Journal of Clinical Epidemiology 184 (2025) 111816

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

# Variation in sensitivity and specificity of diverse diagnostic tests across health-care settings: a meta-epidemiological study

Natasja D. Vijfschagt[a], Huibert Burger[a], Marjolein Y. Berger[a], Thomas R. Fanshawe[b], Ann van den Bruel[c], Mariska M.G. Leeflang[d], Michiel R. de Boer[a], Gea A. Holtman[a,*]

[a]*Department of Primary- and Long-term Care, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands*
[b]*Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom*
[c]*Department of Public Health and Primary Care, University of Leuven, Leuven, Belgium*
[d]*Department of Clinical Epidemiology, University of Amsterdam, Amsterdam Medical Center, Amsterdam, The Netherlands*

Accepted 29 April 2025; Published online 6 May 2025

## Abstract

**Objectives:** Diagnostic test accuracy (DTA) may vary among health-care settings, which among other reasons may be due to referral from primary to secondary care. The true magnitude and direction of any difference is not certain. We analyzed the results of meta-analyses of DTA to compare sensitivity and specificity between patients in nonreferred and referred care settings.

**Study Design and Setting:** We systematically searched EBSCOhost MEDLINE for systematic reviews that included at least ten original studies of the same diagnostic test, with at least three studies each performed in nonreferred and referred care. Random-effects models, with setting as a binary covariate, were used to calculate pooled sensitivity and specificity estimates per test. Sensitivity analyses were conducted limiting the analyses to studies from countries with gatekeeping systems only.

**Results:** In total, nine systematic reviews evaluating thirteen diagnostic tests were included. For signs and symptoms (seven tests), the differences in sensitivity and specificity ranged from +0.03 to +0.30 and from −0.12 to +0.03, respectively; for biomarkers (four tests) differences in sensitivity ranged from −0.11 to +0.21 and specificity from −0.01 to −0.19. Differences in sensitivity and specificity for one questionnaire test were +0.1 and −0.07 respectively and for one imaging test were −0.22 and −0.07. Sensitivity analyses limited to countries with gatekeeping health care systems produced similar results.

**Conclusion:** Sensitivity and specificity vary in both direction and magnitude between nonreferred and referred settings, depending on the test and target condition, with no universal patterns governing performance differences. © 2025 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

*Keywords:* Sensitivity and specificity; Primary health care; Secondary care centers; Diagnostic tests (eg, ultrasound; ); Systematic review; Rapid diagnostic tests; Ultrasonography; Patient health questionnaire; Signs and symptoms

Trial registration: PROSPERO CRD42021212302.

Funding: Gea A Holtman is supported by a personal Veni grant from the Dutch Research Council (09150161810046). Thomas R Fanshawe receives funding from the National Institute for Health Research (NIHR) Community Healthcare MedTech and In Vitro Diagnostics Co-operative at Oxford Health NHS Foundation Trust (MIC-2016-018) and the NIHR Applied Research Collaboration Oxford and Thames Valley at Oxford Health NHS Foundation Trust. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

* Corresponding author. Department of Primary- and Long-term Care, University Medical Center Groningen, Hanzeplein 1, Groningen 9713 GZ, The Netherlands.

*E-mail address:* g.a.holtman@umcg.nl (G.A. Holtman).

https://doi.org/10.1016/j.jclinepi.2025.111816

0895-4356/© 2025 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

**Plain Language Summary**

Doctors use diagnostic tests to help assess the likelihood if a patient has a certain condition. However, the accuracy of these tests may vary depending on where they are used—such as in primary care (where patients first seek help) or in specialist care (after being referred by a doctor). We wanted to find out how much test accuracy changes between these settings. To do this, we analyzed previous studies that reviewed the accuracy of different diagnostic tests. We compared how well these tests worked in patients who had not yet been referred to a specialist vs those who had. Our analysis included results from thirteen different diagnostic tests, covering symptoms, biomarkers (such as blood tests), a questionnaire, and an imaging test. We found that test accuracy varied depending on the type of test and the condition being diagnosed. Some tests had higher sensitivity (correctly identifying patients with the disease) or specificity (correctly identifying healthy individuals) in primary care, while in specialist care, the same test could perform better, worse, or similarly. There was no clear pattern that applied to all tests. This suggests that researchers should consider how test accuracy may differ across health-care settings when conducting and interpreting diagnostic test accuracy studies.

## 1. Introduction

Diagnostic testing in symptomatic patients is central to the identification of numerous health conditions. An essential characteristic of a diagnostic test's clinical utility is its accuracy, for which sensitivity and specificity serve as vital measures during early evaluation [1,2]. Sensitivity gauges a test's ability to identify true positives, while specificity measures its capacity to identify true negatives. However, both measures are known to vary across populations and health-care settings [3–7]. Setting is considered a proxy for patient selection, clinician expertise, and test and reference standard characteristics, which differ by specialism, country, and health-care system. Health-care settings range from primary care, where nonreferred patients seek initial assessment, to specialized care, where a selected group of referred patients suspected of having a particular disease undergo further evaluation. Tests evaluated after referral have in general been reported to have a higher sensitivity and lower specificity when compared with the same test evaluated before referral [1]. In a referred population, the development of symptoms of the target condition is expected to be more pronounced with a higher sensitivity as a result. In addition, the distribution of disease changes after referral, with more similarities between those with and without the disease increasing the likelihood of false positives and therefore lowering specificity. Yet, sensitivity and specificity may also vary due to other factors, for example, differences in operator expertise in each health-care setting. Although researchers have explored variations between settings for specific tests and target conditions, mainly based on theoretical reasoning and simulation studies, our understanding of variability across different tests and conditions remains limited [4,6,8]. Greater insight into the direction and magnitude of differences and into the observed patterns of variation could help when translating sensitivity and specificity results across different settings and aids in research conduct and interpretation.

