who have a negative index test result).

A test’s sensitivity and specificity are commonly believed not to vary with disease prevalence. Yet a number of studies have shown that differences in diagnostic accuracy often accompany differences in prevalence between study groups (see Table 5.1 for examples). For example, Flicker and colleagues used a consensus diagnosis as the reference standard in assessing the diagnostic accuracy of different checklists for dementia. They found a lower sensitivity as well as a lower specificity in study groups with a greater prevalence. The opposite effect has also been reported. A study of Magnetic Resonance Imaging to diagnose multiple sclerosis, reported both a higher sensitivity as well as a higher specificity in the study group with a greater prevalence. On the other hand, when the general results of this study are compared with those of another study with greater prevalence of multiple sclerosis, the latter study reported a lower sensitivity. Lachs and colleagues, studied dipstick tests in patients suspected of urinary tract infection and found a higher sensitivity and a lower specificity with greater prevalence.

Greater prevalence can be associated with both higher as well as lower sensitivity and specificity. In this paper, we explain some of the underlying mechanisms that can lead to changes in both disease prevalence and in diagnostic accuracy (see Figure 5.1). Prevalence variability itself, as well as the study characteristics that cause prevalence differences, can result either in clinical or artefactual variation.

<table>
<thead>
<tr>
<th>First Author (year)</th>
<th>Target Condition</th>
<th>Index Test</th>
<th>Reference Standard</th>
<th>Prevalence</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flicker (1997)</td>
<td>Dementia</td>
<td>Checklists</td>
<td>Consensus diagnosis</td>
<td>41% vs. 72%</td>
<td>78 to 73 (%)</td>
<td>88 to 71 (%)</td>
</tr>
<tr>
<td>O’Connor (1996)</td>
<td>MS</td>
<td>MRI</td>
<td>Expert panel</td>
<td>‘higher probability’</td>
<td>20 to 70% (↑)</td>
<td>80 to 93% (↑)</td>
</tr>
<tr>
<td>Lee (1991)</td>
<td>MS</td>
<td>MRI</td>
<td>Clinical follow-up</td>
<td>43% vs. 53%</td>
<td>84 to 58% (↑)</td>
<td>63 to 91% (↑)</td>
</tr>
<tr>
<td>Lachs (1992)</td>
<td>Urinary tract infection</td>
<td>Dipstick</td>
<td>Culture</td>
<td>7% vs. 50%</td>
<td>56 to 92% (↑)</td>
<td>78 to 42% (↑)</td>
</tr>
</tbody>
</table>

Abbreviations: MS=Multiple Sclerosis; MRI=Magnetic Resonance Imaging; ↓= lower; ↑= higher.
in test accuracy, the latter being a consequence of imperfections in study design or execution.

We will first discuss clinical variation, then artefactual variation. In the closing section we will provide guidance for both readers and researchers about how to deal with prevalence differences in study populations and the translation into practice.

**5.2 Clinical variability in prevalence and test accuracy**

Clinical variability refers to diagnostic test accuracy varying with prevalence because of differences in the patients or the characteristics of the setting in which those patients are being assessed.

**5.2.1 Patient Spectrum**

Both disease prevalence and test accuracy may be associated with patient spectrum, a term that denotes the severity of disease or the range of comorbidities in the patients studied. Flicker and colleagues studied the diagnostic accuracy of different checklists for dementia in two different settings: a screening group that consisted of elderly people, with memory difficulties, from the general population, with a dementia prevalence of 41%, and a diagnostic care group that consisted of people who were already more or less mentally disabled, with a dementia prevalence of 72%. It is likely that distinguishing patients with dementia from those without...
Dementia was more difficult in the diagnostic care setting. The more the underlying conditions in these patients look alike, the more false positive results as well as false negative results will be encountered. This is reflected by the lower sensitivity (73% versus 78%) and the lower specificity (71% versus 88%) in the diagnostic care group.

Not only comorbidities affect a test’s ability to distinguish people with the target condition from those without this target condition. Many target conditions represent an underlying continuum, ranging from ‘barely present’ to ‘clearly present’. It is possible that the shape of the distribution of the underlying continuum varies with disease prevalence. For example, in situations where disease is common, the distribution may be skewed towards the ‘clearly present’ end of the spectrum, and sensitivity is higher.

