Increasing value in diagnostic research: Publication and reporting of test accuracy studies

Korevaar, D.A.

Increasing value in diagnostic research: Publication and reporting of test accuracy studies

Daniël A. Korevaar

Increasing value in diagnostic research: Publication and reporting of test accuracy studies
Daniël A. Korevaar

Increasing value in diagnostic research:
Publication and reporting of test accuracy studies

Daniël A. Korevaar

This thesis was prepared at the Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Printing of this thesis was financially supported by Opvoedkundige Praktijk Het Leerhuis Abcoude, and Stichting tot Bevordering van de Klinische Epidemiologie Amsterdam.

Layout:	Daniël A. Korevaar
Cover illustration:	Jan Korevaar
Cover design:	Fred Thoolen
Printed by:	Ipskamp Drukkers, Enschede, the Netherlands
ISBN:	978-94-028-0299-3

Increasing value in diagnostic research:

Publication and reporting of test accuracy studies

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op vrijdag 18 november 2016, te 14.00 uur

door Daniël Arnoldus Korevaar

geboren te Amsterdam

Promotiecommissie

Promotor:	prof. dr. P.M.M. Bossuyt	Universiteit van Amsterdam
Copromotor:	dr. L. Hooft	Universiteit Utrecht
Overige leden:	prof. dr. L.M. Bouter	Vrije Universiteit Amsterdam
	prof. dr. K.G.M. Moons	Universiteit Utrecht
	prof. dr. P.J. Sterk	Universiteit van Amsterdam
	prof. dr. J. Stoker	Universiteit van Amsterdam
	prof. dr. ir. H.C.W. de Vet	Vrije Universiteit Amsterdam
	dr. G. ter Riet	Universiteit van Amsterdam

Faculteit der Geneeskunde

Contents

Part A: Publication of full study reports

Part B: Prospective registration of study protocols

Part C: Informative reporting of study reports

Part D: Diagnostic tests in respiratory medicine

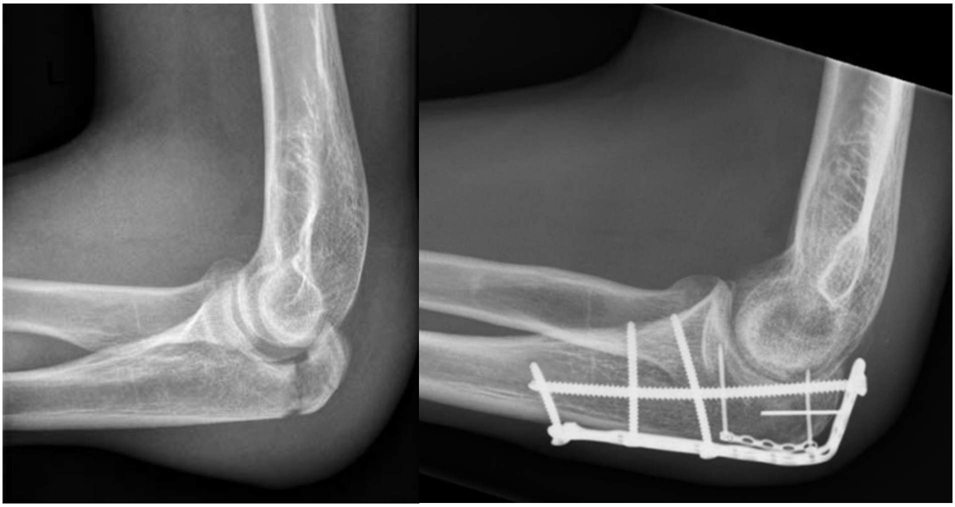
Addendum

General introduction

Case report

After a crash in a cycling race in Woensdrecht in the Netherlands during the second year of my PhD, I was brought to the emergency department of the nearest hospital. I had a painful and swollen left elbow, a painful and bruised left hip, and a painful lower back. The local clinician performed physical examination. Based on this, he estimated the probability of an elbow fracture to be high, the probability of a hip fracture to be low, and the probability of a vertebral fracture to be nihil. X-rays were taken; they demonstrated an olecranon fracture (Figure 1), and no abnormalities in the hip region. Imaging tests of the vertebral column were considered unnecessary. The elbow got surgically repaired (Figure 1), and I was sent home.

Figure 1. X-ray of the left elbow: pre-surgery (left) and post-surgery (right).

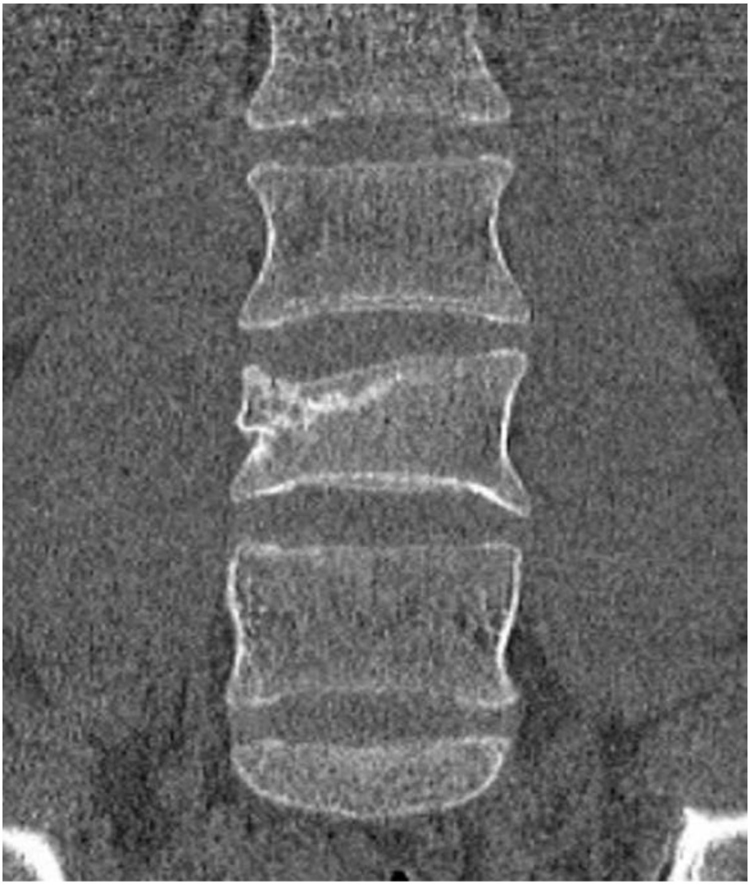


Two painful weeks went by. Based on quite a few previous experiences with crashes at high speed in cycling races, I knew that it generally took me a week to fully recover from a lower back contusion. This time, I experienced no improvement at all. With the greatest difficulty, I managed to get out of bed and ready for work every morning, after which I shuffled from my house to the metro station, and from the metro station to my office in the Academic Medical Center in Amsterdam, where I stayed seated most of the day.

This abnormal course started to worry me, so I went back to the hospital. A CT (computed tomography) scan was performed, which demonstrated a fracture of the fourth lumbar vertebra (Figure 2). Apparently, physical examination at the emergency department had produced a false negative result, a common phenomenon in clinical practice.[1]

The clinician at the emergency department was probably aware of the fact that physical examination is an imperfect test for detecting vertebral fractures. What makes his job extremely difficult, however, is that the results of scientific studies are often ambiguous. One study, for example, concluded that "clinical examination is a sensitive screening method for significant thoracolumbar spine injury",[2] whereas another study found that "clinical examination is insufficient to rule out thoracolumbar spine injuries".[3] Such conflicting conclusions can lead to clinical uncertainty. Every clinician needs to decide which studies are most relevant and trustworthy for answering his or her clinical question. But what if not all studies have been published, or if the study reports do not contain sufficient details to assess relevance and trustworthiness?

Figure 2. Fracture of the fourth lumbar vertebra.



Background

Medicine is "the science or practice of the diagnosis, treatment, and prevention of disease".[4] As this definition highlights, diagnosis of disease forms an essential part of daily clinical practice. Clinicians rarely initiate therapeutic interventions before performing at least one, but usually multiple, diagnostic tests. A test can be any method or tool for obtaining insights into a patient's health status. Tests aid in the detection of diseases and other medical conditions.

In the fourth century BC, the Greek physician Hippocrates wrote: "In acute diseases the physician must conduct his inquiries in the following way. First he must examine the face of the patient, and see whether it is like the faces of healthy people, and especially whether it is like its usual self. Such likeness will be the best sign, and the greatest unlikeness will be the most dangerous sign".[5] Fortunately, nowadays we have access to tests that are able to detect conditions with much more accuracy.

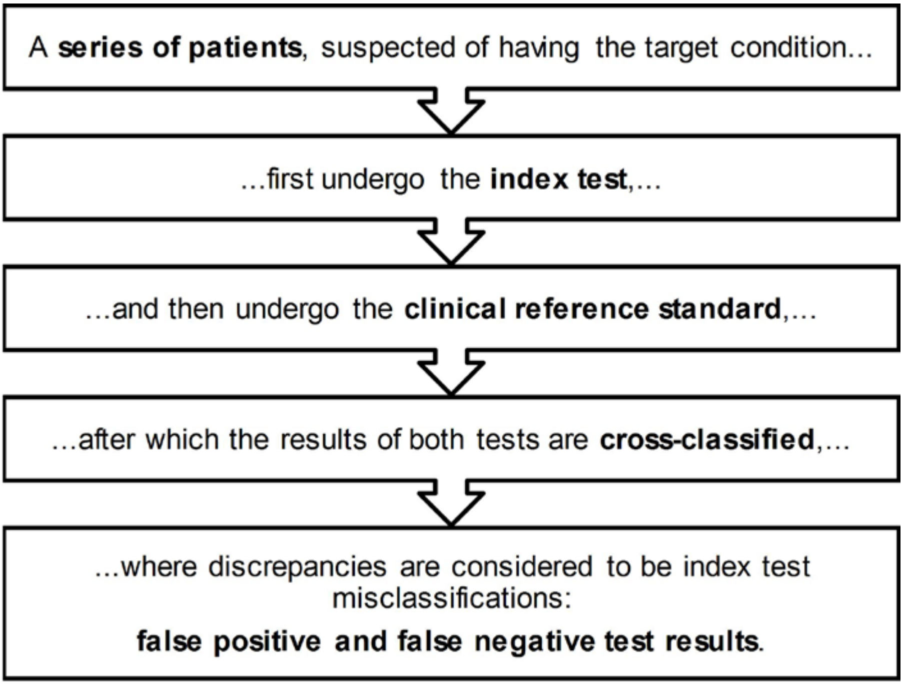
However, as the case report above illustrates, most tests are still not perfectly accurate: they usually produce a proportion of false positive and false negative results.[6] Such results may lead to delays in the administration of optimal treatments, to adverse events from the application of unnecessary interventions, and to inflated healthcare costs. For this reason, a test's accuracy should be critically evaluated before the test can be implemented in clinical practice, and clinicians should be aware of the accuracy of the tests they apply.

Numerous diagnostic accuracy studies are being initiated each year. These studies evaluate a test's ability to determine whether individuals have a specific target condition. Figure 3 illustrates the typical design of a diagnostic accuracy study. A series of study participants that are clinically suspected of having the target condition first undergo the test under evaluation, which is referred to as the index test. The same participants then undergo a test that is considered to have a very high - or, preferably, perfect - ability for establishing the presence or absence of the target condition. This test is referred to as the clinical reference standard. The results of these two tests are cross-classified, and discrepancies represent index test misclassifications: false positive index test results and false negative index test results. From this cross-classification estimates of measures of diagnostic accuracy can then be calculated, such as sensitivity, specificity, predictive values and likelihood ratios.[6]

Ideally, after researchers complete their diagnostic accuracy study, they publish a corresponding study report that not only contains the study findings, but also provides sufficient details about the study's rationale, objectives, design, methods, and analyses. The Declaration of Helsinki, which provides ethical principles for medical research, specifies that “researchers have a duty to make publicly available the results of their research on human subjects and are accountable for the completeness and accuracy of their reports”.[7] Unfortunately, this ethical obligation is not always fulfilled: researchers may fail to publish their study,[8] or, if published, the study report may not be as informative as it could be.[9]

Nowadays, clinicians are trained to practice according to the principles of evidence-based medicine.[10] Clinical decisions should not solely be based on a clinician’s professional experience or convictions, but must be supported by the best available scientific evidence. In performing evidence-based medicine, clinicians heavily rely on systematic reviews and clinical practice guidelines, which present thorough and critical summaries of the available literature on a specific topic, and commonly provide recommendations for clinical practice based on this. However, if the published literature is incomplete or insufficiently informative in its reporting, such recommendations may not fully represent the best available evidence, which could hamper clinicians in their approaches to perform evidence-based medicine, thereby unnecessarily putting those that seek medical care at risk.

Figure 3. Typical design of a diagnostic accuracy study.



A **series of patients**, suspected of having the target condition…

…first undergo the **index test**,…

…and then undergo the **clinical reference standard**,…

…after which the results of both tests are **cross-classified**,…

…where discrepancies are considered to be index test misclassifications:

false positive and false negative test results.

Aim and outline

The general aim of the studies presented in this thesis was to uncover deficiencies in the process of publishing and reporting diagnostic accuracy studies, with the ultimate goal of increasing value in diagnostic research. This thesis consists of four parts.

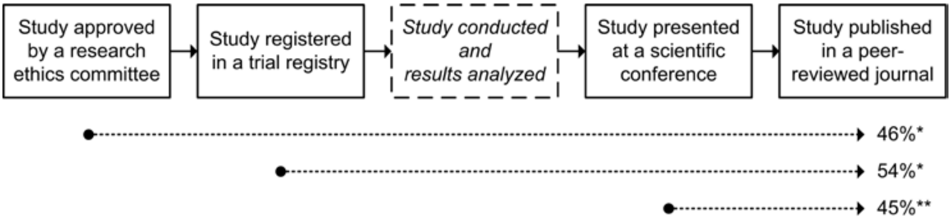
Part A: Publication of full study reports

Many completed biomedical studies take years to get published, while others are never published at all (Figure 4).[11-15] If studies are published, the final study report often only contains a selection of the outcomes that were pre-defined in the original study protocol. This failure to publish study results in a timely manner is problematic for reasons that have been thoroughly documented.[8,16] Other research groups, unaware of the existence of these unpublished studies, may for example duplicate research efforts with no or limited added value, which will lead to a waste of research resources.

Failure to publish can also negatively impact patient care if this is driven by the the nature and direction of results, a phenomenon that has been defined as 'reporting bias'.[17] There is convincing evidence that studies with statistically significant findings favoring the therapeutic intervention under investigation are more likely to be published then those that do not.[12-15] This results in a published literature base that is more optimistic about the efficacy of therapeutic interventions than can be justified based on all existing evidence, both published and unpublished. Whether similar problems exist among diagnostic accuracy studies has rarely been investigated.[18-20]

We evaluated to what extent diagnostic accuracy studies that had been registered in a clinical trial registry were subsequently published in full, and how often there were undisclosed major discrepancies in registered and published outcomes (**Chapter 1**). We also analyzed whether diagnostic accuracy studies that had been presented as conference abstracts at a major international ophthalmology meeting were subsequently published in full, and examined associations between the magnitude of the reported accuracy estimates and full publication (**Chapter 2**). We then studied the time from completion of participant recruitment to publication among published diagnostic accuracy studies, and whether this was associated with the magnitude of the reported accuracy estimates (**Chapter 3**).

Figure 4. Proportions of trials of therapeutic interventions reaching full-text publication in a peer-reviewed journal.