We aimed to study the variation in sensitivity and specificity between settings by reanalyzing data from earlier meta-analyses. This included studies of different types of diagnostic tests performed in nonreferred and referred care.

## 2. Material and methods

### 2.1. Study design

This metaepidemiological study included systematic reviews of diagnostic test accuracy (DTA) to meta-analyze estimates of the sensitivity and specificity of tests in nonreferred and referred care settings as reported in the primary studies (International Prospective Register of Systematic Reviews: CCRD42021212302) [9]. The following definitions were used.

- Nonreferred: patients presenting with symptoms for the first time without prior diagnostic testing, categorized as those presenting to 1) clinicians in community care facilities (community care), 2) general practitioners (GPs) or family doctors (primary care), or 3) general specialists (eg, general internist), including self-referral to emergency departments in nongatekeeping countries (outpatient care).
- Referred: patients referred to specialist care based on clinical suspicion by, for example, a GP, family doctor, or general internist (specialist care).

Gatekeeping systems provide a clearer distinction between nonreferred and referred care, therefore we categorized health care into systems with and without gatekeeping (Supplementary File A).

### 2.2. Search strategy

A literature search was performed in EBSCOhost MEDLINE to identify systematic reviews of DTA published

**What is new?**

**Key findings**

- Sensitivity and specificity vary in both direction and magnitude between settings.

**What this adds to what is known?**

- Differences do not follow a specific pattern; it varies across tests and conditions.
- Differences in sensitivity were larger than those in specificity.

**What is the implication and what should change now?**

- Consider setting in diagnostic accuracy interpretation and research design.

between 2010 and September 2020 (Supplementary File B). The Cochrane library was checked for reviews and we checked references of Horizon Scan Reports on new point of care diagnostic tests relevant to primary care settings [10].

### 2.3. Systematic review and primary study selection

We included systematic reviews with ≥10 primary studies of diagnostic index tests suitable for use in nonreferred care (eg, noninvasive, cheap, easily accessible), ≥3 primary studies in both nonreferred and referred care, and written in English [11]. Two reviewers (N.D.V. with H.B., M.d.B., or G.A.H.) independently screened the titles and abstracts for eligibility and reviewed the full texts of selected systematic reviews. Disagreements were resolved through discussion, and if needed, with a third reviewer. Primary studies incorporated in the systematic reviews were included only if we could classify the setting and calculate sensitivity and specificity. We excluded primary studies performed in screening settings or with case-control designs (except for nested studies in which both cases and controls were sampled from the target population) [12].

### 2.4. Systematic review and primary study data extraction

Two reviewers (G.A.H., N.D.V.) extracted the data, checked whether the search strategy of the systematic reviews was sufficiently broad, using ROBIS tool item 2.1, and whether the systematic reviews evaluated factors influencing sensitivity or specificity [13]. For each primary study, two reviewers independently selected and extracted data (N.D.V. with G.A.H., M.d.B., H.B., M.B., A.v.B., M.L., or T.R.F.), with disagreements resolved by a third reviewer. Data were extracted using a standardized form (Supplementary File C). Risk of bias was assessed with the Quality Assessment of Diagnostic Accuracy Studies-2, using either outcomes from the systematic review or ratings by ourselves (N.D.V., T.R.F.).

### 2.5. Data analysis

We aimed to determine variation in sensitivity and specificity for each index test between studies performed in nonreferred and referred care settings. First, to assess heterogeneity visually, we presented the sensitivity and specificity for each test in forest plots. Second, to calculate pooled estimates of setting-specific sensitivity and specificity, we used bivariate random-effects models with setting as a binary covariate [14]. Third, the pooled estimate of prevalence for each target condition per index test in each setting was calculated using Clopper−Pearson confidence limits for binomial proportions [15]. The pooled prevalence, sensitivity, and specificity of each test in each setting, together with the risk of bias outcomes, were presented graphically. We decided a priori not to pool setting effects across systematic reviews, as the direction of biases could be different in each meta-analysis [16]. In cases with ≥3 studies per setting, we performed sensitivity analyses for each index test in gatekeeping countries only; sensitivity analyses were also performed to evaluate the influence of outbreak settings (eg, community care, schools, and festivals) by excluding these settings. The statistical software Stata/SE (version 18.0; Stata Corp, College station, TX) was used with the MetaDTA package for calculations of pooled sensitivity and specificity, and the Metaprop package was used to calculate the pooled prevalence [14,15,17]. RStudio Team (2020; RStudio, Integrated Development for R, RStudio, PBC, Boston, MA) was used for data visualization [18].