Weiner and colleagues reported on the diagnostic accuracy of an exercise test for coronary artery disease in the coronary artery surgery study. Patients in this study were divided into three groups, based on their symptoms: definite angina, probable angina or nonischemic pain. The prevalence of coronary artery disease was 89% in males with definite angina, 70% in males with probable angina and 22% in males with nonischemic pain. The higher the chances that the symptoms are a manifestation of coronary artery disease, the higher the likelihood that the coronary artery disease is severe, and that a person can be correctly identified as having coronary artery disease. In definite angina, sensitivity will be higher. On the other hand, it will be more difficult to correctly classify persons with definite angina as not having coronary artery disease, so specificity can be expected to be lower. The sensitivity in the group of men who displayed definite angina was 85% while their specificity was 67%, in the group with probable angina sensitivity was 75% and specificity was 74%, and in the group with nonischemic pain, the sensitivity was 54% and specificity 76%.

5.2.2 Referral Filter

Differences in patient spectrum may be caused by differences in study population and clinical setting, but also by prior testing of patients before they are enrolled in the study. Possible effects of prior testing of patients were nicely shown in two studies on the diagnostic accuracy of a diagnostic protocol for children suspected of having appendicitis.

Kosloske and colleagues reported a relatively high sensitivity (99%) in their study, compared to other appendicitis studies. This coincided with a greater prevalence (59%). Although Kosloske claimed that prior testing of children more likely to have appendicitis did not affect sensitivity and specificity, Swarr and Keren pointed out in a comment that prescreening and the related prevalence change had influenced the diagnostic accuracy estimates in this appendicitis study. The severity of the appendicitis is associated with the displayed symptoms, and with the ability of either a general practitioner or an emergency doctor to correctly refer only those
children that really have appendicitis (see Figure 5.2). Children that are more likely to have appendicitis may express clearer symptoms, resulting in a higher sensitivity. On the other hand, Peña and colleagues sent the children that were most likely to have appendicitis directly to surgery and did not include them to study the accuracy of the diagnostic protocol. They therefore reported a lower prevalence than Kosloske et al.: 36%. This may have resulted in less clear appendicitis cases in the study group, thus leading to more false negatives and a lower sensitivity (94%).

In the above example, prior testing of children led to differences in patient spectrum and thus to variation in prevalence and diagnostic accuracy. More generally, the referral filter is the diagnostic pathway that determines which patients will be referred to the setting where they will be enrolled in the study. Even without result-
Diagnostic accuracy may vary with prevalence

ing in apparent differences in patient spectrum, a different referral filter can lead to changes in prevalence and in diagnostic test accuracy.

In a study on the diagnostic accuracy of ultrasound for the detection of breast cancer in young symptomatic women, using mammography as referral filter may affect the prevalence of breast cancer in the study group as well as the diagnostic accuracy of ultrasound. Tables 5.2a and 5.2b show the effect of mammography as a referral filter when assessing the accuracy of ultrasound. Test accuracies derived from a study by Houssami et al. have been applied to a hypothetical population with a breast cancer prevalence of 9%\(^{10}\). The sensitivity of ultrasound is 82% (82/100) in the overall population. If only mammography positive women are referred for ultrasound, the prevalence increases to 39%. The sensitivity of ultrasound in the mammography-positive group (62/76: 82%) is identical to the overall sensitivity before referral. This is because the errors of mammography and ultrasound in detecting disease are not associated. However, the errors of mammography and ultrasound in declaring breast cancer to be absent are correlated. Hence, the specificity of ultrasound in the overall group differs markedly (880/1000: 88%) from that in women who were positive on mammography (90/120: 75%). With positive mammography as the referral filter, the specificity of ultrasound is lower. These correlated errors may well occur for other reasons than spectrum differences.

5.2.3 Reader Expectation

In 1990, Gianrossi and colleagues reported a meta-analysis on cardiac fluoroscopy to diagnose coronary artery disease\(^{11}\). They found a lower sensitivity in studies with greater prevalence, without any apparent reason for this difference. They reasoned that clinicians who were used to a lower disease prevalence than the prevalence in this study population, were less likely to indicate a patient as having coronary artery disease based on fluoroscopy abnormalities.