Study approved by a research ethics committee
Study registered in a trial registry
Study conducted and results analyzed
Study presented at a scientific conference
Study published in a peer-reviewed journal
46%*
54%*
45%**

*Results are from Schmucker et al.[12] **Results are from Scherer et al.[14]

Part B: Prospective registration of study protocols

To prevent some of the negative effects of failure to publish, publically accessible trial registries have been established, such as ClinicalTrials.gov and the Netherlands Trial Registry.[21-23] The International Committee of Medical Journal Editors (ICMJE) now requires that clinical trials are included in such a registry before initiation of participant enrollment.[24] If this requirement is not fulfilled, the trial will not be considered for publication in ICMJE's member journals. More and more governmental bodies, funders, and academic institutions have implemented similar policies to enforce trial registration.[25]

The Cochrane Collaboration promotes the performance of high-quality systematic reviews. In its Handbook for Systematic Reviews of Interventions it specifies that "prospective trial registration […] has the potential to substantially reduce the effects of publication bias",[17] which is a form of reporting bias. If all study protocols are registered, unpublished studies can be identified, selective reporting can be prevented, and unnecessary duplication of research efforts can be avoided.[21-24] The Declaration of Helsinki urges researchers that registration is also an ethical obligation: "Every research study involving human subjects must be registered in a publicly accessible database before recruitment of the first subject".[7] It is largely unknown, however, how many diagnostic accuracy studies are currently being registered.

We evaluated to what extent published diagnostic accuracy studies were registered in clinical trial registries (**Chapter 4**). We also performed a survey among journal editors, with the aim of assessing whether journals currently adhere to ICMJE's trial registration policy (**Chapter 5**).

Part C: Informative reporting of study reports

Published reports of biomedical studies are often insufficiently informative.[9] This certainly also applies to diagnostic accuracy studies.[26] Although the design of a diagnostic accuracy study, as described in Figure 3, may seem straightforward, the interpretation of the results of these studies can be challenging. The main reason for this is that diagnostic accuracy studies often host sources of bias and variation.[27,28] Bias may result from methodological shortcomings and usually leads to overestimations of the index test's ability to detect the target condition. Variation refers to the phenomenon that a test may not be equally accurate if applied under different clinical circumstances. It is crucial that those that read a report of a diagnostic accuracy study are able to assess the extent to which bias may have occurred and to whom the study results apply.

To guide authors in writing fully informative reports of diagnostic accuracy studies, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) reporting guideline was developed, and first published in 2003.[29,30] The STARD Group recently developed STARD 2015, an update of the original guideline.[31] In preparing this update, the STARD Group considered it useful to find out to which extent the quality of reporting improved after the launch of STARD, and what the current state of reporting is.

First, we performed a systematic review of studies that evaluated adherence of reports of diagnostic accuracy studies to STARD (**Chapter 6**). Then we performed our own evaluation of adherence of reports of diagnostic accuracy studies to STARD, and analyzed whether improvements in reporting quality were made over time (**Chapter 7**). We also analyzed how informative journal and conference abstracts of diagnostic accuracy studies are (**Chapter 8** and **9**). We also reported in detail on the methods used in the development of STARD 2015 (**Chapter 10**).

Part D: Diagnostic tests in respiratory medicine

Where the previous parts have focused on issues in the publication and reporting of diagnostic accuracy studies in general, this fourth part focuses on the performance of several diagnostic tests in respiratory medicine. Asthma and lung cancer are highly prevalent airway diseases.[32,33] In the clinical work-up of these diseases, clinicians use many different tests, not only for diagnosis, but also when selecting patients that are likely to respond to specific treatments. Asthma is a heterogeneous disease that consists of multiple phenotypes.[34] Disease prognosis varies across these phenotypes, as does the optimal treatment strategy. In patients with lung cancer, the stage of disease directly determines the prognosis and treatment options.[33,35] Accurate phenotyping of patients with asthma and accurate staging of patients with lung cancer is therefore important for prognostic and treatment selection purposes, illustrating the central role that tests can have in clinical practice.

Both for asthma phenotyping and for lung cancer staging, the number of available tests has been growing rapidly over the past decades. In using these tests, clinicians rely on diagnostic accuracy studies, trusting the published literature to be complete and sufficiently informative.

We performed a systematic review of minimally invasive markers for airway eosinophilia, a specific phenotype in patients with asthma (**Chapter 11**). In this review, we thoroughly searched for unpublished research results as well, and compared whether these differed from published results. Next to this review, we performed a diagnostic accuracy study in which we aimed to establish thresholds for minimally invasive markers for ruling-in and ruling-out airway eosinophilia in patients with asthma, and to develop a multivariable model with improved accuracy (**Chapter 12**). By adhering to STARD, we aimed to be as complete as possible in our reporting. We also performed a systematic review in which we evaluated to which extent the combined use of endobronchial and esophageal endosonography improves accuracy for mediastinal nodal metastases in patients with lung cancer, as compared to using either test alone (**Chapter 13**). Finally, we assessed whether improved staging of lung cancer results in a survival benefit, by analyzing follow-up data from a randomized clinical trial that compared two staging strategies, one of them incorporating the combined use of endobronchial and esophageal endosonography (**Chapter 14**).

Part A

Publication of full study reports

Chapter 1

Publication and reporting of test accuracy studies registered in ClinicalTrials.gov

Daniël A. Korevaar
Eleanor A. Ochodo
Patrick M. Bossuyt
Lotty Hooft

Clinical Chemistry 2014;60(4):651-9

Appendices and supplemental material are available online at:
http://www.clinchem.org/content/vol60/issue4

Abstract

Background

Failure to publish and selective reporting are recognized problems in the biomedical literature, but their extent in the field of diagnostic and prognostic testing is unknown. We aimed to identify non-publication and discrepancies between registered records and publications among registered test accuracy studies.

Methods

We identified studies evaluating a test's accuracy against a reference standard that had been registered in ClinicalTrials.gov between January 2006 and December 2010. We included studies that were completed before October 2011, allowing at least 18 months until publication. We searched PubMed, Embase, and Web of Science, and contacted investigators for publications. We examined associations between study characteristics and publication status for studies that had been completed at least 30 months prior to our searches.

Results

Overall, we included 418 studies, of which 224 (54%) had been published by mid-2013, with a median time to publication of 18 months (IQR 9 to 28). Among studies that provided an exact completion date and had been completed at least 30 months prior to our searches, 45% (128/282) were published within 30 months after their completion. After removing studies that were registered after study completion, and studies with an unknown (instead of completed) recruitment status, study duration was the only characteristic significantly associated with publication, with lower rates in studies lasting up to one year (20/51; 39%) compared to studies of 13-24 months (34/55; 62%) or longer (29/43; 67%) (p=0.01). In the 153 published studies that had been registered before completion, 49 (32%) showed discrepancies between the registered record and publication regarding the inclusion criteria for study participants (n=19), the index test or corresponding positivity threshold (n=9), or the outcomes (n=32).

Conclusions

Failure to publish and selective reporting are prevalent in test accuracy studies. Their registration should be further promoted among researchers and journal editors.

Introduction

In recent years, failure to publish studies and selective reporting of research findings, each related to the strength and direction of outcomes,[17,36] have been demonstrated several times in the biomedical literature.[15,37] Studies with favorable results were shown to be more likely to be published than studies with negative or disappointing ones.[15,38] This is regrettable for several reasons. The non-reporting of research results may lead to unnecessary duplication of research efforts, wasting time and money. Furthermore, the absence of information in the public domain can affect the evidence base on which clinical decisions are made.[39] Systematic reviews, which have now achieved a fundamental role in modern evidence-based healthcare, are especially sensitive to the selective absence of study findings, since unpublished research results are difficult to find and include. This may lead to skewed syntheses of the evidence, biased estimates of the effectiveness of healthcare interventions, and eventually, unnecessary exposure of patients to potentially ineffective or harmful interventions.[40]

In 2005 the International Committee of Medical Journal Editors (ICMJE) decided to require researchers to register essential information about the design of randomized controlled trials before study initiation before study initiation,[24] in a publicly accessible register such as ClinicalTrials. gov.[21] Study registration is seen as an important solution for recognizing reporting biases since, in principle, non-publication and selective reporting can easily be identified.

So far, most demonstrations of failure to publish and selective reporting have targeted randomized clinical trials.[13] The extent to which similar mechanisms are active in research estimating the accuracy of diagnostic and prognostic medical tests and markers is largely unknown.[19,20,41] Although registration of test accuracy studies is currently not required by the ICMJE, increasing numbers of these studies seem to be registered. The main objectives of our study were to identify non-publication and discrepancies between registered records and corresponding publications in a cohort of test accuracy studies registered in ClinicalTrials.gov. We also explored associations between study characteristics and non-publication.

Methods

Creation of ClinicalTrials.gov cohort

A cohort of test accuracy studies registered in ClinicalTrials.gov was identified. Details on the registry search are provided in Supplemental Methods 1, available online. The search was limited to studies that had a ‘first received date’ between January 1, 2006, and December 31, 2010.

Studies were included if their objective was to evaluate the accuracy of a medical test (the index test) in correctly classifying human subjects as having the target condition, evaluated against a clinical reference standard. The results of such studies are typically reported in terms of sensitivity and specificity, predictive values, or area under the receiver operating characteristic curve. Because outcomes are sometimes vaguely registered, or not registered at all, we also included studies that did not explicitly mention any accuracy measure but for which one could be calculated on the basis of the information included in the registry. Studies that only evaluated the analytical or technical performance of a test were excluded.

We selected studies for which the study 'completion date' was set before October 2011, thereby allowing at least 18 months between intended completion and publication. Also, studies with an unknown (instead of completed) 'recruitment status' in ClinicalTrials.gov were included if they met this criterion. The status of these studies is characterized as unknown by the registry if it has not been verified within the past two years. If no 'completion date' was provided, the 'primary completion date' was used (n=18). If neither date was reported, a study was included only if the 'recruitment status' had been updated to completed before October 2011, as recorded in the 'history of changes' option in ClinicalTrials.gov (n=8). Studies were excluded whenever their 'recruitment status' in the registered record indicated that the study had been withdrawn, terminated, or suspended. One author (D.A.K.) scanned the search results to identify studies meeting the inclusion criteria. If there was any doubt that a study met the inclusion criteria, the case was discussed with a second author (P.M.B.).

Identification of publications

In April and May 2013, one author (D.A.K.) undertook the following steps to identify matching publications in peer-reviewed biomedical journals. First, the 'publications' field in the ClinicalTrials.gov record was examined. If no reference of an article was reported, Medline (through PubMed) and Embase were searched by use of the trial registration number, registry title, name of the principal investigator or other contact person, index test, and/or target condition (including synonyms). When no publication was identified, Web of Science was searched by combining the search strategy with the institution, city, and/or country where a study was performed.

In May 2013, a second author (E.A.O.) repeated the search for studies for which no publication had been identified. If still no publication was found, one author (D.A.K.) tried to contact the principal investigator of the study by email, by use of contact information provided in the registry, or, if this was outdated or not available, by searching for corresponding email addresses through previously published studies or Google. Emails were sent in June 2013. Contact attempts were limited to three emails, each a week apart. If no answer was received, a study was considered as unpublished. If there was any doubt that a study identified through our searches matched with the registered record, the investigators were also contacted for confirmation. If no response was received, the case was again discussed with another author (P.M.B.). If a study had other objectives besides investigating the accuracy of a test, the study was considered as published only if the results on test accuracy were reported in a publication.

Data extraction

One author (D.A.K.) performed the data collection. Included studies were categorized as registered before initiation (defined as registered before or in the same month as the registered 'start date'), after completion (defined as registered in the same month as or after the registered 'completion date'), or between initiation and completion (defined as registered after the month of the registered 'start date' but before the month of the registered 'completion date'). We also extracted whether the study's 'recruitment status' was completed or unknown. Based on the registered information, we also classified studies as those evaluating imaging tests, laboratory tests, or other types of tests, and as diagnostic accuracy studies or prognostic accuracy studies.

The 'funder type' of a study is categorized in ClinicalTrials.gov into National Institutes of Health (NIH), other US federal agency, industry, or all others (including individuals, universities, and non-profit organizations). We further categorized NIH and other US federal agency into 'government'. We categorized the country in which the study was performed as US versus Canada, Australia, and New Zealand versus European Union and Switzerland versus other country.

The study duration was obtained by calculating the number of months between the registered 'start date' and 'completion date'. We also extracted the anticipated sample size from the 'enrollment' or 'estimated enrollment' field in the registry. If this number had changed during the course of the study, we took the first one recorded, according to the 'history of changes' option.

Identification of discrepancies between registries and publications

One author (D.A.K.) compared each registered record with the final publication regarding inclusion criteria for study participants, index tests, and corresponding positivity thresholds, and primary outcomes and secondary accuracy outcomes. Studies registered after completion were excluded from this analysis, as a formal assessment of selective reporting would not be possible in these studies. Studies were subdivided into those with clear discrepancies between the registry and publication, those where an unambiguous appraisal of discrepancies was not possible due to vague, retrospectively added, or unavailable information in the registered record, and those with no or minor discrepancies. If a reason for deviating from the registered information was provided in the publication, it was not considered as a discrepancy.

When comparing outcomes, clear discrepancies were defined as those where a registered primary outcome had been omitted from the publication, a registered primary outcome had become secondary, a registered secondary outcome had become primary, an outcome absent in the registry had become primary, a registered secondary accuracy outcome had been completely omitted, or the timing of assessment had changed. In this classification, minor discrepancies were not taken into account. We considered discrepancies in the outcomes as minor when all the registered accuracy measures were reported in the publication, but other unregistered measures of accuracy were reported as well (e.g., only sensitivity and specificity were registered, but likelihood ratios were reported in the publication in addition to these), and when only a selection of the registered accuracy measures were reported in the publication (e.g., only sensitivity and specificity were reported, whereas likelihood ratios were not).

The classification of each study was confirmed by a second author (L.H.), with disagreements resolved by consensus.

Statistical analysis

Descriptive statistics are reported as frequencies and percentages, and as medians with IQRs for skewed continuous variables.

We performed Kaplan-Meier survival analysis to estimate the time to publication, defined as the time (in months) between study completion (as reported in the registry) and publication (defined as the date when the article was first published in full online or appeared in print, whichever came first). Only studies that had registered a 'completion date' were included in this analysis, thereby removing those that had only registered a 'primary completion date', or no date at all. Publication times for unpublished studies were considered censored on March 2013, the month before we started our searches. Studies published before the completion date (n=23) were considered as published at time 0 in these analyses.

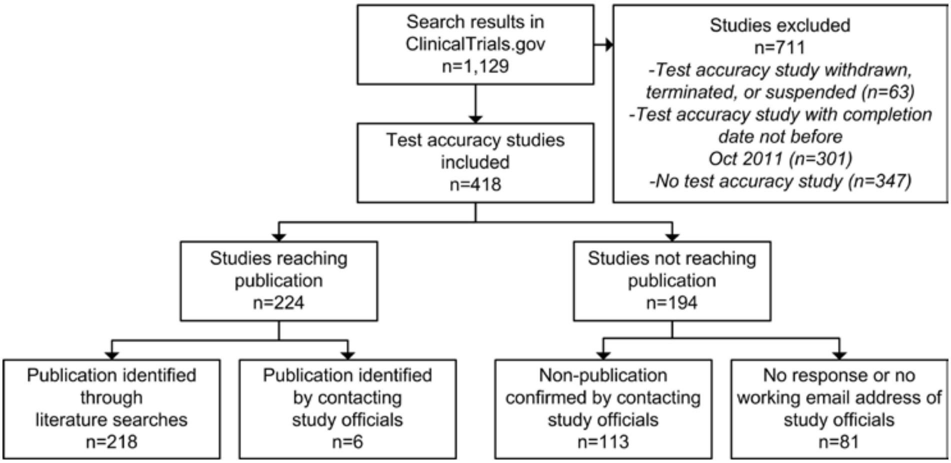
We used χ^2 tests to examine associations between study characteristics and publication status, allowing at least 30 months between study completion and publication. We also analyzed these data when excluding studies that were registered after completion or with an unknown (instead of completed) 'recruitment status', since these studies were highly likely to respectively over- and underestimate publication rates. Data were analyzed in SPSS version 20 (IBM, Armonk, NY, USA).