## 3. Results

Figure 1 shows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart of included systematic reviews, and Supplementary File D provides an overview of the included and excluded primary studies extracted from the reviews [19]. We included nine systematic reviews including 156 original studies investigating thirteen different types of index tests categorized into: sign or symptom (seven tests), biomarker (four tests), questionnaire (one test), and imaging (one test) (Fig 1, Supplementary File D). Table 1 provides an overview of the included reviews and is expanded upon in Supplementary File E. Authors from the included systematic reviews performed appropriate literature searches. Of the nine reviews, four evaluated the influence of setting on DTA, six evaluated risk of bias items, and all evaluated

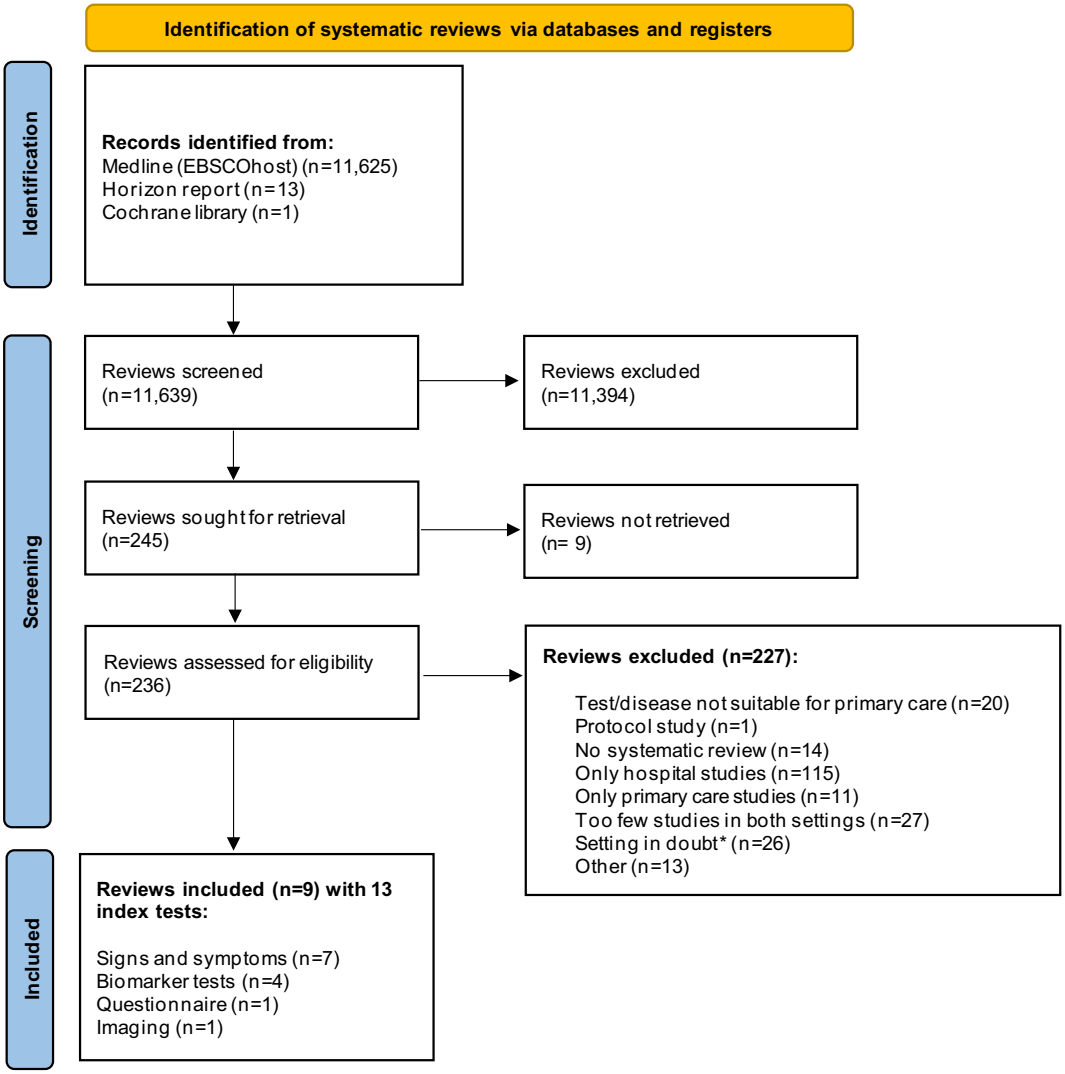


**Figure 1.** Preferred Reporting Items for Systematic Reviews and Meta-Analyses Flowchart of included systematic reviews. *Setting of interest not reported or focused explicitly on nonreferred settings, but did not include primary care studies (eg, studies conducted only in the general population or municipalities).

characteristics (eg, age, disease severity or prevalence). We agreed with the setting categories in most instances, but did categorize some studies differently (Supplementary File F).