Variation in prevalence can be a cause of accuracy differences when it influences the implicit threshold that clinicians use when they judge for example radiological images. This is called reader expectation and may be expected when clinicians switch to a setting with a different prevalence than they were used to. In response to

<table>
<thead>
<tr>
<th>5.2a Breast Cancer</th>
<th>5.2b No Breast Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>M+</td>
<td>62</td>
</tr>
<tr>
<td>M-</td>
<td>20</td>
</tr>
<tr>
<td>Totals</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 5.2a displays the results for women having breast cancer. Table 2b displays the results for women having no breast cancer. US = ultrasound. M=mammography.
prevalence, the clinicians may alter their threshold for declaring a perceived characteristic as abnormal. Compared to those who work in screening, physicians who are more involved in diagnostic examinations (and less involved in screening) may expect higher underlying rates of cancer when reading screening mammograms. This would lead to a lower false negative rate and a higher false positive rate, thus leading to a lower sensitivity and a higher specificity in reading.

5.3 Artefactual variability in prevalence and test accuracy

Artefactual variability refers to changes in prevalence due to imperfections in the design and execution of a study. Crucial study design features that relate to both prevalence and diagnostic accuracy are a distorted inclusion of participants in the study and misclassification in the reference standard used for verification of the index test results.

5.3.1 Distorted Inclusion of Participants

Ideally a diagnostic accuracy study includes all patients within a specific period who are suspected of having the target condition and in whom using the test would be considered (consecutive enrollment). This will result in a patient spectrum that reflects as much as possible the range of patients that a clinician will see in practice. Distortion of this ideal inclusion pattern may artefactually affect the prevalence of the target condition in the study group as well as the accuracy of the diagnostic test under study.

The most extreme form of distorted patient inclusion occurs when persons who have the target condition are sampled from a completely different population than the persons who do not have the target condition. Such a design approach, often called a case-control design, can be used without bias if there is appropriate sampling. On the other hand, these two-gate designs can be a serious source of bias when cases and controls are sampled from two different populations.

Medeiros and colleagues demonstrated the effects of a two-gate design in a study on the diagnostic accuracy of several tests to diagnose glaucoma. They compared two different methods of patient recruitment: the first method comprised consecutive enrollment of patients and the other method was a two-gate design. With the two-gate design, sensitivity was calculated in a group of relatively severe glaucoma patients, while specificity was calculated in a group of healthy volunteers. With the consecutive enrollment design a study group was assembled consisting of patients all suspected of having glaucoma. The Glaucoma Probability Score, which is an automated device to detect glaucomatous damage, had a higher sensitivity and spe-
Diagnostic accuracy may vary with prevalence

cificity with the two-gate design (sensitivity of 64% and specificity of 95%) than in the consecutive enrollment design (sensitivity of 35% and specificity of 86%).

In a study with the two-gate design, the researchers selected the cases and the controls themselves and they determined the apparent prevalence. In such an example, the prevalence in the study group is determined by the study design and may not reflect the population prevalence as seen in practice. In other studies, the effect of distorted selection may be more subtle and the prevalence in the study group may seem to reflect the prevalence as seen in practice.

The latter is demonstrated in a study of ultrasound for diagnosing epididymitis with retrospective selection of patients\(^\text{15}\). Four different strategies to select patients with epididymitis from existing data files resulted in prevalences ranging from 23% when the broadest selection method was used to 76% with the narrowest selection method. With greater prevalence, sensitivity decreased from 83% to 76% and specificity decreased from 97 to 79%.

If more conditions were included, such as testicular torsion, orchitis or testicular carcinoma, prevalence was lower and it was easier for the readers of the ultrasound images to differentiate between patients who had epididymitis and patients who had another scrotal disease. When they only looked at patients with epididymitis or epididymo-orchitis in the differential diagnosis, the prevalence of epididymitis was greater but it also became more difficult to differentiate between patients with and without epididymitis.

Exclusion of patients can also have effects in the opposite direction. By excluding related target conditions that challenge a test’s ability to detect the target condition as well as the ability to identify the patients without the target condition, a test’s sensitivity and specificity will be higher but so will prevalence. By excluding these related conditions from the study or the subsequent analyses, a test may seem to perform better, a phenomenon known as limited challenge\(^\text{16}\). Note that what will be called limited challenge in one situation may be called a difference in patient spectrum in another situation. Excluding obese patients in a study on the accuracy of ultrasound can be regarded as an example of limited challenge, as it is more difficult to distinguish abnormalities by ultrasound in obese patients than in non-obese patients. On the other hand, if the ultrasound is not used in obese patients, the exclusion of obese patients and the resulting diagnostic accuracy of the ultrasound will fairly reflect the clinical situation.