Results

Search and selection

The search in ClinicalTrials.gov identified 1,129 studies, of which 711 had to be excluded (Figure 1). In the excluded studies, 63 had been withdrawn (n=17), terminated (n=36), or suspended (n=10). Of these, 54 studies provided in total 56 reasons for not completing the study. Reported reasons referred to recruitment problems (n=17); sponsor decisions, financial problems, or lack of time or manpower (n=15); design problems (n=5); a principal investigator who had moved (n=4); institutional review board (IRB) decisions (n=2); a product being withdrawn from the market (n=2); or other reasons (n=11).

Figure 1. Flowchart for selection of studies and identification of corresponding publications.


Search results in ClinicalTrials.gov n=1,129
Studies excluded n=711
-Test accuracy study withdrawn, terminated, or suspended (n=63)
-Test accuracy study with completion date not before Oct 2011 (n=301)
-No test accuracy study (n=347)
Test accuracy studies included n=418
Studies reaching publication n=224
Studies not reaching publication n=194
Publication identified through literature searches n=218
Publication identified by contacting study officials n=6
Non-publication confirmed by contacting study officials n=113
No response or no working email address of study officials n=81

Study characteristics

Characteristics of the 418 included studies are summarized in Table 1. In short, almost a quarter of the studies had been registered after completion, and a quarter of the studies had not changed their status to completed, although the completion date indicated that they should have been finished. Nearly half of the studies investigated an imaging test, with very few prognostic tests. A quarter of the studies were funded by industry, and only 4% by government. The remaining 297 studies had another type of funder, mostly universities (n=171). The great majority of the studies had been performed in the US (36%), or in the European Union and Switzerland (32%). The median study duration was 19 months (IQR 11 to 30), and the median sample size was 172 (IQR 80 to 412).

Table 1. Characteristics of included studies.

	n
Total	418
Registration	
Before initiation	144 (34%)
Between initiation and completion	180 (43%)
After completion	94 (23%)
Recruitment status	
Completed	315 (75%)
Unknown	103 (25%)
Type of test evaluated	
Imaging test	179 (43%)
Laboratory test	119 (29%)
Other type of test	109 (26%)
Multiple test categories	11 (3%)
Aim of test evaluated	
Diagnostic test	385 (92%)
Prognostic test	28 (7%)
Both diagnostic and prognostic test	5 (1%)
Funder type	
Industry	106 (25%)
Government	15 (4%)
Other	297 (71%)
Country[a]	
USA	147 (36%)
Canada, Australia, New Zealand	30 (7%)
European Union, Switzerland	131 (32%)
Other country	74 (18%)
Multiple categories	31 (8%)

[a]Five studies were excluded from this analysis because country was unknown.

Publication

No publication could be found for 194 of the 418 included studies. The investigators of 113 of these studies (58%) confirmed that the study had not been published (Figure 1). Five of them indicated that the study was accepted for publication, four manuscripts had been rejected, and 12 had been submitted and were awaiting editorial decisions. Three studies were still ongoing, seven had been stopped early and their results had not (yet) been published, and one had never started. The other responding investigators only answered that the study had not (yet) been published.

Of the included studies, 224 (54%) had resulted in one (n=184), two (n=18), three (n=14), four (n=4), five (n=3), or seven (n=1) published articles. A reference to at least one of the published articles was provided in the 'publications' field in the registry for 80 (36%) published studies. Of the 224 published studies, 129 (58%) included the ClinicalTrials.gov registration number in at least one corresponding article. Of these, only 65 (50%) provided the registration number in the abstract. Seventy-four published studies (33%) both provided a reference to the publication in the registry and included a registration number in the publication. The estimated median time to publication was 35 months (95%CI 29.0 to 41.0) in the Kaplan-Meier analysis. For published studies, the median time to publication was 18 months (IQR 9 to 28).

Overall, studies registered after completion were more likely to be published (76%) than studies registered before initiation (42%) or studies registered between initiation and completion (51%, p<0.001) (Supplemental Table 1). Studies for which the 'recruitment status' indicated that the study was completed were more often published (59%) than studies with an unknown status (36%, p<0.001) (Supplemental Table 1).

When we excluded studies registered after completion and studies with an unknown recruitment status, we found that 120 of 228 studies (53%) had been published, with an estimated median time to first publication of 33 months (95%CI 23.7 to 42.3) in the Kaplan-Meier analysis (Figure 2). The median time to publication of published studies in this subgroup was 15 months (IQR 8 to 25).

In studies that had registered a 'completion date', 49 of 67 (73%) completed before 2008, 41 of 59 (70%) completed in 2008, 48 of 90 (53%) completed in 2009, 53 of 107 (50%) completed in 2010, and 22 of 69 (32%) completed in 2011 were published (p<0.001). When we excluded studies registered after completion and studies with an unknown (instead of completed) recruitment status, these rates were 65% (11/17), 64% (25/39), 50% (25/50), 56% (38/68), and 38% (18/47), respectively (p=0.119).

Publication by study characteristics

Publication rates by study characteristics for 302 studies that had been completed at least 30 months prior to our searches are presented in Table 2. Overall, 179

(59%) of these had been published. Of these 302 studies, 282 provided an exact 'completion date', of which 128 (45%) were published within 30 months after their completion. Besides timing of registration (before initiation versus after completion versus in between) and recruitment status (completed versus unknown), the country where the study had been performed was significantly associated with study publication, but after excluding studies registered after completion and studies with an unknown (instead of completed) recruitment status, this association was no longer present.

Study duration was significantly associated with publication, after exclusion of studies registered after completion and studies with an unknown recruitment status; fewer studies lasting up to one year were published (39%) than studies of one to two years (62%), or longer (67%; p=0.01). We observed no significant associations between publication rates and other study characteristics, such as type and aim of the test, funder type, or sample size.

Figure 2. Time from study completion to publication, excluding studies registered after completion and those with an unknown (instead of completed) 'recruitment status'.

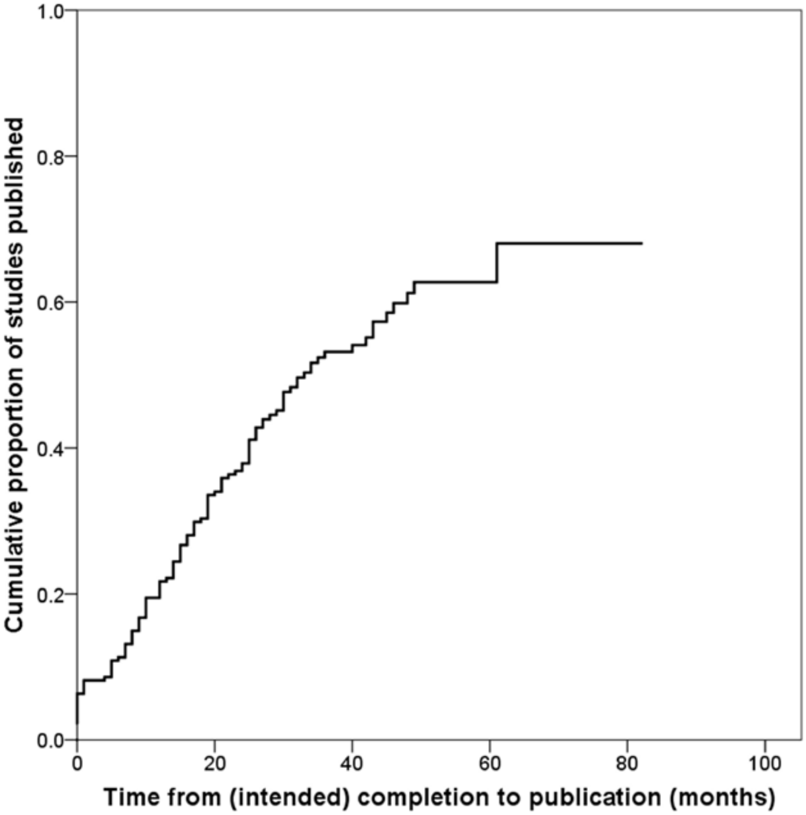


Table 2. Summary of publication rates by study characteristics for studies completed before October 2010, at least 30 months prior to our searches.

		All studies			**Excluding studies registered after completion and studies with an unknown recruitment status**	
	n	**Published n**	***p-value***	**n**	**Published n**	***p-value***
Total	302	179 (59%)		154	86 (56%)	
Registration						
Before initiation	90	45 (50%)	0.001	66	37 (56%)	0.96
Between initiation and completion	122	66 (54%)		88	49 (56%)	
After completion	90	68 (76%)			-	
Recruitment status						
Completed	237	150 (63%)	0.007	154	86 (56%)	-
Unknown	65	29 (45%)			-	
Type of test evaluated						
Imaging test	125	72 (58%)	0.08	63	38 (60%)	0.14
Laboratory test	83	47 (57%)		43	20 (47%)	
Other type of test	86	58 (67%)		43	27 (63%)	
Multiple test categories	8	2 (25%)		5	1 (20%)	
Aim of test evaluated						
Diagnostic test	276	161 (58%)	0.29	140	77 (55%)	0.44
Prognostic test	23	15 (65%)		12	7 (58%)	
Both	3	3 (100%)		2	2 (100%)	
Funder type						
Industry	75	39 (52%)	0.34	49	25 (51%)	0.71
Government	8	5 (63%)		5	3 (60%)	
Other	219	135 (62%)		100	58 (58%)	
Country						
USA	110	57 (52%)	0.01[a]	67	33 (49%)	0.40
Canada, Australia, New Zealand	22	9 (41%)		15	7 (47%)	
European Union, Switzerland	92	65 (71%)		36	24 (67%)	
Other country	54	37 (69%)		20	13 (65%)	
Multiple categories	19	11 (58%)		16	9 (56%)	
Study duration						
0-12 months	105	56 (53%)	0.09[b]	51	20 (39%)	0.01[c]
13-24 months	90	53 (59%)		55	34 (62%)	
25 months or more	87	60 (69%)		43	29 (67%)	
Sample size						
Below or equal to median	145	84 (58%)	0.72[d]	81	49 (61%)	0.20[e]
Above median	145	87 (60%)		70	35 (50%)	

[a] Five studies were excluded from this analysis because no country was registered. [b] Twenty studies were excluded from this analysis because no 'completion date' was registered. [c] Five studies were excluded from this analysis because no 'completion date' was registered. [d] Twelve studies were excluded from this analysis because no '(estimated) enrollment' was registered; median sample size was 153. [e] Three studies were excluded from this analysis because no '(estimated) enrollment' was registered; median sample size was 200.

Comparison between registries and publications

Seventy-one of 224 (32%) published studies had been registered after completion and were excluded from our comparisons between registries and corresponding publications. Of the remaining 153 published studies, 49 (32%) showed clear discrepancies between the registry and the publication regarding the inclusion criteria for study participants, the index test or corresponding positivity threshold, or the outcomes.

The inclusion criteria had changed in 19 (12%) studies. An unambiguous appraisal of discrepancies was difficult in 10 (7%) other studies, because the inclusion criteria were much more vaguely reported in the registered record than in the final publication.

Nine (6%) studies showed discrepancies in the index test or corresponding positivity threshold: in eight studies, one or more registered index test(s) were not reported in the publication, and in one study, the registered positivity threshold value for the index test differed from the published one. We were unable to completely exclude discrepancies in another 23 (15%) studies: information on the index test was more vaguely reported in the registry than in the publication for eight studies, and among 19 studies that reported a predefined positivity threshold value in the methods section of their publication, 15 did not register this value.

A comparison of registered and published outcomes was not possible in 22 studies (14%) because no outcomes had been registered (n=6), outcomes had been registered after study completion (n=4), or outcomes had been registered much more vaguely (n=12). Of the remaining 131 studies, 32 (24%) showed clear discrepancies: a registered primary outcome had been omitted in the publication (n=14), a registered primary outcome had become secondary (n=7), a registered secondary outcome had become primary (n=6), an outcome absent in the registry had become primary (n=7), a registered secondary accuracy outcome had completely been omitted (n=11), or the timing of assessment had changed (n=2).

Many studies showed discrepancies that we considered as minor, and were not taken into account in the classification above. Registered outcomes were often unspecific regarding the accuracy measures that would be calculated (n=43); instead, vague or general terms such as “diagnostic value” or “diagnostic accuracy” were used. Of the studies that were specific about the accuracy measures that would be calculated, 23 reported all the registered accuracy measures in the publication but added several unregistered others in the publication, and for seven, only a selection of the registered accuracy measures were reported in the publication. Primary and secondary outcomes were often not explicitly distinguished in published papers, and in 11 publications, the registered primary outcome seemed at least equally important as the registered secondary outcome.

Discussion

We evaluated failure to publish and selective reporting in a cohort of test accuracy studies registered in ClinicalTrials.gov. Only slightly more than half of the studies that had been completed 18 months or longer before our analyses were found to be published in a peer-reviewed biomedical journal. Although publication rates increased over time, about one-third of the studies completed before 2009 were still unpublished by mid-2013. Discrepancies between the registered record and publication regarding the inclusion criteria for study participants, index tests, positivity thresholds, and/or primary outcomes or secondary accuracy outcomes appeared in one-third of the published studies that had been registered before study completion. Unfortunately, an unambiguous assessment of selective reporting was not always possible due to scarce, absent, or retrospectively registered information.

We acknowledge that our study has several potential limitations. We may have missed publications, despite our efforts to identify published reports. Two authors thoroughly searched three databases, and responses to our email survey were satisfactory; non-publication was confirmed for 58% of the studies for which we did not identify a publication. For our analyses we relied on information provided in the registry but this information has proven to be not always accurate. For example, the registered completion date differed sometimes from the published completion date, and several studies were published before the registered completion date. The estimated time from completion to publication may therefore be inaccurate and can be expected to be longer. We cannot generalize our results to unregistered test accuracy studies without some form of caution. The ICMJE currently does not require registration for test accuracy studies, and such registration may therefore happen selectively. However, it seems unlikely that publication rates are higher among unregistered studies. Authors who are willing to register their test accuracy study may be more aware of the negative effects of failure to publish and, consequently, be more motivated to publish results, even when unsatisfactory. Several studies registered their outcomes near the end of the study, when the direction of the results was probably already known. This may have affected our estimates of discrepancy rates between registries and publications.

Failure to publish and selective outcome reporting are widely recognized problems in the biomedical literature. Evidence from large cohorts of registered clinical trials, both randomized and non-randomized, suggests that only between 46% and 63% get published.[42-44] We found similar results among test accuracy studies: 54% of the studies completed at least 18 months before our analyses, and 59% completed at least 30 months before our analyses, were published. These numbers may be considered as optimistic, because we excluded studies that had enrolled patients but discontinued before study completion. Two recent evaluations with study designs similar to ours assessed time to publication among registered clinical trials funded by the NIH and the National Heart, Lung, and Blood Institute.[45,46] Respectively 46% (294/635) and 57% (132/244) of the trials had been published within 30 months of completion, which is comparable to our findings. In other research fields, industry-funded trials have been associated with lower publication rates.[42,43] Besides study duration, we did not identify study characteristics significantly associated with publication.

In previous studies, inconsistencies between registered and published primary outcomes varied between 18% and 49%.[42,47,48] In our cohort, 24% of the studies that had been registered before study completion showed inconsistencies between the registered and published primary outcomes and/or secondary accuracy outcomes. Performing multiple statistical inferences increases the risk of false-positive findings. The selective reporting of multiple outcomes does not allow the reader of a study to be aware of the magnitude of this risk. In addition, not providing a predefined threshold, but rather estimating one on the basis of the collected data, gives room for manipulation and will usually lead to inflated estimates of test accuracy that are hard to reproduce.[49]

To our knowledge, this is the first investigation of failure to publish and selective reporting in such a large and general cohort of test accuracy studies. We are aware of only one similar project. Brazelli et al. Investigated publication rates in a much smaller cohort of conference abstracts of test accuracy studies in stroke research and found that 76% (121/160) were subsequently published in full.[41] They did not evaluate discrepancies between protocols and publications.