Overall, the pooled prevalence of target conditions was higher for tests in referred settings (median 32%, range 4%–49%) compared to nonreferred settings (median 19%, range 4%–44%), although there were exceptions (Fig 2). These included the following tests: abdominal pain, weight loss, and change in bowel habit for colorectal cancer; enzyme immune assay (EIA) and optical immune assay (OIA) for Group A β-hemolytic streptococcus; and rapid influenza diagnostic test (RIDT) for influenza A infection. Forest plots for each index test showed different patterns of heterogeneity in sensitivity and specificity (Supplementary File G). Sensitivity was typically more heterogeneous with lower prevalence and smaller samples of patients with the disease (eg, all signs and symptoms, OIA/EIA, RIDT, mini-mental state examination [MMSE], ultrasound), whereas specificity was typically high, more homogenous, and more precisely estimated (eg, anosmia, visual inspection, weight loss, all biomarkers, and MMSE). Most tests had similar risks of bias in nonreferred and referred settings. The main differences were in the QUADAS domains ''reference standard'' and/or ''flow and timing'' for visual inspection, and fecal calprotectin 50 μg/g and 100 μg/g; here, more studies performed in the nonreferred setting had a higher risk of bias (Fig 2).

### 3.1. Index test by type

#### 3.1.1. Signs and symptoms

Seven index tests (60 studies; $n = 53{,}111$) were included for four target conditions: paroxysmal cough (10 studies;

**Table 1.** Descriptive characteristics of included systematic reviews

| Author, year | Population | Index test(s) | Reference standard(s) | Target condition | Settings (number of studies) |
|---|---|---|---|---|---|
| Signs and symptoms | | | | | |
| Moore A, 2017 [19] | Patients of any age attending any health-care setting with clinical characteristics that might be associated with pertussis | Paroxysmal cough<br>Posttussive vomiting | Laboratory investigation (culture, serology, RT-PCR, and/or DFA test) | Bordetella pertussis infection | Nonreferred primary care ($n$ = 2)<br>Nonreferred outpatient care ($n$ = 2)<br>Referred specialist care ($n$ = 9) |
| Jellema P, 2010 [20] | Adults with nonacute (>2 wks) lower abdominal symptoms | Abdominal pain<br>Change in bowel habit<br>Weight loss | Colonoscopy, barium, enema, or clinical follow-up | Colorectal cancer | Nonreferred primary care ($n$ = 7)<br>Referred specialist care ($n$ = 9) |
| Struyf T, 2021 [21] | Patients with clinically suspected COVID-19 | Anosmia | RT-PCR, clinical expertise, and imaging | COVID-19 | Nonreferred community ($n$ = 4)<br>Nonreferred primary care ($n$ = 3)<br>Nonreferred outpatient care ($n$ = 1)<br>Referred specialist care ($n$ = 3) |
| Dinnes J, 2018 [22] | Adults with lesions suspicious for melanoma | Visual inspection, in-person | Histopathological diagnosis or clinical follow-up (to confirm benignity) | Any form of invasive cutaneous melanoma or atypical intraepidermal melanocytic variants | Nonreferred primary care ($n$ = 3)<br>Referred specialist care ($n$ = 19) |
| Biomarkers | | | | | |
| Merckx J, 2019 [23] | Children and adults with clinically suspected influenza during periods of influenza activity | Rapid influenza diagnostic tests | RT-PCR | Influenza A infection | Nonreferred community ($n$ = 5)<br>Nonreferred outpatient care ($n$ = 9)<br>Referred specialist care ($n$ = 3) |
| An YK, 2019 [24] | Patients (adults or children) presenting with lower gastrointestinal symptoms | FCal cutoff 50 μg/g<br>FCal cutoff 100 μg/g | Colonoscopy or cross-sectional imaging | Organic gastrointestinal disorder | Nonreferred primary care ($n$ = 5)<br>Referred specialist care ($n$ = 9) |
| Cohen JF, 2016 [25] | Children with acute pharyngitis | Enzyme immunoassay and optical immunoassay | Throat culture on a blood agar plate | Group A β-hemolytic streptococcus | Nonreferred primary care ($n$ = 9)<br>Nonreferred outpatient care ($n$ = 17)<br>Referred specialist care ($n$ = 6) |
| Questionnaire | | | | | |
| Tsoi KKF, 2015 [26] | Participants studied for the detection of dementia associated with Alzheimer disease, vascular dementia, or Parkinson disease in any clinical or community setting | MMSE | Standard diagnostic criteria for defining dementia | Dementia | Nonreferred community ($n$ = 1)<br>Nonreferred primary care ($n$ = 3)<br>Nonreferred outpatient care ($n$ = 2)<br>Referred specialist care ($n$ = 11) |

(*Continued*)

**Table 1.** Continued

| Author, year | Population | Index test(s) | Reference standard(s) | Target condition | Settings (number of studies) |
|---|---|---|---|---|---|
| Imaging | | | | | |
| Ebell MH, 2016 [27] | Adults and children with clinically suspected sinusitis or acute respiratory tract infection | Ultrasound (A or B mode) | Radiography, ultrasound, computed tomography, or MRI for acute rhinosinusitis, and antral puncture revealing purulent fluid or fluid yielding a positive culture for acute bacterial rhinosinusitis | Acute rhinosinusitis | Nonreferred community ($n$ = 1)<br>Nonreferred primary care ($n$ = 2)<br>Nonreferred outpatient care ($n$ = 3)<br>Referred specialist care ($n$ = 5) |

COVID-19, coronavirus disease 2019; DFA, direct fluorescent antibody; FCAL, fecal calprotectin; MMSE, mini-mental state examination; MRI, magnetic resonance imaging; RT-PCR, Reverse transcription polymerase chain reaction.