5.3.2 Verification Bias

Diagnostic test accuracy is assessed by verifying the results of the test under evaluation with the result of a single reference standard in every patient in the study. Verification bias occurs if not all patients are verified, or if some patients are verified by a second or a third reference standard.
An example of verification bias can be found in the meta-analysis of Mol and colleagues. They assessed the accuracy of nuchal translucency measurement for Down syndrome detection. Some studies used two reference standards: fetal karyotyping in fetuses with an increased nuchal translucency, whereas pregnancy outcome was awaited in fetuses that showed a normal measurement. In the studies with such a verification bias, the prevalence ranged from 0.1 to 0.9% (pooled estimate 0.4%). In studies without verification bias, the prevalence ranged from 0.2 to 2.3% (pooled estimate 1.1%). The pooled sensitivity in the last group of studies was found to be lower (55%) than that in the studies with verification bias (77%), whereas specificity remained unaffected (96% and 97%).

In these studies, test positives were verified with another reference standard than the test negatives. This is called differential verification: some participants receive a different reference standard, conditional on the index test result. The reference standard in the test negatives was follow-up. Between testing and birth, fetuses may be aborted (and thus be excluded from the analysis) or Down syndrome may not have been recognized directly at birth. In general, the effects of differential verification very much depend on the uniformity in the characteristics of all used reference standards.

Another form of verification bias is partial verification. Partial verification occurs when not all participants receive the reference standard. The effects on prevalence and diagnostic accuracy depend on whether the reference standard was randomly allocated to patients or not. In the majority of studies with partial verification, most index test positive cases are verified, whereas index test negative cases are likely to be verified only in case of an increased pretest suspicion. In that case, a number of false negatives are not detected as such, and even more true negatives drop out of the analysis. The result is an overestimation of prevalence and of sensitivity.

Based on a systematic review of sources of bias in accuracy studies, Whiting et al. reported that differential verification led to overestimation of overall accuracy and that partial verification always led to overestimation of sensitivity but not always on specificity.

### 5.3.3 Reference standard misclassification

Another artefactual modifier of prevalence is the use of an imperfect reference standard. Because a perfect reference standard is unlikely, reference standard imperfections will play a role in most accuracy studies and this effect may have been present in all examples mentioned so far. The study of Lachs, for example, resulted in a letter from Boyko, who raised the possibility that the spectrum bias Lachs and colleagues described was partly due to the imperfect reference test they used. The same issue was mentioned by Evans, some ten years later, in response to another article on spectrum variability.
Diagnostic accuracy may vary with prevalence

The issue of reference standard misclassification had already been raised in 1966 by Buck and Gart\textsuperscript{21,22}. Using a hypothetical example, they showed that in the presence of an imperfect reference standard, the reported accuracy will always be lower than the true accuracy, but sensitivity will increase towards its original value as prevalence increases, while specificity decreases. The same was also described by Brenner and Gefeller in 1997 and by Miller in 1998\textsuperscript{23,24}. In addition, imperfect reference standards also bias the reported prevalence. The following example illustrates this effect.

Let us assume that the prevalence of pulmonary embolism in hospitalized patients is 10\% and that the sensitivity of a D-dimer test is 95\% and its specificity 60\%. In a study with 1000 patients, this may lead to 455 patients with a positive index test and 545 patients with a negative index test. In a study where patients were verified by a ventilation perfusion scan, with a sensitivity of 95\% and a specificity of 90\%, the estimated prevalence of pulmonary embolism will be around 18.5\%. The estimated sensitivity will then be 68\% and specificity 60\%. Because estimates of both prevalence and accuracy are affected by reference standard misclassification, accuracy may artefactually seem to vary among subgroups of patients with different prevalences\textsuperscript{25}.