Our results indicate that the problems of failure to publish and selective reporting also appear among test accuracy studies. This would mean that study registration is equally important in this field of research. Although the fact that more and more test accuracy studies are being registered is promising, our results also show that, at this point, study registration for test accuracy studies needs more guidance. The majority of studies are not registered before initiation, registered information varies widely between studies, and essential information (including outcomes) is often registered vaguely or after study completion, if at all. A registration number was provided in slightly more than half of the publications of registered studies, making it difficult to assess whether a study has been previously registered. A reference to a publication was provided in the registry for only one-third of published studies, which hampers determination of whether a registered study is published. Many authors also seem to forget to change the status of their study to completed, even among published studies.

Failure to publish and selective reporting among test accuracy studies threaten patient safety because adoption of medical tests into clinical practice on the basis of an incomplete evidence base may lead to inadequate medical decision making.[18,50] Patients may be subjected to the side effects of unnecessary medical interventions on the basis of an erroneous diagnosis or withdraw from an intervention on the basis of an erroneous prognosis. In addition, tests may have potential complications and side effects, and inadequate testing increases pressure on healthcare funds. This is a particular worry in times of economic recession and a continuous increase in healthcare costs. Healthcare policymakers should be able to make an objective appraisal of any given test, on the basis of all the available evidence.

We recommend more research into the extent, drivers, and consequences of failure to publish and selective reporting in test accuracy studies. An obvious next step would be to follow up on a cohort of IRB-approved protocols of test accuracy studies. This way, a more exact estimation of publication rates and a more adequate assessment of discrepancies between the original protocols and the final publications can be made. Above all, we strongly recommend that study registration becomes a requirement for test accuracy studies. However, before implementing such a requirement, guidelines specifically designed for registration of test accuracy studies should be developed.

Acknowledgments

We thank René Spijker for his comments on the search in ClinicalTrials.gov.

Chapter 2

Reported estimates of diagnostic accuracy in ophthalmology conference abstracts were not associated with full-text publication

Daniël A. Korevaar
Jérémie F. Cohen
René Spijker
Ian J. Saldanha
Kay Dickersin
Gianni Virgili
Lotty Hooft
Patrick M. Bossuyt

Journal of Clinical Epidemiology 2016
E-published ahead of print

Appendices and supplemental material are available online at:
http://www.jclinepi.com/article/S0895-4356(16)30173-1/addons

Abstract

Objective

To assess whether conference abstracts that report higher estimates of diagnostic accuracy are more likely to reach full-text publication in a peer-reviewed journal.

Methods

We identified abstracts describing diagnostic accuracy studies, presented between 2007 and 2010 at the Association for Research in Vision and Ophthalmology (ARVO) Annual Meeting. We extracted reported estimates of sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and diagnostic odds ratio (DOR). Between May and July 2015, we searched Medline and Embase to identify corresponding full-text publications; if needed, we contacted abstract authors. Cox regression was performed to estimate associations with full-text publication, where sensitivity, specificity, and AUC were logit transformed, and DOR was log transformed.

Results

A full-text publication was found for 226 out of 399 (57%) included abstracts. There was no association between reported estimates of sensitivity and full-text publication (hazard ratio (HR) 1.09 (95%CI 0.98 to 1.22)). The same applied to specificity (HR 1.00 (95%CI 0.88 to 1.14)), AUC (HR 0.91 (95%CI 0.75 to 1.09)), and DOR (HR 1.01 (95%CI 0.94 to 1.09)).

Conclusions

Almost half of the ARVO conference abstracts describing diagnostic accuracy studies did not reach full-text publication. Studies in abstracts that mentioned higher accuracy estimates were not more likely to be reported in a full-text publication.

Introduction

There is abundant evidence that many biomedical studies never reach full-text publication in a peer-reviewed journal.[12,14,51] Studies with statistically significant results are published more often than those with non-significant results.[12-15,36] The resulting overrepresentation of 'positive' findings in the biomedical literature may introduce publication bias when researchers try to synthesize the available evidence, such as in systematic reviews and clinical practice guidelines.[52] Failure to publish studies jeopardizes adequate patient care and stifles scientific progress, while violating the ethical responsibility to disseminate study findings and use healthcare and research funds appropriately.[8]

Diagnostic accuracy studies assess how well a medical test differentiates between patients with and without a specific target condition. Accurate tests are important in clinical practice because false positive results could expose patients to unnecessary medical interventions, while false negative results could lead to withholding needed treatments. Although there are many investigations of the extent and drivers of publication bias among studies of therapeutic interventions, similar investigations are rare for diagnostic accuracy studies.[18,19]

Failure to reach full-text publication has recently been shown to be problematic among diagnostic accuracy studies, but the mechanisms of such failures are largely unknown.[41,53,54] Statistical significance is unlikely to be a major determinant, because most of these studies present their results in terms of accuracy estimates such as sensitivity, specificity, and area under the receiver operating characteristic curve (AUC), without clear hypothesis tests and accompanying p-values.[20,49,55,56] Nevertheless, when promising findings in conference abstracts, reflecting strong performance of diagnostic tests, more easily reach full-text publication, overoptimistic impressions of a test's accuracy can result. This could invite premature adoption or inappropriate clinical use of tests.

The objective of this study was to evaluate the extent to which diagnostic accuracy studies presented as conference abstracts at an international ophthalmology meeting reached full-text publication in a peer-reviewed journal and to assess associations between reported accuracy estimates and full-text publication. We hypothesized that abstracts reporting higher estimates of diagnostic accuracy would more often lead to full-text publication.

Methods

Selection of conference abstracts

Conference abstracts were considered for inclusion in our study if they were presented between 2007 and 2010 at the annual meeting of the Association for Research in Vision and Ophthalmology (ARVO), the world's largest gathering of eyes and vision researchers. This timeframe was selected to ensure that we would include a large number of around 400 abstracts and that abstract authors would have a sufficient amount of time for full-text publication.

Abstracts were eligible for inclusion if the authors reported on a study that assessed the accuracy of one or more diagnostic tests to establish a clinical diagnosis in humans, and calculated - or announced the calculation of - at least one of the following measures of diagnostic accuracy: sensitivity, specificity, predictive values, likelihood ratios, AUC, diagnostic odds ratio (DOR), or total accuracy. We excluded abstracts reporting on the prognostic or predictive accuracy of tests, evaluated against a future event or the outcome of treatment. We also excluded abstracts for which updated results were reported in another included abstract.

Potentially eligible abstracts were identified by searching ARVO's online abstract proceedings. The full search strategy was developed by two investigators (D.A.K., in consultation with R.S., a medical information specialist). It consists of 34 different terms that are commonly used in reports of diagnostic accuracy studies (Appendix A, available online). All retrieved abstracts were screened for inclusion by one investigator (D.A.K.). Whenever there was any doubt whether an abstract fulfilled the inclusion criteria, the case was discussed with a second investigator (J.F.C. and/or P.M.B.).

Data extraction

One investigator (D.A.K.) extracted data from the included ARVO abstracts, and a second investigator (J.F.C.) verified all extracted data. Disagreements were resolved through discussion.

We extracted the first author, year of presentation at ARVO, number of authors, continent of first author, international affiliations (authors from multiple countries versus all authors from one country). We also extracted declared conflicts of interest (at least one author versus none of the authors), acknowledged funding for the study (industry versus non-industry only versus none), whether a trial registration number was provided, study design (cohort versus case-control versus unclear), data collection (prospective versus retrospective versus unclear), research field (glaucoma versus ocular surface and corneal diseases (keratoconus and dry eye) versus common chorioretinal diseases (diabetic retinopathy and age-related macular degeneration) versus other), and number of participants and eyes.

For each abstract, we also extracted the highest reported estimate of sensitivity, specificity, AUC, and DOR.[6] The DOR is a single statistic summarizing the results of a 2x2 table; higher values represent better performance.[57] Because not all diagnostic accuracy studies report a DOR, we recalculated this from reported pairs of estimates of sensitivity and specificity, positive and negative predictive value, or positive and negative likelihood ratio, or from the AUC, using standard formulas.[57,58] In this recalculation, a correction needed to be applied to accuracy estimates of 0 or 1; these were considered to be 0.01 and 0.99, respectively.[57]

Identification of full-text publications

Between May and July 2015, at least five years after each included abstract was presented at ARVO, we undertook the following steps to identify corresponding full-text publications in peer-reviewed journals. Similar search strategies were used in previous related projects:[41,53,59]

1. For each abstract, one investigator (D.A.K.) searched Medline (through PubMed) and Embase (through Ovid) by separately using the abstract's first, second, and last author's name, combined with (synonyms of) the test(s) under investigation and/or (synonyms of) the target condition. First, the titles and abstracts of retrieved articles were screened; if potentially corresponding to the ARVO abstract, the full-text was assessed.
2. If unsuccessful, a second investigator (R.S.) repeated the search, and additionally searched Google Scholar, using the same strategy.
3. If still no full-text publication could be identified, we tried contacting abstract authors via email. One investigator (D.A.K.) searched for email addresses of two abstract authors through their previous publications and institutional websites. These two authors were successively contacted, each with two reminders, if necessary. If no response was received or if no working email address of any authors could be identified, the abstract was considered to not have reached full-text publication.

We matched ARVO abstracts and full-text publications by comparing authors' names, dates of participant recruitment, participant characteristics, and technical details about the diagnostic tests applied. Abstracts were considered to have reached full-text publication if at least some of the presented diagnostic accuracy data were reported in the corresponding publication. This means that if an abstract reported on the accuracy of two tests and the publication only reported on the accuracy of one of these, the abstract was considered to have reached full-text publication. However, if the abstract and publication corresponded to the same study, but the publication did not report on test accuracy or only reported on the accuracy of a test that was not presented in the abstract, the abstract was considered to not have reached full-text publication.

If an abstract corresponded to a publication, but reported results were discrepant, the abstract was still considered to have reached full-text publication. This was also the case if an abstract corresponded to a publication that described a lower or higher number of participants. If multiple abstracts corresponded to a single publication, all were considered have reached full-text publication.

If there was any doubt whether an abstract and full-text publication matched, the case was discussed within the research team (D.A.K., with G.V. and/or P.M.B.). If doubt persisted, study investigators were contacted by email for confirmation. For each full-text publication, we considered the date the article was added to the PubMed database as the publication date. If multiple full-text publications corresponded to one abstract, the date of the first publication was selected.

Statistical analysis

We calculated the overall proportion of abstracts reaching full-text publication and the median time from presentation at ARVO to full-text publication.

Univariable Cox proportional hazards regression analyses were performed, and hazard ratios (HR) were calculated, to analyze whether the accuracy estimates reported in ARVO abstracts were associated with full-text publication. For these analyses, sensitivity, specificity, and AUC were logit transformed, and DOR was log transformed. These transformed accuracy estimates were added as continuous variables to the regression model. Abstracts without accuracy estimates were included in the regression model by adding an indicator of missingness. Full-text publications that were published before or at the date of presentation of the abstract were arbitrarily considered published one month after presentation. Publication times for abstracts that did not reach full-text publication were considered censored at May 2015, the month in which we started our searches for corresponding publications. A p-value of ≤ 0.05 was considered statistically significant.

In addition, we used χ^2 tests to explore the association between other abstract characteristics and full-text publication. In this explorative analysis, a Bonferroni correction was applied to adjust for multiple testing, where a $p \leq 0.004$ was considered statistically significant. All statistical analyses were performed in SPSS version 22 (IBM, Armonk, NY, USA).

Results

Search and selection

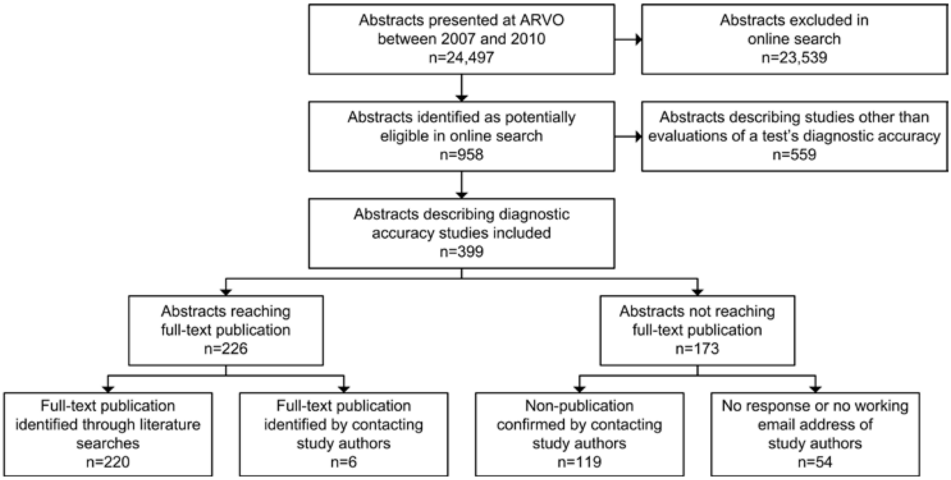
In total, 24,497 abstracts were presented at ARVO between 2007 and 2010, of which 958 were identified in our search (Figure 1). After screening the abstracts, 399 could be included. References of included abstracts are provided in Appendix B.

Abstract characteristics

Characteristics of the included ARVO abstracts are reported in Table 1. Disagreements between the two reviewers occurred in 1% (75/6,783) of extracted data elements. The median number of authors was five (IQR 4 to 7), and most first authors were affiliated with organizations in the USA (n=151; 38%), followed by Germany (n=34; 9%) and the UK (n=29; 7%). Some abstracts (n=75; 19%) contained authors from multiple countries. In 133 (33%) abstracts, at least one author declared a conflict of interest, but industry funding for the study was acknowledged in only 37 (9%) abstracts.

A trial registration number was provided in 26 (7%) abstracts, all referring to ClinicalTrials.gov. Most abstracts described a case-control study (n=219; 55%), and almost half referred to glaucoma research (n=186; 47%). The median number of participants was 107 (IQR 55 to 223), with a median number of 140 eyes (IQR 75 to 267).

Figure 1. Flowchart for selection of abstracts and identification of corresponding full-text publications.


Abstracts presented at ARVO between 2007 and 2010 n=24,497
Abstracts excluded in online search n=23,539
Abstracts identified as potentially eligible in online search n=958
Abstracts describing studies other than evaluations of a test's diagnostic accuracy n=559
Abstracts describing diagnostic accuracy studies included n=399
Abstracts reaching full-text publication n=226
Abstracts not reaching full-text publication n=173
Full-text publication identified through literature searches n=220
Full-text publication identified by contacting study authors n=6
Non-publication confirmed by contacting study authors n=119
No response or no working email address of study authors n=54

Full-text publication

For 226 of 399 (57%) ARVO abstracts, we found a corresponding full-text publication in a peer-reviewed journal. Of these, 220 (97%) were identified through our literature searches, and 6 (3%) by contacting study authors (Figure 1). For 15 of 226 (7%) abstracts that reached full-text publication, the number of participants in the abstract was more than 10% greater than the number of participants reported in the corresponding full-text publication. Among abstracts that reached full-text publication, the median time from presentation to publication was 17 months (IQR 8 to 29) (Figure 2). Thirteen full-text publications were published before the date of presentation of the corresponding abstract.

We confirmed non-publication by email contact with the authors of 119 of 173 (69%) abstracts for which we were unable to identify a matching full-text publication (Figure 1). The number of participants was reported for 138 of 173 abstracts that did not reach full-text publication and totaled 50,500. An overview of proportions of abstracts reaching full-text publication across subgroups defined by abstract characteristics is provided in Table 1.

Figure 2. Time from presentation to full-text publication among ARVO abstracts describing diagnostic accuracy studies.