$n$ = 4013 nonreferred, $n$ = 601 referred) and posttussive vomiting (seven studies; $n$ = 827 nonreferred, $n$ = 419 referred) for *Bordetella pertussis*; abdominal pain (fifteen studies; $n$ = 1737 nonreferred, $n$ = 16,297 referred), change in bowel habit (fourteen studies; $n$ = 1405 nonreferred, $n$ = 14,304 referred), and weight loss (nine studies; $n$ = 1734 nonreferred, $n$ = 6280 referred) for colorectal cancer; anosmia (eleven studies; $n$ = 8310 nonreferred, $n$ = 1310 referred) for coronavirus disease 2019 (COVID-19); and [4] visual inspection (22 studies; $n$ = 1373 nonreferred, $n$ = 18,664 referred) for malignant melanoma. All pooled sensitivities were higher in nonreferred care, with a median difference of 0.15 percentage points, ranging from 0.03 (weight loss for colorectal cancer) to 0.30 (paroxysmal cough for *B. pertussis*) (Fig 2, Supplementary File H). Pooled specificity was lower in nonreferred care for five of seven index tests, differing from 0.01 (abdominal pain for colorectal cancer) to 0.31 (paroxysmal cough for *B. pertussis*). All confidence intervals overlapped, except for the pooled sensitivity and specificity of anosmia in COVID-19.

### 3.1.2. Biomarkers

In nonreferred care, fecal calprotectin testing for organic gastrointestinal disease at cutoffs of 50 µg/g (12 studies; $n$ = 2125 nonreferred, $n$ = 1324 referred) and 100 µg/g (nine studies; $n$ = 2124 nonreferred, $n$ = 965 referred) had lower pooled sensitivities at 0.07 and 0.11, respectively. Pooled specificity with the 50 µg/g cutoff was 0.19 percentage points lower in nonreferred care, but it was the same for the 100 µg/g cutoff (Fig 2, Supplementary File H).

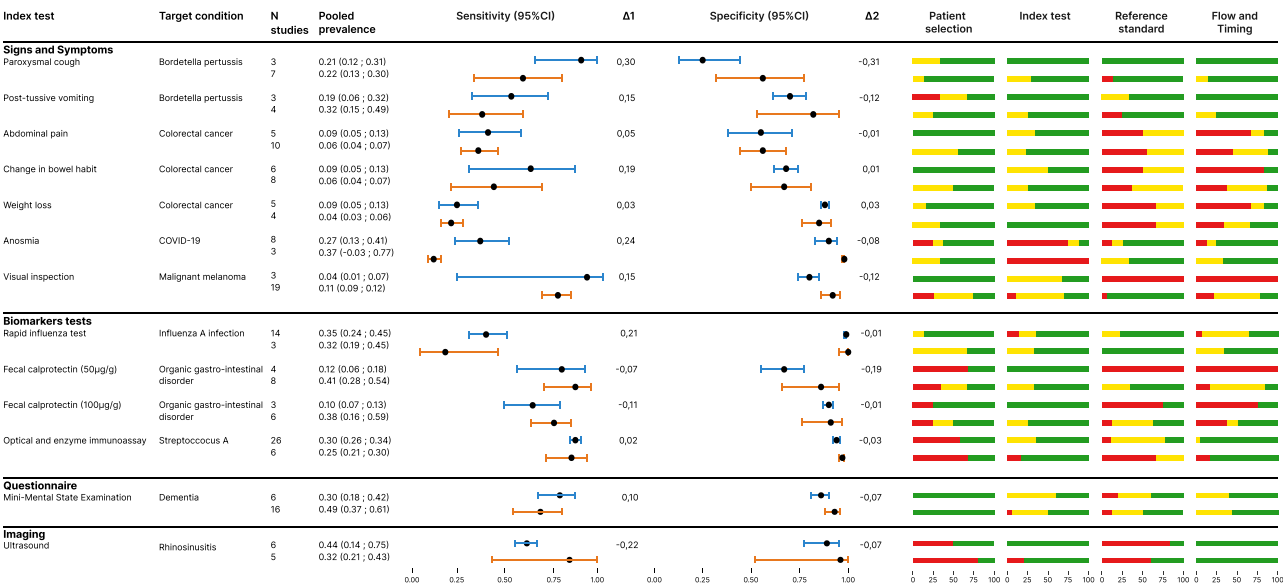


**Figure 2.** Variation in pooled sensitivity and specificity for all tests across both settings, together with the risk of bias assessment. The blue line represents the nonreferred population, while the orange line represents the referred population. Δ1 difference in sensitivity between nonreferred and referred settings. Δ2 difference in specificity between nonreferred and referred settings.