5.4 Conclusions

By their mathematical definition, sensitivity and specificity do not depend on the disease prevalence. Yet we have shown a series of examples that prevalence and diagnostic test accuracy may covary with prevalence. These examples were from both systematic reviews, which showed variation between studies, and from individual studies, which showed variation between patient subgroups. The parallel variability of prevalence and accuracy can occur through clinical mechanisms, such as patient spectrum, referral filter, or reader expectation, and artefactual mechanisms, which include distorted inclusion of participants, verification bias and reference standard misclassification. An awareness of these mechanisms and the way they can affect diagnostic accuracy is essential for a balanced translation of study results into clinical practice.

In clinical practice, Bayes’ rule is often applied, in which the likelihood ratio of a test is used to translate the pre-test probability to post-test probability. The pre-test probability is often based on the disease prevalence. The likelihood ratio of a test is a function of the test’s sensitivity and specificity and is not a fixed test property. A likelihood ratio calculated from a study with a prevalence of 5\% can therefore not be blindly used to calculate the post-test probability in a population with a prevalence of 20\%.
Chapter 5

The latter was demonstrated in the study of Van der Schouw and colleagues on diagnostic accuracy of ultrasonography for epididymitis mentioned earlier. In the patient group that had clearly epididymitis or epididymo-orchitis in the differential diagnosis, prevalence of epididymitis was 81%, post-test probability was 94% and the positive likelihood ratio was 4. In the patient group with diseases mimicking epididymitis in the differential diagnosis, the prevalence of epididymitis was 39%, the post-test probability was 91% and the positive likelihood ratio was 16. If we would have had only the results of the group with a prevalence of 81% (and a positive likelihood ratio of 4) and applied those to the group with a prevalence of 39% by using Bayes’ rule and the positive likelihood ratio of 4, we would have estimated a post-test probability of 72%. The latter differs markedly from the actual post-test probability in that group, which is 91%.

Other authors have also emphasized the relation between study characteristics and changes in diagnostic test accuracy. Unfortunately, study design features and characteristics of the population or referral filter are still badly reported. Prevalence is therefore the most apparent key feature of studies. The examples and mechanisms in this paper illustrate how prevalence can be used to signal study design deficiencies and crucial differences in patient characteristics. Hopefully a more widespread dissemination and implementation of the Standards for the Reporting of Diagnostic accuracy studies (STARD) by authors and journals will enable readers to signal study characteristics directly.

Clinicians who use the diagnostic literature in their daily practice should carefully define their clinical question first: in what population is the test going to be used, what is the clinical setting, and what is the referral filter. Studies not addressing that question and studies with obviously improper designs, such as those relying on comparisons between healthy controls and severely diseased, are unlikely to be helpful and may not be considered further.

Studies being considered further can be expected to show variability in test accuracy. By examining how accuracy varies with prevalence, an understanding of more subtle biases and sources of between-study variability in accuracy can be of help. Reasons for artefactual variability, as discussed in this paper, should be identified first. Were certain patient groups excluded from the study; was the same reference standard used in all patients? Although their effect on prevalence may vary, as seen in the examples we used in this paper, both limited challenge and verification will most often lead to higher diagnostic accuracy. The magnitude of the overestimation due to flaws in study design will vary. The severity of this overestimation will depend on clinical question and the decision that has to be made. The effect of reference standard misclassification will lead to more predictable changes in test accuracy. In case of an imperfect reference standard, sensitivity will be less underestimated and specificity will be more underestimated with greater prevalence.
Diagnostic accuracy may vary with prevalence

When the reader is ensured of the absence of bias, reasons for clinical variability, such as differences in patient groups, have to be identified. The central question here is: do the patients in this study reflect my clinical population? The combination of setting, referral filter and prevalence can be used to select those studies that are most appropriate for the clinical question. If referral filter and setting are badly reported, prevalence can serve as a guiding tool: does the prevalence of the studied population reflect my own patient population?

When both clinical and artefactual mechanisms, with possibly conflicting effects are present, the net result may be difficult to predict. Systematic reviews of diagnostic accuracy studies that take variability in prevalence into account, may throw some more light on these mechanisms and their effects.

Sensitivity and specificity are not fixed test characteristics, but test properties, that describe the behaviour of the test in a particular situation. As the setting, filter, or patient group changes, prevalence and accuracy may change. For this reason, variation in disease prevalence and test accuracy between studies can act as a flag for clinicians to detect important differences in study population or study design, affecting accuracy.
Chapter 5

References