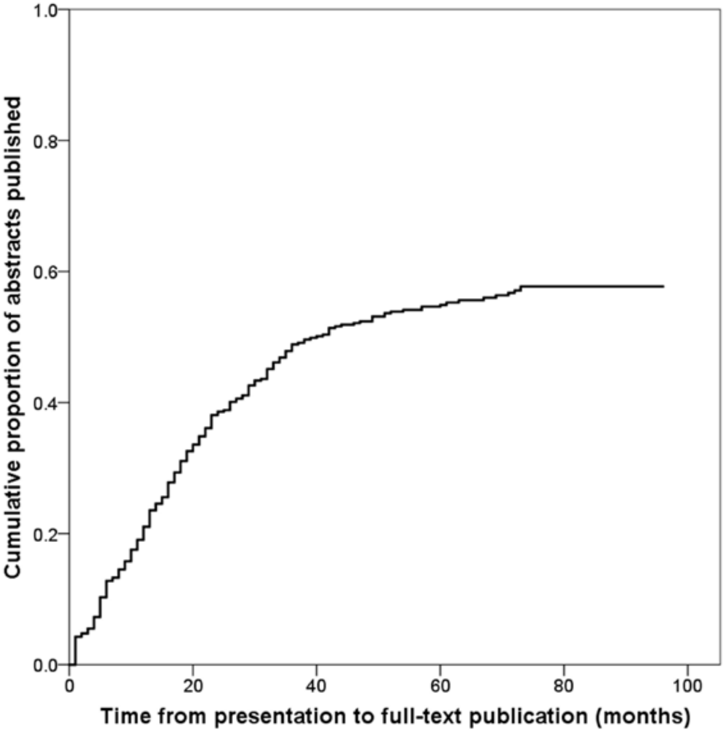


Table 1. Characteristics of ARVO abstracts and association with full-text publication.

	All abstracts	Abstracts that reached full-text publication	Abstracts that reached full-text publication[a]
	n	n	%
Total	399 (100%)	226	57%
Year of presentation at ARVO			
2007	75 (19%)	41	55%
2008	102 (26%)	65	64%
2009	96 (24%)	52	54%
2010	126 (32%)	68	54%
Number of authors			
< 5	144 (36%)	91	63%
≥ 5	255 (64%)	135	53%
Continent of first author			
Asia	64 (16%)	44	69%
Europe	130 (33%)	66	51%
North America	165 (41%)	86	52%
Oceania	18 (5%)	15	83%
South America	22 (6%)	15	68%
International affiliations			
Authors from multiple countries	75 (19%)	45	60%
All authors from one country	324 (81%)	181	56%
Conflicts of interest			
At least one author	133 (33%)	74	56%
None of the authors	266 (67%)	152	57%
Funding for study			
Industry	37 (9%)	20	54%
Non-industry only	194 (49%)	117	60%
None	168 (42%)	89	53%
Trial registration number provided			
Yes	26 (7%)	16	62%
No	373 (94%)	210	56%
Study design			
Cohort	139 (35%)	83	60%
Case-control	219 (55%)	124	57%
Unclear	41 (10%)	19	46%
Data collection			
Prospective	54 (14%)	35	65%
Retrospective	37 (9%)	20	54%
Unclear	308 (77%)	171	56%
Research field			
Glaucoma	186 (47%)	99	53%
Ocular surface and corneal diseases	35 (9%)	17	49%
Common chorioretinal diseases	44 (11%)	24	55%
Other	134 (34%)	86	64%
Number of participants			
<100	150 (38%)	88	59%
100-1000	146 (37%)	81	56%
≥1000	31 (8%)	20	65%
Not reported	72 (18%)	37	51%
Number of eyes			
<100	71 (18%)	35	49%
100-1000	118 (30%)	70	59%
≥1000	18 (5%)	12	67%
Not reported	192 (48%)	109	57%

[a] None of these abstract characteristics were significantly associated with full-text publication, after applying a Bonferroni correction to adjust for multiple testing.

Accuracy estimates and full-text publication

We grouped abstracts by quartiles of accuracy estimates (Table 2). Across the ARVO abstracts, 63% of those reporting a sensitivity in the highest quartile reached full-text publication, compared to 48% of those reporting a sensitivity in the lowest quartile. These proportions were 61% and 62% for specificity, 58% and 71% for AUC, and 55% and 56% for DOR, respectively.

In Cox proportional hazards regression analyses, there was no statistically significant association between reported estimates of sensitivity and full-text publication (HR 1.09 (95%CI 0.98 to 1.22)) (Table 3). The same applied to specificity (HR 1.00 (95%CI 0.88 to1.14)), AUC (HR 0.91 (95%CI 0.75 to 1.09)), and DOR (HR 1.01 (95%CI 0.94 to 1.09)).

Table 2. Accuracy estimates reported in ARVO abstracts and association with full-text publication.

	All abstracts n (%)	Abstracts that reached full-text publication n	Abstracts that reached full-text publication %
Overall	399 (100%)	226	57%
Sensitivity[a]			
<0.78	62 (16%)	30	48%
0.78-0.87	61 (15%)	33	54%
0.87-0.95	64 (16%)	35	55%
≥ 0.95	65 (16%)	41	63%
No sensitivity reported	147 (37%)	87	59%
Specificity[a]			
<0.82	55 (14%)	34	62%
0.82-0.90	45 (11%)	24	53%
0.90-0.98	81 (20%)	39	48%
≥0.98	61 (15%)	37	61%
No specificity reported	157 (39%)	92	59%
AUC[a]			
<0.86	38 (10%)	27	71%
0.86-0.91	31 (8%)	17	55%
0.91-0.96	42 (11%)	28	67%
≥0.96	43 (11%)	25	58%
No AUC reported	245 (61%)	129	53%
DOR[a]			
<16.1	78 (20%)	44	56%
16.1-48.6	86 (22%)	51	59%
48.6-168.6	80 (20%)	46	58%
≥168.6	84 (21%)	46	55%
No DOR reported	71 (18%)	39	55%

[a] Accuracy estimates were categorized by quartiles.

Table 3. Accuracy estimates reported in ARVO abstracts (n=399) and hazard ratios of full-text publication.

	Hazard ratio[a] (95%CI)	***p-value***
Sensitivity		
Sensitivity (logit transformed)	1.09 (0.98-1.22)	0.126
No sensitivity reported[b]	1.31 (0.91-1.89)	0.151
Specificity		
Specificity (logit transformed)	1.00 (0.88-1.14)	0.951
No specificity reported[b]	1.07 (0.70-1.62)	0.763
AUC		
AUC (logit transformed)	0.91 (0.75-1.09)	0.291
No AUC reported[b]	0.61 (0.36-1.03)	0.065
DOR		
DOR (log transformed)	1.01 (0.94-1.09)	0.753
No DOR reported[b]	0.99 (0.62-1.58)	0.971

[a]Estimated using Cox proportional hazards regression analyses. [b]Abstracts without accuracy estimates were included in the regression model by adding an indicator of missingness.

Discussion

Almost half of the conference abstracts describing diagnostic accuracy studies presented at the annual ARVO meeting between 2007 and 2010 did not reach full-text publication, five years or more after presentation. This represents diagnostic accuracy data collected in at least 50,500 study participants for which findings were not fully reported.

This massive failure to reach full-text publication is in line with previous evaluations of conference abstracts in different fields of biomedical research. A Cochrane systematic review summarized 79 such evaluations, and found that on average only 45% of abstracts reached full-text publication.[14] One of these evaluations was performed among 327 abstracts that were randomly selected from all studies presented in 1985 at ARVO; a full-text publication could be identified for 63%.[60] Another evaluation was performed among 93 abstracts describing randomized trials that were presented between 1988 and 1989 at ARVO or the American Academy of Ophthalmology; a full-text publication could be identified for 66%.[61] More recently, it was found that among 513 abstracts describing randomized trials that were presented between 2001 and 2004 at ARVO, 45% reached full-text publication.[62]

Unfortunately, failure to publish is not a random phenomenon. There is overwhelming evidence of publication bias, caused by an overrepresentation of positive and favorable results in full-text publications, and leading to an overoptimistic literature base.[13,15] Such bias can also be observed for conference abstracts. The Cochrane systematic review cited previously found that conference abstracts reporting at least one statistically significant result were 30% (95%CI 14% to 47%) more likely to reach full-text publication than those that did not.[14] In our analysis, no associations between the accuracy estimates reported in the ARVO abstracts and full-text publication were observed.

Although investigations of failure to publish and its determinants in diagnostic research are scarce, our findings are in line with what has been found to date. Among 418 diagnostic accuracy studies that were registered between 2006 and 2010 in ClinicalTrials.gov, a full-text publication could be identified for 54%.[53] In an evaluation of 160 conference abstracts describing diagnostic accuracy studies that were presented between 1995 and 2004 at two international stroke meetings, a full-text publication was found for 76%; no association was observed with reported accuracy estimates.[41] In a similar evaluation of 250 abstracts describing diagnostic accuracy studies that were presented in 2009 at three dementia conferences, a full-text publication was identified for only 39%, but potential associations with reported accuracy estimates were not assessed.[54]

We found no evidence of publication bias in the process of publishing diagnostic accuracy studies in ophthalmology, but also examined the possibility that our findings may have been influenced by limitations in the design of our study. It is possible that the selective reporting of studies with favorable results already took place when deciding to submit an abstract to ARVO. If that is the case, bias would only have been detected if publication proportions had been assessed among (a selection of) all initiated diagnostic accuracy studies, not only those presented at ARVO.

We decided to focus our analysis on the highest accuracy estimates reported in each abstract, but many abstracts contained multiple accuracy outcomes, reporting performance for multiple tests, for different target conditions or across subgroups. It is possible that a study's highest accuracy estimates are not the ones that stimulate writing, submitting, or publishing a corresponding full study report. Ideally, we would have focused our assessment on each abstract's most important accuracy estimate. Unfortunately, this almost always is ambiguous in diagnostic accuracy studies because "primary" or "main" outcomes are rarely explicitly defined in abstracts or in full-texts.[49,53,63]

Although our sample size was relatively large, not all abstracts reported accuracy estimates, which limited the power to detect significant associations with full-text publication. Because abstract selection was done by only one investigator (DAK), some relevant abstracts may have been excluded by mistake.

Despite our efforts to contact authors of abstracts for which we did not find a full-text publication, 30% of those authors did not respond to our requests to confirm non-publication. This proportion is much lower than in previous related projects, and only six of the 125 (5%) authors who responded provided a full-text publication that we had missed in our literature searches. When extrapolating this to the 54 abstracts for which we did not receive a response, it is estimated that we have missed three full-text publications, which is less than 1% overall.

If publication bias is much less of a problem for diagnostic accuracy studies, a reason could be that these studies are fundamentally different from other types of studies. Most diagnostic accuracy studies lack an explicit, predefined hypothesis, and corresponding statistical testing of these hypotheses is a rarity.[49,55] It has been suggested that non-significant results are regarded as disappointing or uninteresting, and that investigators are less likely to spend time writing articles describing such findings,[64] whereas journal editors are less inclined to publish them.[65] Yet if a distinction between statistically significant and non-significant results is rarely made, authors have far greater freedom in interpreting the results and to "spin" them in a positive way, a phenomenon that is highly prevalent in diagnostic accuracy studies.[49,66] This may explain the absence of a strong association between high accuracy estimates and full-text publication; even lower accuracy estimates may be regarded as positive, not hampering writing a longer study report or submitting it to a journal.

To allow the identification of ongoing, terminated, unpublished, or selectively published clinical trials, registries such as ClinicalTrials.gov have been initiated.[21] In 2005, the International Committee of Medical Journal Editors (ICMJE) decided that, for future clinical trials submitted to its member journals, only those that had been registered in a trial registry before initiation of the study would be considered for publication.[24] Although the implementation of this policy by ICMJE journals can be improved,[67] the existence of policy has led to a dramatic increase in the number of registered trials.[22,68]

Diagnostic accuracy studies are not generally considered to be clinical trials: only 7% of the diagnostic ARVO abstracts included in this analysis provided a registration number, which is in line with a recent evaluation, in which we reported that only 15% of 351 diagnostic accuracy studies published in high-impact journals had been registered.[69] To prevent research waste, the scientific community should strongly consider enforcing registration of all diagnostic accuracy studies, or at least those that are prospective.[18,19,70,71] This would allow researchers and funders to avoid unnecessary duplication of research efforts and improve collaborations, whereas systematic reviewers and guideline developers can uncover all potentially eligible unpublished studies or study materials, and journals and peer reviewers can help minimize selective publication by identifying discrepancies between the registered record and the submitted study report.[18]

Inaccessible research is widely considered to be one of the largest sources of research waste.[8] Evidence is now accumulating that many diagnostic accuracy studies never reach full-text publication.[41,53] Although we found no evidence of publication bias in the process of publishing these studies, this failure to publish them cannot be justified for ethical, economic, and scientific reasons.[18,70] Changing this will need concerted action from all stakeholders, but it is an absolute must.[8,25]

Chapter 3

Time to publication among completed diagnostic accuracy studies: associated with reported accuracy estimates

Daniël A. Korevaar
Nick van Es
Aeilko H. Zwinderman
Jérémie F. Cohen
Patrick M. Bossuyt

BMC Medical Research Methodology 2016;16(1):68

Appendices and supplemental material are available online at:
https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0177-4

Abstract

Background

Studies evaluating the effectiveness of therapeutic interventions are not always reported, and those with statistically significant results are published more rapidly than those without. We analyzed whether diagnostic accuracy studies that report promising results about test performance are also published more rapidly.

Methods

We obtained all diagnostic accuracy studies included in meta-analyses of Medline-indexed systematic reviews that were published between September 2011 and January 2012. For each study, we extracted estimates of diagnostic accuracy (sensitivity, specificity, Youden's index), the completion date of participant recruitment, and the publication date. We calculated the time from completion to publication and assessed associations with reported accuracy estimates.

Results

Forty-nine systematic reviews were identified, containing 92 meta-analyses and 924 unique primary diagnostic accuracy studies, of which 756 could be included. Study completion dates were missing for 285 (38%) of these. Median time from completion to publication in the remaining 471 studies was 24 months (IQR 16 to 35). Primary studies that reported higher estimates of sensitivity (Spearman's rho=-0.14; p=0.003), specificity (rho=-0.17; p<0.001), and Youden's index (rho=-0.22; p<0.001) had significantly shorter times to publication. When comparing time to publication in studies reporting accuracy estimates above versus below the median, the median number of months was 23 versus 25 for sensitivity (p=0.046), 22 versus 27 for specificity (p=0.001), and 22 versus 27 for Youden's index (p<0.001). These differential time lags remained significant in multivariable Cox regression analyses with adjustment for other study characteristics, with hazard ratios of publication of 1.06 (95%CI 1.02 to 1.11; p=0.007) for logit-transformed estimates of sensitivity, 1.09 (95%CI 1.04 to 1.14; p<0.001) for logit-transformed estimates of specificity, and 1.09 (95%CI 1.03 to 1.14; p=0.001) for logit-transformed estimates of Youden's index.

Conclusions

Time to publication was significantly shorter for studies reporting higher estimates of diagnostic accuracy compared to those reporting lower estimates.

Introduction

Many completed biomedical studies take years to get published, if they get published at all.[12,14] Over the past decades, there have been increasing concerns about the resulting bias for those relying on a synthesis of the available literature in getting summary estimates of the effectiveness of therapeutic interventions.[8,40,52] There is now overwhelming evidence that studies with statistically non-significant results are less likely to result in a publication in a peer-reviewed journal than those with statistically significant results.[12-15] Evaluations have also shown that it takes more time before negative studies are published.[11,72-74] There are multiple reasons for non- or delayed publication of studies with non-significant findings. Researchers, anticipating low scientific impact, may be reluctant to write and submit the study report; journals, foreseeing low citation rates, may be less interested in publishing them.[64,65]

Diagnostic accuracy studies evaluate the ability of medical tests to differentiate between patients with and without a target condition. It is unknown whether such studies are also susceptible to differential publication processes, with studies that document disappointing results about a test's performance being less likely to be published in full, or published later, compared to studies reporting more promising findings.[41,53,54,75] In itself, statistical significance is unlikely to be a major determinant of time to publication among diagnostic accuracy studies; these studies typically present results only in terms of estimates of sensitivity and specificity, and most do not have specific hypothesis tests and accompanying p-values.[20,49,55,56] It is possible, however, that the sheer magnitude of the reported accuracy estimates can be seen as a measure of the favorability of the study findings, and that studies reporting higher accuracy estimates are published sooner than studies reporting lower accuracy estimates.