EIA and OIA (32 studies; $n$ = 16,440 nonreferred, $n$ = 2761 referred) had similar pooled sensitivities and specificities in both settings. By contrast, RIDT (17 studies; $n$ = 7065 nonreferred, $n$ = 2467 referred) showed a pooled sensitivity 0.21 percentage points higher, and a pooled specificity 0.01 percentage points lower, in nonreferred care (Fig 2, Supplementary File H).

#### 3.1.3. Questionnaire

The MMSE for dementia (22 studies; $n$ = 1215 nonreferred, $n$ = 3395 referred) had a pooled sensitivity 0.10 percentage points higher, and a pooled specificity 0.07 percentage points lower, in nonreferred care (Fig 2, Supplementary File A).

#### 3.1.4. Imaging

Ultrasound for rhinosinusitis (11 studies; $n$ = 982 nonreferred, $n$ = 554 referred) had pooled sensitivity and specificity values 0.22 and 0.07 percentage points lower, respectively, in nonreferred care (Fig 2, Supplementary File H).

### 3.2. Sensitivity analysis

In sensitivity analyses for RIDT and EIA/OIA, results were unchanged after excluding studies performed during outbreaks (Fig 2, Supplementary File H and I).

For studies done in gatekeeping countries, the sensitivity analyses were performed for abdominal pain, change in bowel habit, visual inspection, OIA/EIA, and fecal calprotectin 50 μg/g. They generally showed similar directions and magnitudes of sensitivity and specificity by setting (Supplementary File J). However, magnitude differed for change in bowel habit for colorectal cancer (smaller difference of 0.09 percentage points in sensitivity) and fecal calprotectin 50 μg/g (smaller difference of 0.07 percentage points in specificity).

## 4. Discussion

### 4.1. Summary

Comparing nonreferred and referred settings, the magnitude and direction of sensitivity and specificity for different tests and conditions varied. Of note, no test showed the frequently described pattern of lower sensitivity and higher specificity in nonreferred care, while sensitivity varied more than specificity. Compared to the referred population, in the nonreferred population, 54% of the tests had higher sensitivity with lower specificity, 15% had higher sensitivity with higher specificity, and 31% had lower sensitivity with lower specificity. Sensitivity analysis produced similar results for gatekeeping systems that better differentiate nonreferred and referred settings.

### 4.2. Explanation of findings

Absence of the frequently described pattern of lower sensitivity and higher specificity in nonreferred care contradicts the results of previous metaepidemiological studies [3,28]. When combining meta-analyses of diagnostic tests, an overall higher prevalence has been associated with higher sensitivity and lower specificity [28], albeit with effects in the opposite direction for several index tests [3,28]. Most of these studies included in previous metaepidemiological studies were performed in referred care (hospital) settings.

The differences with our results could be explained by several mechanisms described in literature. First, the higher sensitivity and lower specificity in nonreferred care for all signs and symptoms, in some instances, could reflect ''over reading'' (overdiagnosis from overinterpretation) by clinicians when compared with referred care [4]. This might particularly be the case for meta-analyses of conditions such as colorectal cancer and malignant melanoma, which if incorrectly ruled out in primary care may have serious consequences for the patient. The GP may therefore implement a low threshold for test positivity. Also, in situations that require subjective interpretation of index test results, there could be an effect of prevalence on implicit thresholds of how clinicians interpret test results [29]. Second, referred care may have a higher proportion of equivocal cases that are more difficult to diagnose (eg, malignant melanoma) resulting in lower specificity [3,22]. Third, the lower sensitivities of fecal calprotectin for organic gastrointestinal disease and of ultrasound for acute rhinosinusitis in nonreferred care could reflect the lower severity of disease and thereby less developed symptoms, and the relative inexperience of examiners. Finally, the similar sensitivity and specificity of EIA and OIA for Group A β-hemolytic streptococcus in each setting could reflect similarities in prevalence and test execution. Unfortunately, we could not quantify these explanatory factors due to poor reporting, the limited number of studies, and the relationships between factors.

In contrast to our results, earlier studies showed more variation in specificity than in sensitivity. A meta-analysis evaluating several rules for pulmonary embolism showed that all strategies had a sensitivity exceeding 90% in all settings, but that specificity varied considerably [30]. Furthermore, a metaepidemiological study of DTA more often showed associations between prevalence and specificity when compared to sensitivity [5]. There are several possible explanations for the discrepancy between our results and those of others. First, the goal of the test is key when deciding on whether a high sensitivity (eg, triage tests to avoid missing severe disease) or a high specificity (eg, tests to decrease invasive follow-up or treatment) is most important. A high sensitivity or specificity provides less room for variation (ie, the ceiling effect). In our study, eight

of thirteen tests had a specificity >80% in both settings, but no test had a sensitivity >80% in both settings. In contrast, the metaepidemiological study that showed an association between prevalence and specificity found that eight of 23 tests had a sensitivity >80% [5]. A more recently published metaepidemiological study of 553 meta-analyses found associations between prevalence and both sensitivity and specificity, explaining that earlier studies were too small and underpowered to find an association with sensitivity [28].

Remarkably, although we expected prevalence to be higher in referred care, this was not the case for some target conditions, where prevalence was actually higher in nonreferred care. For example, this occurred for target conditions that are almost exclusively managed in primary care such as Group A β-hemolytic streptococcus and influenza A infection. Therefore, the relationship between prevalence and setting is complex and is likely to differ for different target conditions.

### 4.3. Future directions

The differences in direction and magnitude of sensitivity and specificity between settings should be considered by clinicians using tests and by researchers designing DTA studies or performing DTA reviews. Clinicians who wish to apply the Bayesian properties of diagnostic tests and translate pretest probabilities to posttest probabilities need accurate estimates of pretest probability, sensitivity, and specificity in their target population. An important finding of our study is that the setting is not always clearly described or defined in systematic reviews or in primary studies. Researchers performing DTA studies should clearly state the setting, as stipulated in the Standards for the Reporting of Diagnostic Accuracy Studies checklist [31], while those performing DTA reviews should conduct subgroup analysis by setting. Using individual patient data (IPD) instead of aggregate data improves the interpretation of DTA results across different settings and enhances information quality by facilitating comparison of participant and test characteristics across studies. Standardizing datasets improves statistical analyses, and incorporating setting-specific or patient-level factors reduces inconsistency and heterogeneity.

### 4.4. Study limitations

Our study has several limitations. First, we identified only nine eligible systematic reviews of thirteen different tests because only these systematic reviews included at least three studies in both settings. It was deemed not efficient to search for primary studies and we therefore decided to confine ourselves to systematic reviews. As a result, we may have missed important studies. Second, we lacked studies with data about patients initially tested in nonreferred care and who underwent repeat testing in referred care. This approach would offer more direct insights into setting specific DTA for different tests and target conditions. Third, it appeared difficult to define setting, which differs by specialism, health-care system, and country. Therefore, our extensive assessment of these criteria for each included study, including the sensitivity analysis in countries with gatekeeping systems, could inform future metaepidemiological studies. Fourth, due to the small number of studies in some metaregressions and the presence of small sample sizes in certain cases, these results may be susceptible to small sample bias. This could give rise to publication bias, outcome reporting bias, or clinical heterogeneity across studies, all of which may distort pooled estimates when based on limited data. Last, metaepidemiological studies cannot typically provide insight into the causes of variability due to the poor descriptions of other contributing factors in meta-analyses. Specifically, inadequate information about, for example, presenting symptoms, index test or reference standard on IPD hindered comprehensive examination of variation in test characteristics, which is a limitation inherent to utilizing aggregated data.

## 5. Conclusion

Sensitivity and specificity vary considerably in direction and magnitude between nonreferred and referred settings. Notably, performance is not always in the same direction and varies with the test and target condition. It can therefore be concluded that there are no definitive or universally applicable patterns regarding the performance differences between nonreferred and referred care.

## CRediT authorship contribution statement

**Natasja D. Vijfschagt:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Huibert Burger:** Writing – review & editing, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Marjolein Y. Berger:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Thomas R. Fanshawe:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Ann van den Bruel:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Mariska M.G. Leeflang:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Michiel R. de Boer:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Gea A. Holtman:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

There are no competing interests for any author.

## Acknowledgments

The authors thank Lisa Powaga (Department of Public Health and Primary Care, University of Leuven, Belgium) for assisting with the data extractions. Special gratitude to Thijmen Kupers (data scientist, Department of Primary- and Long-term Care, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands) for constructing the results in R software. Finally, we thank Dr Rob Sykes (www.doctored.org.uk) for manuscript editing services.

## Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2025.111816.

## Data availability

No data available. We did not collect and process any primary data. As part of our systematic review and meta-analysis, we only used results that were reported in published literature

## References

[1] Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. J Clin Epidemiol 2002; 55(12):1201−6. https://doi.org/10.1016/S0895-4356(02)00528-0.

[2] Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clinica Chim Acta 2014;427:49−57. https://doi.org/10.1016/j.cca.2013.09.018.

[3] Leeflang MMG, Bossuyt PMM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62(1):5−12. https://doi.org/10.1016/j.jclinepi.2008.04.007.

[4] Sackett DL, Haynes RB. Evidence base of clinical diagnosis: the architecture of diagnostic research. Br Med J 2002;324(7336):539−41. https://doi.org/10.1136/bmj.324.7336.539.

[5] Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt PMM. Variation of a test's sensitivity and specificity with disease prevalence. Can Med Assoc J 2013;185(11):E537−44. https://doi.org/10.1503/cmaj.121286.

[6] Gambino B. Test performance variation between settings and populations. J Gambl Stud 2018;34(4):1085−108. https://doi.org/10.1007/s10899-017-9728-9.

[7] Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med 1997; 16(9):981−91. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N.

[8] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New Engl J Med 1978; 299(17):926−30. https://doi.org/10.1056/NEJM197810262991705.

[9] Puljak L, Makaric ZL, Buljan I, Pieper D. What is a meta-epidemiological study? Analysis of published literature indicated heterogeneous study designs and definitions. J Comp Eff Res 2020; 9(7):497−508. https://doi.org/10.2217/cer-2019-0201.