The objective of this study was to evaluate whether reported accuracy estimates were associated with time to publication among published diagnostic accuracy studies.

Methods

Selection of diagnostic accuracy studies

We relied on a set of 114 Medline-indexed systematic reviews of diagnostic accuracy studies, published in English between September 2011 and January 2012. These reviews were identified in a previous meta-epidemiological project from our research group. The search and selection process have been described in full elsewhere.[56]

These systematic reviews were included in the current evaluation if they contained one or more meta-analyses and provided 2x2 tables for the primary studies included in these meta-analyses, describing the number of true and false positive and negative results for the diagnostic test under investigation. For each primary study included in the meta-analyses, we then obtained the full study report or, if not available, the abstract.

Data extraction

For each primary study, two investigators (D.A.K., J.F.C.) independently extracted the test under evaluation and the 2x2 tables reported in the meta-analyses. These investigators also independently identified the publication date.

For Medline-indexed studies, the date on which the citation was added to the PubMed database was used as the publication date. For studies not indexed in Medline, we tried to obtain the publication date through Google Scholar, the journal website, or the full study report. Primary studies for which no publication date could be identified were excluded from further analysis, as were conference abstracts.

One investigator (D.A.K. or N.v.E.) then extracted additional data from the articles in which the primary studies were reported. A random 10% of this data extraction was independently verified by the other investigator; discrepancies occurred in 3 out of 632 (0.5%) verified characteristics.

We extracted the start date and completion date of participant recruitment, the date of first submission to the publishing journal, and the date the study was accepted for publication. If only the months but not the exact dates of participant recruitment were provided, start dates were rounded to the first day of that month, whereas completion dates were rounded to the last day of that month. If only years of participant recruitment were provided, start dates were rounded to January 1 of the starting year, and completion dates to December 31 of the completion year.

We also extracted the journal in which the study was published and corresponding 2014 impact factor (through Web of Knowledge), number of authors, country of first author, and type of data collection (prospective versus retrospective). Data extraction from study reports published in non-English language was performed with the help of native speakers, or using Google Translate. Any disagreements in the data extraction process were resolved through discussion.

Statistical analysis

For each included primary study, we recalculated estimates of sensitivity and specificity from the extracted 2x2 tables. Because tests may have a high sensitivity but a low specificity, or the other way around, we also calculated Youden's index (sensitivity *plus* specificity *minus* 1). This is a single measure of diagnostic accuracy that takes the whole 2x2 table into account.[76] If multiple 2x2 tables were available for one primary study - which could happen, for example, because multiple tests had been evaluated - the highest reported estimates of sensitivity, specificity, and Youden's index were used in the analyses.

Our analysis focused on time from completion to publication, defined as the time interval between the completion date and the publication date. This was further subdivided in time from completion to submission, and time from submission to publication.

We calculated Spearman's rho correlation coefficients between accuracy estimates and time from completion to publication; a negative correlation coefficient meaning that studies reporting higher estimates had shorter times to publication. To further quantify potential delays, estimates of sensitivity, specificity, and Youden's index were then dichotomized by a median split, and median times from completion to publication were compared using Mann-Whitney U tests. To explore more specifically in which phase potential delays in time from completion to publication occurred, this analysis was repeated for time from completion to submission, and for time from submission to publication. Studies with partially missing dates were only excluded from the analyses for that specific time interval.

We performed multivariable Cox proportional hazards regression analysis to evaluate the unconditional and conditional effect of accuracy estimates on the hazard of publication, adjusting for year of publication, journal impact factor (≥4 versus <4 or not available), number of authors, continent (Europe, North America, or Oceania versus Asia, Africa, or South America), type of test (imaging versus other), type of data collection (prospective versus retrospective or not reported), study duration (time interval between the start date and completion date), and number of participants in the 2x2 table, adding a frailty term per meta-analysis to account for systematic differences in time to publication between meta-analyses. In this analysis, accuracy estimates were logit transformed, where a correction was applied for accuracy estimates of exactly 0 or 1; these were considered to be 0.001 or 0.999. Other continuous variables were not transformed before adding them to the models. This analysis was also repeated for time from completion to submission, and for time from submission to publication.

Sensitivity analysis

We performed sensitivity analysis by excluding primary studies that only provided the year, but not the month or exact date of completion of participant recruitment, as these calculations of time from completion to publication were likely to be less accurate. We also performed sensitivity analysis by excluding studies that did not provide both a completion date and a submission date, thereby restricting the analysis to studies for which we had both time from completion to publication, time from completion to submission, and time from submission to publication (complete case analysis). Data were analyzed in SPSS version 22 (IBM, Armonk, NY, USA) and R version 3.0 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Selection of diagnostic accuracy studies

Details on the selection of studies and a list of included systematic reviews are provided in Additional files 1 and 2, available online. In total, 49 systematic reviews could be included in the current evaluation, containing 92 meta-analyses. Together, these meta-analyses contained 924 unique primary diagnostic accuracy studies. Of these, 168 (18%) had to be excluded because no publication date could be obtained (n=163), because they were conference abstracts (n=4), or because they had been retracted (n=1).

The remaining 756 primary diagnostic accuracy studies were included, corresponding to 1,088 2x2 tables, as some studies were included in multiple meta-analyses. A full study report could be obtained for 751 of these; for the other 5 studies, data extraction was performed using the abstract only.

Study characteristics

Nineteen diagnostic accuracy studies (3%) were published before 1990; 133 (18%) between 1990 and 2000; 527 (70%) between 2000 and 2010; and 77 (10%) between 2010 and 2012. They were published in 322 different journals, most frequently in *European Journal of Nuclear Medicine and Molecular Imaging* (n=30; 4%), *Radiology* (n=27; 4%), *American Journal of Roentgenology* (n=20; 3%), and *Journal of Clinical Microbiology* (n=20; 3%). The median impact factor was 3.1 (IQR 2.0 to 5.4).

Study reports were in 10 different languages, most frequently in English (n=726; 96%). The median number of authors was six (IQR 5 to 8). First authors were from

64 different countries, most frequently USA (n=153; 20%), Germany (n=60; 8%), and Japan (n=54; 7%).

The type of test under investigation was an imaging test for 387 studies (51%) and another type of test for 369 studies (49%). Data collection was prospective in 307 studies (41%), retrospective in 125 studies (17%), and not reported in 324 studies (43%). The median study duration was 22 months (IQR 12 to 37), with a median number of participants of 100 (IQR 49 to 255). The median accuracy estimates were 0.875 (IQR 0.73-0.97) for sensitivity, 0.899 (0.76-0.97) for specificity, and 0.684 (0.45-0.83) for Youden's index.

Time to publication: association with reported estimates of diagnostic accuracy

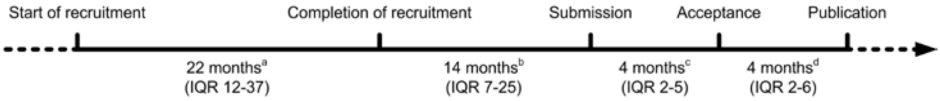
Of the included primary studies, 520 (69%) reported a submission date, 564 (75%) an acceptance date, 474 (63%) a start date, and 471 (62%) a completion date. Median times between study stages are summarized in Figure 1.

The median time from completion to publication (available for 471 studies) was 24 months (IQR 16 to 35). Sensitivity (rho=-0.14; p=0.003), specificity (rho=-0.17; p<0.001), and Youden's index (rho=-0.22; p<0.001) were each negatively correlated with time from completion to publication (Figure 2).

When comparing time from completion to publication in studies reporting accuracy estimates above versus below the median, the median number of months was 23 versus 25 for sensitivity (p=0.046), 22 versus 27 for specificity (p=0.001), and 22 versus 27 for Youden's index (p<0.001) (Table 1; Figure 3). Median time from completion to publication stratified by other categories of study characteristics is provided in Table 2.

These differential time lags remained significant in multivariable Cox regression analyses, with hazard ratios of publication of 1.06 (95%CI 1.02 to 1.11; p=0.007) for logit-transformed estimates of sensitivity, 1.09 (95%CI 1.04 to 1.14; p<0.001) for logit-transformed estimates of specificity, and 1.09 (95%CI 1.03 to 1.14; p=0.001) for logit-transformed estimates of Youden's index (Table 3).

Figure 1. Median times between study stages.



Median times missing for: [a]246; [b]426; [c]275; and [d]192 of 756 included studies.

When subdividing time from completion to publication, we observed significant associations between accuracy estimates and time from completion to submission (available for 330 studies), but not between accuracy estimates and time from submission to publication (available for 520 studies) (Table 1, with multivariable Cox regression analyses in Additional files 3 and 4).

Sensitivity analysis

The sign and significance of the association between estimates of diagnostic accuracy and time from completion to publication, time from completion to submission, and time from submission to publication remained the same when excluding studies that only reported the year of completion of participant recruitment but not the month or exact date, and when excluding studies that did not report both a completion date and a submission date (Additional file 5).

Figure 2. Correlations between reported estimates of diagnostic accuracy and time from completion to publication

Figure 2a. Sensitivity.

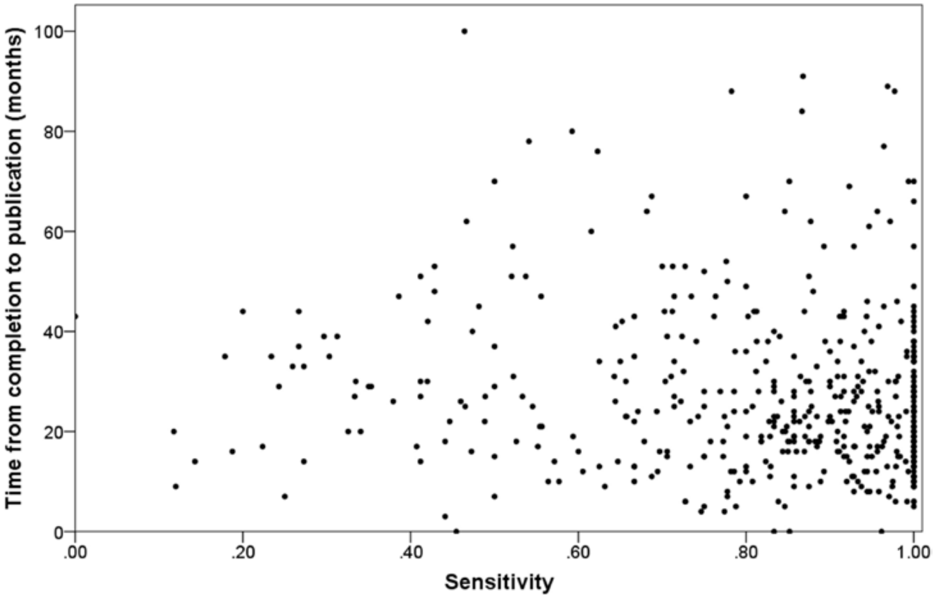


Figure 2. *Continued.*

Figure 2b. Specificity.

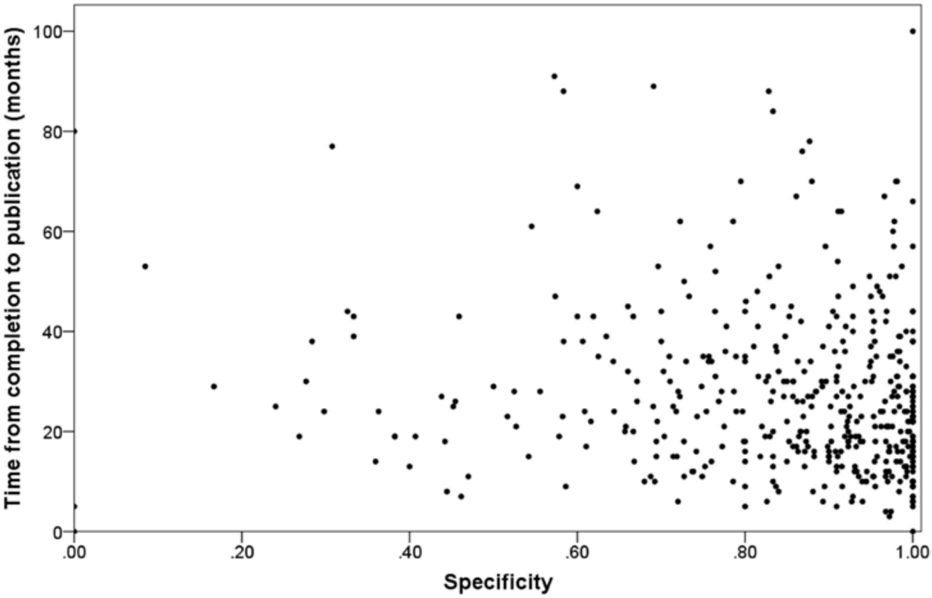


Figure 2c. Youden's index.

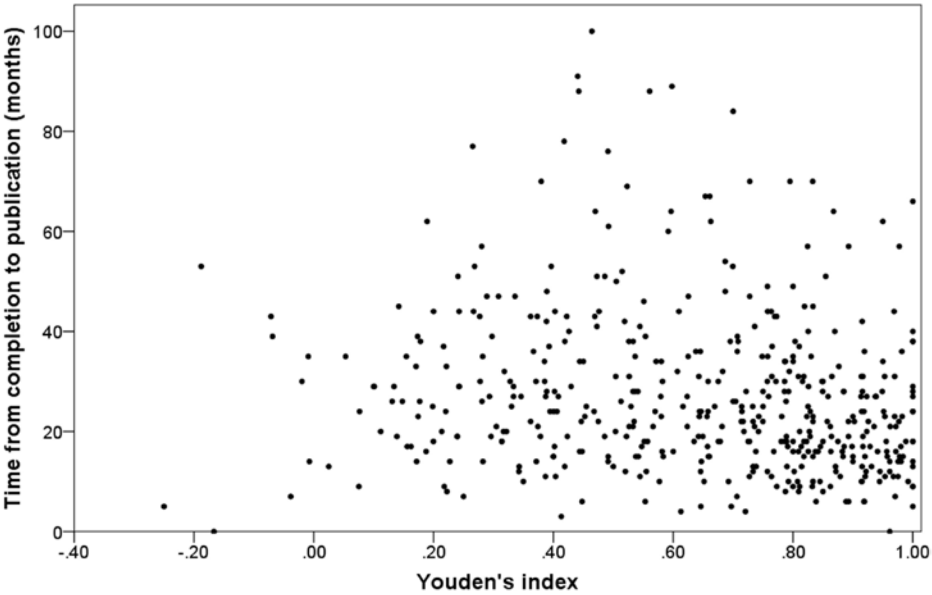


Table 1. Time to publication: association with dichotomized accuracy estimates.

	Time from completion to publication[a]			Time from completion to submission[b]			Time from submission to publication[c]		
	Studies n	**Months Median (IQR)**	***p-value***	**Studies n**	**Months Median (IQR)**	***p-value***	**Studies n**	**Days Median (IQR)**	***p-value***
Overall	471 (100%)	24 (16-35)		330 (100%)	14 (7-25)		520 (100%)	238 (177-329)	
Sensitivity[d]									
<0.875	226 (48%)	25 (16-39)	0.046	149 (45%)	16 (7-30)	0.037	225 (43%)	240 (176-324)	0.755
≥0.875	239 (51%)	23 (15-32)		179 (54%)	13 (7-22)		293 (56%)	238 (183-335)	
Specificity[e]									
<0.899	209 (44%)	27 (18-38)	0.001	152 (46%)	17 (9-30)	0.001	262 (50%)	238 (176-317)	0.494
≥0.899	250 (53%)	22 (15-31)		173 (52%)	12 (6-22)		252 (48%)	238 (180-332)	
Youden's index[f]									
<0.684	225 (48%)	27 (18-39)	<0.001	157 (48%)	17 (10-30)	<0.001	245 (47%)	235 (176-321)	0.251
≥0.684	230 (49%)	22 (14-31)		166 (50%)	11 (6-21)		267 (51%)	244 (180-332)	

Median times missing for: [a]285, [b]426, and [c]236 of 756 included studies. [d]Sensitivity, [e]Specificity, and [f]Youden's index missing for 7, 14, and 18 of 756 included studies, respectively.

Figure 3. Time from completion to publication.

Figure 3a. Sensitivity.

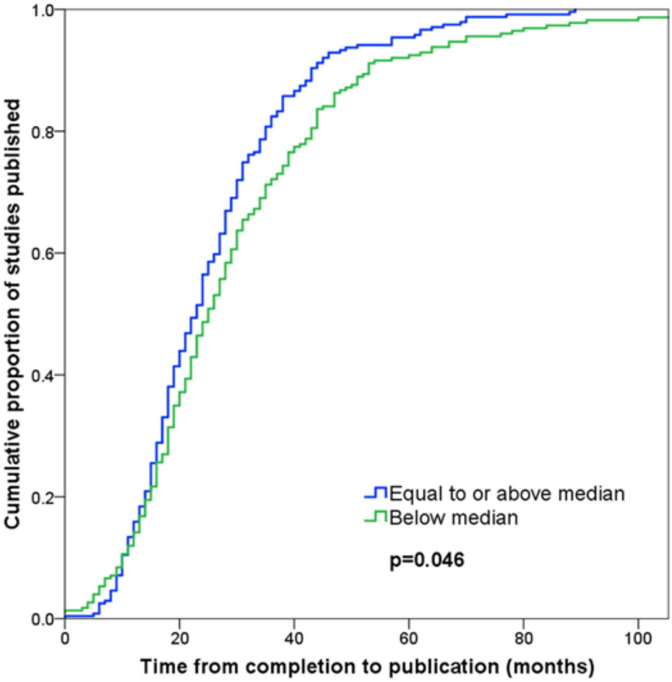


Figure 3b. Specificity.

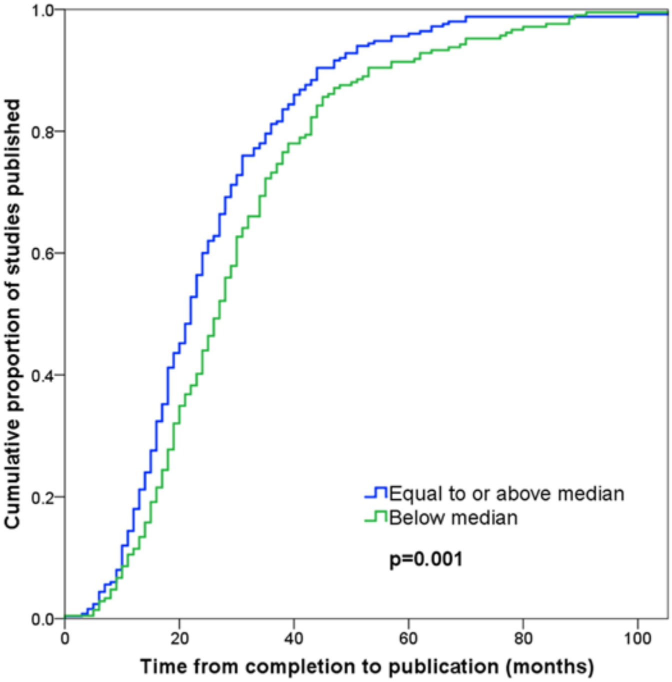
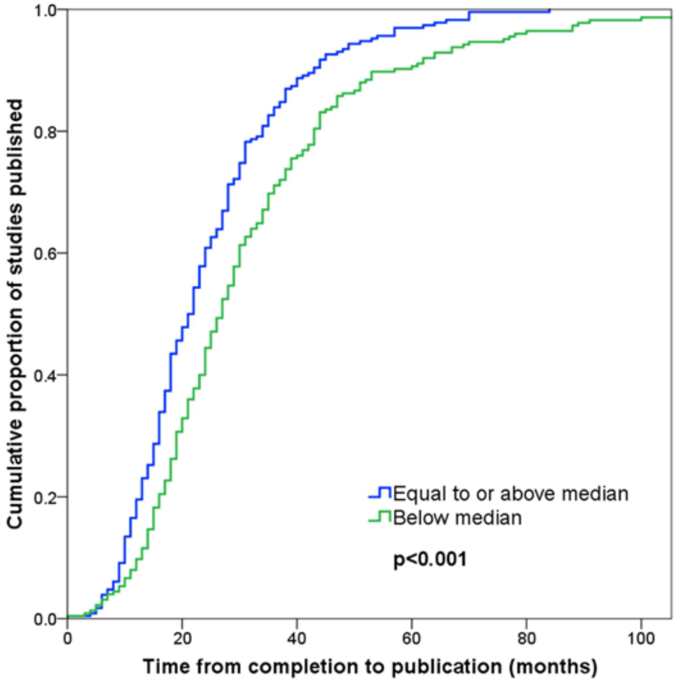


Figure 3. *Continued.*

Figure 3c. Youden's index.



Discussion

In a large sample of published diagnostic accuracy studies, we found that it took authors on average two years to publish study findings after completing participant recruitment. Time from completion to publication was significantly shorter for studies reporting higher estimates of diagnostic accuracy compared to those reporting lower estimates, a delay that could not be attributed to differences in speed of processing within the journals that eventually published the studies.

Some elements deserve consideration. Many reports of diagnostic accuracy studies contain multiple accuracy outcomes, for example, for different tests, target conditions, and subgroups. We only obtained the 2x2 tables that were used in the selected meta-analyses, but the primary studies may have focused on other accuracy outcomes as well. Whenever a study reported multiple 2x2 tables, we selected the highest accuracy estimates in our analysis, because in our personal experience authors have a tendency to emphasize these in their conclusions. However, whether the highest accuracy estimates in a study are indeed the ones that drive time to publication is unknown. In our analysis, we focused on dichotomized accuracy estimates, as this allowed us to provide a straightforward quantification of the delays that can be anticipated in the publication of results that are relatively disappointing in diagnostic research. We acknowledge that a dichotomization in terms of a median-split is arbitrary, and that this may not reflect the difference between statistically significant and non-significant results based on the p-value.

Table 2. Time from completion to publication: association with other study characteristics.

	Studies n	Months Median (IQR)
Overall[a]	471 (100%)	24 (16-35)
Year of publication		
<1990	10 (2%)	26 (17-53)
1990-1994	15 (3%)	18 (14-38)
1995-1999	44 (9%)	23 (15-36)
2000-2004	99 (21%)	26 (18-37)
2005-2009	243 (52%)	23 (16-35)
≥2010	60 (13%)	22 (15-35)
Journal impact factor		
<4 or not available	312 (66%)	25 (16-37)
4-9	129 (27%)	22 (15-33)
≥10	30 (6%)	22 (16-34)
Number of authors		
<6	170 (36%)	23 (14-38)
≥6	301 (64%)	24 (17-34)
Continent of first author		
Africa	32 (7%)	31 (20-48)
Asia	125 (27%)	20 (13-29)
Europe	183 (39%)	24 (17-35)
North America	116 (25%)	26 (16-39)
Oceania	10 (2%)	28 (21-44)
South America	5 (1%)	19 (14-50)
Type of test		
Imaging	229 (49%)	24 (16-34)
Other	242 (51%)	24 (16-37)
Type of data collection		
Prospective	192 (41%)	24 (17-34)
Retrospective	96 (20%)	25 (15-36)
Not reported	183 (39%)	24 (15-38)
Study duration[b]		
<13 months	115 (24%)	24 (15-34)
13-24 months	139 (30%)	25 (19-35)
≥25 months	216 (46%)	24 (14-37)
Number of participants		
<100	208 (44%)	24 (15-35)
100-999	232 (49%)	24 (16-35)
≥1000	31 (7%)	27 (20-39)

[a] Time from completion to publication missing for 285 of 756 included studies. [b] Study duration missing for 1 of 471 studies included in this analysis.

Table 3. Time from completion to publication: multivariable Cox regression analyses.

	Hazard ratio (95%CI)[a]	***p-value***
Model 1: Sensitivity (n=464)		
Sensitivity (logit transformed)	1.06 (1.02-1.11)	0.007
Year of publication (per 5 years)	0.98 (0.88-1.08)	0.660
Journal impact factor		
≥4	1.24 (0.99-1.55)	0.064
<4 or not available	1	
Number of authors	0.99 (0.96-1.02)	0.550
Continent of first author		
Europe, North America or Oceania	0.69 (0.55-0.87)	0.002
Africa, Asia or South America	1	
Type of test		
Imaging	1.15 (0.88-1.50)	0.300
Other	1	
Type of data collection		
Prospective	1.19 (0.96-1.48)	0.120
Retrospective or not reported	1	
Study duration (per year)[b]	1.00 (0.98-1.03)	0.820
Number of participants (per 1000)	1.01 (0.99-1.03)	0.430
Model 2: Specificity (n=458)		
Specificity (logit transformed)	1.09 (1.04-1.14)	<0.001
Year of publication (per 5 years)	1.01 (0.91-1.11)	0.910
Journal impact factor		
≥4	1.34 (1.07-1.67)	0.011
<4 or not available	1	
Number of authors	0.99 (0.96-1.02)	0.360
Continent of first author		
Europe, North America or Oceania	0.72 (0.57-0.90)	0.003
Africa, Asia or South America	1	
Type of test		
Imaging	1.15 (0.90-1.46)	0.260
Other	1	
Type of data collection		
Prospective	1.23 (1.00-1.52)	0.050
Retrospective or not reported	1	
Study duration (per year)[b]	1.00 (0.98-1.02)	0.980
Number of participants (per 1000)	1.01 (0.99-1.03)	0.540
Model 3: Youden's index (n=454)		
Youden's index (logit transformed)[g]	1.09 (1.03-1.14)	0.001
Year of publication (per 5 years)	0.98 (0.89-1.09)	0.730
Journal impact factor		
≥4	1.28 (1.02-1.61)	0.031
<4 or not available	1	
Number of authors	0.98 (0.95-1.01)	0.280
Continent of first author		
Europe, North America or Oceania	0.69 (0.55-0.87)	0.002
Africa, Asia or South America	1	
Type of test		
Imaging	1.16 (0.90-1.51)	0.250
Other	1	
Type of data collection		
Prospective	1.24 (1.00-1.54)	0.052
Retrospective or not reported	1	
Study duration (per year)[b]	1.00 (0.98-1.02)	0.910
Number of participants (per 1000)	1.01 (0.99-1.03)	0.460

[a] Frailty term added per meta-analysis to account for systematic differences in time from completion to publication between meta-analyses; variance of frailty terms was: model 1=0.103; model 2=0.064; model 3=0.094. [b] One study was excluded from the Cox regression analysis because of a missing study duration.

Although the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement invites authors to report start and completion dates of participant recruitment,[31] these were not provided by more than one-third of the studies. As a consequence, we could not include these studies in our analyses of time from completion to publication. This obviously limited the precision of our findings, but we do not know whether the included sample is a biased one. We included eight additional variables in our Cox regression analyses. It is conceivable that there are other unmeasured confounders in the association between accuracy estimates and time to publication as well. Especially several study characteristics that are associated with study quality and risk of bias, such as blinding of test readers and quality of the reference standard, may be relevant in this respect. We did not include these elements because they are often not reported, and the extent to which they induce bias varies substantially depending on the type of test under investigation and the clinical area in which the test is applied.[77,78] However, we believe that excluding these is more likely to have led to under- rather than overestimations of the associations identified in this study: poor study quality generally leads to inflated accuracy estimates, but will probably also increase time to publication as a result of critical peer reviewers and more journal rejections.

Several previous evaluations found a comparable differential time lag among studies of therapeutic interventions. A Cochrane review that aimed to document the association between statistically significant results and time to publication included two of such evaluations, together analyzing the fate of 196 initiated clinical trials.[11,72,73] On average, trials with significant results were published about two to three years earlier than those with non-significant results. In a similar, more recent evaluation of 785 initiated clinical trials, the estimated median time from completion to publication was 2.1 years for those with significant results, and 3.2 years for those with non-significant ones.[74] A differential time lag was not identified in another evaluation of 1,336 published clinical trials: both those with significant and non-significant outcomes had a median time to publication of 21 months.[79]

Similar evaluations are scarce for diagnostic accuracy studies, and so far limited to abstracts presented at scientific conferences in specific fields of research. In contrast to our findings, no systematic bias could be identified in these previous assessments. One study found a median time from presentation to full publication of 16 months for 160 conference abstracts of diagnostic accuracy studies in stroke research, but the hazard of full publication was not associated with reported estimates of Youden's index.[41] We recently found that the median time from presentation to publication was 17 months among 399 conference abstracts of diagnostic accuracy studies in ophthalmology research; there also, the hazard of publication was not associated with reported estimates of sensitivity and specificity.[80]

In the current evaluation, on average, it took two months less to publish studies with a sensitivity above the median, five months less to publish studies with a specificity above the median, and five months less to publish studies with a Youden's index above the median, compared to studies reporting estimates of these accuracy estimates below the median. Although these time lags can be considered as relatively minor, the potential implications, although difficult to overlook, may be worrisome for multiple reasons.

We believe that the observed differential time lag may reflect a larger underlying problem. A study's chances to reach full publication are likely to fade over time and with every rejection by a journal. This may lead to failure to publish the study and, consequently, to publication bias, since the study will be missing from the evidence base to those relying on databases of published articles.[52] Although there is strong evidence of such bias in syntheses of studies of therapeutic interventions, this topic has been insufficiently investigated for diagnostic accuracy studies.[41,53,54,75]

Even if studies with less favorable results are eventually published, a delay in their publication and associated dissemination can lead to misleading results in systematic reviews. Reporting bias may occur when literature reviewers want to synthesize the available evidence but cannot account for unfavorable results that take substantially longer to get published.[11] To assess time trends in published accuracy estimates, we recently applied cumulative meta-analysis to the same set of systematic reviews as used in the current evaluation.[81] Among 48 meta-analyses included, a total of 12 statistically significant time trends in sensitivity or specificity were identified. The majority of these time trends, 8 out of 12 (67%), were negative, which may indicate that studies that are published earlier sometimes tend to overestimate the accuracy of a test. This may be partially explained by a time lag in the publication of studies that report lower accuracy estimates, as identified in the current evaluation.

The delay in publishing studies with lower accuracy estimates could be attributed to study authors, who may be less motivated to write and submit corresponding study reports, to peer reviewers, who may be more critical towards and less supportive of studies with unfavorable results, or to journal editors, who may be less willing to publish studies reporting disappointing performance of new and existing tests.[64,65] In our evaluation, we did observe a differential time lag from study completion to submission, but not from submission to publication, indicating that delayed publication of studies with lower accuracy estimates was not caused by the journal that eventually published the study report. This suggests two alternative explanations for the delay. One is that authors were less effective, maybe even reluctant, in finalizing and submitting their study report. Another explanation is that the manuscript was not accepted by the journals where authors initially submitted the report to, and this could have been caused, in part, by the less positive findings.

In the multivariable Cox regression analysis for Youden's index, two additional variables were also significantly associated with time from completion to publication. Studies published in journals with a higher impact factor were published more rapidly. An explanation could be that authors first submit their study to higher impact factor journals, going down after each rejection, which would then delay publication. When pooling studies from Africa, Asia, and South America, these were published more rapidly than those from Europe, North America, and Oceania. Little is known about geographical differences in quality and rigorousness of the editorial and peer review processes of biomedical journals.