[10] Verbakel JY, Turner PJ, Thompson MJ, Plüddemann A, Price CP, Shinkins B, et al. Common evidence gaps in point-of-care diagnostic test evaluation: a review of horizon scan reports. BMJ Open 2017; 7(9):e015760. https://doi.org/10.1136/bmjopen-2016-015760.

[11] Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration; 2010. Available at: http://srdta.cochrane.org/. Accessed December 2, 2022.

[12] Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol 2003; 10(6):670−2. https://doi.org/10.1016/S1076-6332(03)80087-9.

[13] Whiting P, Savović J, Higgins JPT, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225−34. https://doi.org/10.1016/j.jclinepi.2015.06.005.

[14] Nyaga VN, Arbyn M. Metadta: a Stata command for meta-analysis and meta-regression of diagnostic test accuracy data − a tutorial. Arch Public Health 2022;80(1):1−15. https://doi.org/10.1186/s13690-021-00747-5.

[15] Nyaga VN, Arbyn M, Aerts M. Metaprop: a Stata command to perform meta-analysis of binomial data. Arch Public Health 2014; 72(1):1−10. https://doi.org/10.1186/2049-3258-72-39.

[16] Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, Van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174(4):469−76. https://doi.org/10.1503/cmaj.050090.

[17] StataCorp. Stata: Release 17. Statistical Software. StataCorp LLC: College Station, TX; 2021.

[18] RStudio Team. RStudio. Boston, MA: Integrated Development for R. RStudio, PBC; 2020.

[19] Moore A, Ashdown HF, Shinkins B, Roberts NW, Grant CC, Lasserson DS, et al. Clinical characteristics of pertussis-associated cough in adults and children: a diagnostic systematic review and meta-analysis. Chest 2017;152(2):353−67. https://doi.org/10.1016/j.chest.2017.04.186.

[20] Jellema P, Van Der Windt DAWM, Bruinvels DJ, Mallen CD, Van Weyenberg SJB, Mulder CJ, et al. Value of symptoms and additional diagnostic tests for colorectal cancer in primary care: systematic review and meta-analysis. BMJ (Online) 2010;340(7750):795. https://doi.org/10.1136/bmj.c1269.

[21] Struyf T, Deeks JJ, Dinnes J, Takwoingi Y, Davenport C, Leeflang MM, et al. Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19. Cochrane Database Syst Rev 2021;2021(3):2−11. https://doi.org/10.1002/14651858.CD013665.pub2.

[22] Dinnes J, Deeks JJ, Grainge MJ, Chuchu N, Ferrante di Ruffano L, Matin RN, et al. Visual inspection for diagnosing cutaneous melanoma in adults. Cochrane Database Syst Rev 2018;2018(12): CD013194. https://doi.org/10.1002/14651858.CD013194.

[23] Merckx J, Wali R, Schiller I, Caya C, Gore GC, Chartrand C, et al. Diagnostic accuracy of novel and traditional rapid tests for influenza infection compared with reverse transcriptase polymerase chain reaction. Ann Intern Med 2017;167(6):395−409. https://doi.org/10.7326/M17-0848.

[24] An YK, Prince D, Gardiner F, Neeman T, Linedale EC, Andrews JM, et al. Faecal calprotectin testing for identifying patients with organic gastrointestinal disease: systematic review and meta-analysis. Med J Aust 2019;211(10):461−7. https://doi.org/10.5694/mja2.50384.

[25] Cohen JF, Bertille N, Cohen R, Chalumeau M. Rapid antigen detection test for group A streptococcus in children with pharyngitis. Cochrane Database Syst Rev 2016;2016(7):CD010502. https://doi.org/10.1002/14651858.CD010502.pub2.

[26] Tsoi KKF, Chan JYC, Hirai HW, Wong SYS, Kwok TCY. Cognitive tests to detect dementia a systematic review and meta-analysis.

JAMA Intern Med 2015;175(9):1450−8. https://doi.org/10.1001/jamainternmed.2015.2152.

[27] Ebell MH, McKay B, Guilbault R, Ermias Y. Diagnosis of acute rhinosinusitis in primary care: a systematic review of test accuracy. Br J Gen Pract 2016;66(650):e612−32. https://doi.org/10.3399/bjgp16X686581.

[28] Murad MH, Lin L, Chu H, Hasan B, Alsibai RA, Abbas AS, et al. The association of sensitivity and specificity with disease prevalence: analysis of 6909 studies of diagnostic test accuracy. CMAJ 2023; 195(27):E925−31. https://doi.org/10.1503/cmaj.221802.

[29] Willis BH. Empirical evidence that disease prevalence may affect the performance of diagnostic tests with an implicit threshold: a cross-sectional study. BMJ Open 2012;2(1):e000746. https://doi.org/10.1136/bmjopen-2011-000746.

[30] Geersing GJ, Takada T, Klok FA, Büller HR, Courtney DM, Freund Y, et al. Ruling out pulmonary embolism across different healthcare settings: a systematic review and individual patient data meta-analysis. PLoS Med 2022;19(1):1−17. https://doi.org/10.1371/journal.pmed.1003905.

[31] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11): e012799. https://doi.org/10.1136/bmjopen-2016-012799.