The findings of this study are relevant for scholars that want to arrive at a synthesis of the available evidence through a search of the literature, and for patients, clinicians, policy makers, and funders, that may rely on these literature syntheses. They should be fully aware that it is very well possible that not all completed diagnostic accuracy studies have been published at the time of the evaluation, and that this could introduce reporting bias. Such bias is likely to be more pronounced if only few published studies are available. As recommended in current guidelines,[75] additional efforts should be made to identify and include unpublished studies in systematic reviews, as this will strengthen the validity and improve the precision and applicability of the results.

Concerns about reporting bias were one factor that prompted the implementation of trial registration policies.[21] The International Committee of Journal Editors now only considers trials for publication if they were registered in a publically accessible trial registry before study start.[24] Unfortunately, currently only 15% of published diagnostic accuracy studies are being registered.[69] Over the past years, evidence that many diagnostic accuracy studies remain unpublished has accumulated,[41,53,54,75] making a strong case for a firmer implementation of registration policies for these studies.[18,19,70,71] The fact that the current evaluation suggests that there may also be bias in the process of publishing diagnostic accuracy studies further amplifies this message.

Registration of diagnostic accuracy studies would enable the identification of all relevant studies in a timely manner, not only those that have been published. Funders, governmental organizations and academic institutions could also require the publication of results within a year after study completion, as currently required by the Food and Drug Administration (FDA) for certain trials.[82] In an era of transparency and open access, stakeholders involved in biomedical research should make efforts to ensure that study results become available in a timely manner; this should apply to all studies, not just those presenting promising, optimistic, and fascinating results.[8,25]

Acknowledgments

The authors thank Eleanor A. Ochodo and W. Annefloor van Enst for the literature search, and Junfeng Wang for performing data extraction on study reports published in non-English language.

Part B

Prospective registration of study protocols

Chapter 4

Infrequent and incomplete registration of test accuracy studies: analysis of recent study reports

Daniël A. Korevaar
Patrick M. Bossuyt
Lotty Hooft

BMJ Open 2014;4(1):e004596

Abstract

Objective

To identify the proportion of articles reporting on test accuracy for which the corresponding study had been registered.

Methods

PubMed was searched for publications in journals with an impact factor of 5 or higher in May and June 2012. Articles were included if they reported on original studies evaluating the accuracy of one or more diagnostic or prognostic tests or markers against a clinical reference standard in humans. Primary outcome was the proportion of registered test accuracy studies. We additionally explored study characteristics associated with registration.

Results

We found 1,941 references; 351 study reports fulfilled the inclusion criteria, of which 52 studies (15%) had been registered. Of these, 27 (52%) provided a registration number in the publication, and 12 (23%) provided a reference to the publication in the registry. Registration rates were similar for studies on diagnostic versus those on prognostic tests, and among studies on imaging tests versus those on laboratory tests. Studies reporting some form of industry involvement were more often registered (33%) than studies reporting another source of funding (11%), and studies without a (reported) source of (external) funding (9%; p<0.001). Of the registered studies, eight (15%) were registered after completion, 14 were registered before initiation (27%), and 30 (58%) between initiation and completion. Only 16 studies (31% of registered studies; 5% of the total sample) had registered the published primary outcome measures before completion.

Conclusions

Few test accuracy studies published in higher impact journals are registered. Only 1 in 22 of such studies register their primary outcomes before study completion. Owing to the reasons for registering studies that investigate the cause-and-effect relationship between health-related interventions and health outcomes also apply to test accuracy studies, prospective study registration of these studies should be further promoted among investigators and journal editors.

Introduction

Since September 2005, the International Committee of Medical Journal Editors (ICMJE) has required researchers to register essential information about the design of their clinical trials in a publicly available trial registry before enrolment of the first patient.[24] By facilitating transparency and completeness of reporting, this policy forms an important measure in preventing negative effects of publication bias and outcome reporting bias, respectively defined as the non-publication and selective reporting of research findings depending on the strength and direction of outcomes.[17,36] This requirement improves the evidence base on which clinical decisions are made. Furthermore, duplication of research efforts can be prevented, research and knowledge gaps can be identified, collaboration can be facilitated, and a more efficient allocation of research funds can be promoted. Full disclosure of study material may also be an ethical obligation, especially to human study participants and future patients.

The ICMJE requires registration of "any research project that prospectively assigns human subjects to intervention and comparison groups to study the cause-and-effect relationship between a medical intervention and a health outcome".[83] The reasons for registration also apply to studies quantifying the accuracy of diagnostic and prognostic tests and markers,[18] especially since failure to publish and selective reporting may also be prevalent among these studies.[19,41] Approval and proper usage of medical tests should be based on a thorough scientific evaluation.[84] Test accuracy studies form an essential part in this process. Such studies evaluate the ability of a test to correctly differentiate between patients with and without a target condition. This can be a disease (screening or diagnosis), a disease stage (staging), a condition in the near future (monitoring and surveillance), response or benefit from therapy (predictive), or an event in the future (prognosis).

At present, many clinical trial registries also include studies that do not fall under ICMJE's registration requirement. Although controversial,[85-87] increasing numbers of observational studies are also being registered.[88] This is illustrated by the fact that 19% of 156,143 records in ClinicalTrials.gov, one of the major trial registries, are tagged as observational (accessed November 27, 2013).

Increasing numbers of test accuracy studies seem to be registered as well. Although most test accuracy studies can be considered as interventional, since consenting participants are prospectively assigned to one or more medical tests, accuracy usually only contributes indirectly to changes in health outcomes. ICMJE's registration requirement, therefore, seems to exclude test accuracy studies. The Food and Drug Administration (FDA), however, requires registration of "controlled trials with health outcomes of devices subject to FDA regulation, other than small feasibility studies."[89] This may imply that studies that indirectly contribute to health outcomes, such as test accuracy studies, should also be registered.

The primary aim of this study was to identify the proportion of articles reporting on test accuracy studies for which the corresponding study had been registered, to evaluate whether registration had preceded study initiation, and to assess whether the registered record included the published primary outcome measures.

Methods

Literature search and study selection

A sample of test accuracy studies was identified by searching PubMed. We searched for studies that were published in May and June 2012 in journals with an impact factor of 5 or higher. A previously validated search filter for test accuracy studies ("sensitivity AND specificity"[MH] OR specificit*[TW] OR "false negative"[TW] OR accuracy[TW])[90] was combined with a list of names and corresponding International Standard Serial Numbers (ISSN) of all the 536 journals that had been assigned an impact factor of 5 or higher in 2011. We applied this cut-off value because we expected the number of registered studies to be larger in higher impact journals. This impact factor cut-off is in line with previously published analyses of test accuracy studies.[49,91] The final search was performed on February 25, 2013.

Articles were included if they reported on studies evaluating the accuracy of one or more tests or markers against a clinical reference standard in human subjects. Tests for screening, diagnosis, staging, monitoring, prediction, or prognosis were all eligible. We limited our search to papers published in English that had an abstract. We excluded studies that did not report an accuracy measure (sensitivity, specificity, likelihood ratio, positive or negative predictive value, diagnostic odds ratio, area under the receiver operating characteristic curve, or c-index), as well as commentaries, discussion articles, and systematic reviews.

One author (D.A.K.) scanned the search results to identify potentially eligible articles. Studies that did not provide an accuracy measure in their abstract, but were deemed likely to publish one in their full-text, were also tagged as potentially eligible. The full-text was then obtained to evaluate whether the study met the inclusion criteria. Two authors (D.A.K., and P.M.B. or L.H.) independently evaluated the potentially eligible articles. Disagreements were resolved through discussion.

Data extraction

Included studies were classified as diagnostic studies, which evaluated the ability of a test to identify a current ((pre-)stage of) disease, or prognostic studies, which used a follow-up period to evaluate the ability of a test to predict a future state or event. Based on the test under investigation, included studies were tagged as imaging studies, laboratory studies, or other. Laboratory studies included all measurements on body fluids or tissues, except for histology and cytology (which were classified as 'other'). We extracted the funding sources from the full publication. Studies that clearly described a source of support were categorized into those reporting some form of industry involvement and those reporting sources of funding not including an industrial party. Studies that did not report a source of support, or only indicated that 'no external funding' was obtained, were categorized as 'no (external) funding reported'.

Identifying registration

The following steps were taken to find out if a study had been registered. First, the full-text of the included articles was checked for a trial registration number. When this number was not reported, the corresponding author was asked through email whether the study had been registered and, if so, in which registry and under which registration number. Contact attempts were limited to three emails, each sent in a week's gap. If no answer was received, the WHO Search Portal, which searches several registries, was used. In addition, we searched ClinicalTrials.gov, the International Standard Randomized Controlled Trial Number Register, and national trial registers of the country of the first author. In these registries, we searched for the names of first, last, and corresponding authors, publication title, evaluated tests, and target disease/outcome. We matched registered records with publications by comparing the data on study design, sample size, country, outcomes, and contact information. If no registration number was found, a study was considered as not registered.

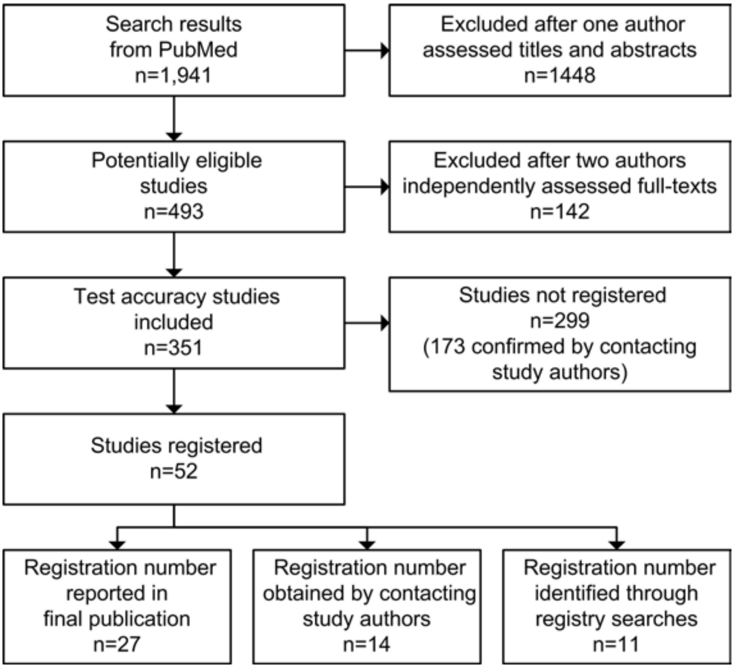
When a paper included in our review was a secondary (post hoc) analysis, we also considered the study as registered if we were able to identify a registered record for the initial study, in which the data had been collected. We categorized studies as those where the data collection had and had not been registered. We further classified studies with a registered data collection as those that had registered the published primary outcomes, those that had registered the published primary aim but vaguer or slightly different, and those that had not registered the primary outcomes or aims.

We checked whether the study had been registered before its initiation by comparing the registration date with the start and completion dates of participant enrolment as reported in the registry. Registration was defined as before initiation if the date of registration fell in or preceded the month of the study's start date as reported in the registry. A study was considered as registered after completion if it had been registered in the same month or after the registered completion date. All other studies were considered as registered in-between initiation and completion.

Statistical analysis

Data are reported as frequencies and percentages. We used χ^2 tests to evaluate associations between study characteristics and registration. Data were analyzed using SPSS version 20 (IBM, Armonk, NY, USA).

Figure 1. Flowchart for selection of studies.


Search results from PubMed n=1,941
Excluded after one author assessed titles and abstracts n=1448
Potentially eligible studies n=493
Excluded after two authors independently assessed full-texts n=142
Test accuracy studies included n=351
Studies not registered n=299 (173 confirmed by contacting study authors)
Studies registered n=52
Registration number reported in final publication n=27
Registration number obtained by contacting study authors n=14
Registration number identified through registry searches n=11

Results

The search identified 1,941 articles of which 351 fulfilled the inclusion criteria (Figure 1). Characteristics of included studies are summarized in Table 1. The majority of studies (71%) evaluated the accuracy of a diagnostic test, while 29%

evaluated a prognostic test. Comparable numbers of studies focused on imaging tests and tests based on a laboratory technique: 33% and 36%, respectively. The remainder focused on another type of test (24%), such as physical examination, electrocardiography (ECG), or pathology, or on (a combination of) tests that were assigned to more than one category (8%). Some form of industry involvement was reported by 19% of the included studies, while 58% reported sources of funding that did not include an industrial party. The remainder (23%) did not have or report an (external) source of funding.

The data collection had been registered in 52 of 351 studies (15%). Of these, 27 provided a registration number in the final publication. We contacted the authors of 324 studies without a registration number in their publication and 187 (58%) responded, providing another 14 registration numbers. Non-registration was confirmed by the authors of 173 studies. We searched the registries for the remaining 137 studies and identified another 11 registered records. Only four of the included studies had a randomized controlled design, and, of these, two (50%) had been registered.

Table 1. Characteristics of included studies and the distribution of registered studies among different characteristics.

	All studies n	Registered studies n
Total	351	52
Aim of the study		
Diagnostic	248 (71%)	38 (15%)
Prognostic	103 (29%)	14 (14%)
Type of test evaluated		
Imaging	114 (33%)	22 (19%)
Laboratory technique	126 (36%)	21 (17%)
Other	83 (24%)	6 (7%)
Combination of categories	28 (8%)	3 (11%)
Funding		
Industry-involvement	67 (19%)	22 (33%)
Other source of funding	203 (58%)	23 (11%)
No funding (reported)	81 (23%)	7 (9%)
Journal impact factor		
Median (range)	6.4 (5.0-53.3)	6.0 (5.1-38.3)

The ‘all studies’ column shows percentages of the total of included studies in parentheses. The ‘registered studies’ column shows percentages of the total per category in parentheses.

Of the 52 registered studies, 27% had been registered before initiation (Table 2). The other studies were registered somewhere between the start and completion date (58%), or after the completion date (15%). Only 23% of the registered studies provided a reference to the full publication in the registered record.

The proportion of registered studies for subgroups defined by study characteristics is shown in Table 1. There was no significant difference between diagnostic and prognostic test studies, or between imaging and laboratory studies. Of the studies reporting some form of industry involvement, 33% had been registered. This was significantly more often than studies reporting another source of funding (11%), and studies without a (reported) source of funding (9%; p<0.001).

Only 16 (31%) registered studies had registered the published primary outcomes before the completion date. Among another 12 (23%), the published primary aim had been registered before the completion date, but it was described more vaguely or somewhat differently than in the study report. Of the remaining studies, the published primary outcome or aim was not registered before study completion, or not registered at all. A majority in the latter group consisted of post hoc analyses, in which the authors had used data from a registered, previously completed study, and reports of substudies that were part of a larger registered project.

Table 2. Characteristics of registered studies.

	n
Total	52
Registration	
Before initiation	14 (27%)
In-between	30 (58%)
After completion	8 (15%)
Registration number reported	27 (52%)
Reference to full publication provided in registry	12 (23%)

Discussion

Using a previously validated sensitive search filter, we found that the data collection of only 15% of test accuracy studies published in journals with an impact factor of 5 or higher in May and June 2012 had been registered. Registration rates were comparable between studies of diagnostic and those of prognostic tests, and among studies of imaging tests and of laboratory tests. Studies reporting some industry involvement were registered more often than studies with other funding sources and studies without reported funding sources.

Adequate assessment of selective reporting among registered test accuracy studies proved difficult: only a quarter of the registered studies - 4% of all published studies - had been registered before initiation, and only one-third of the registered studies - 5% of all published studies - had registered the published primary outcomes before the study completion date. About half of the registered studies reported a trial registration number in the publication, and a reference to the final publication was reported by a quarter of the registered studies.

Our study has some potential limitations. We searched only for test accuracy studies published in journals with an impact factor of 5 or higher. It is possible that studies published in these journals are more likely to be registered than those published in lower impact journals, in which case 15% is an overestimation of the proportion of all registered test accuracy studies. We may have included studies initiated before 2005, when study registration was largely unknown among researchers. We were unable to exclude these because many test accuracy studies do not report their start and ending dates.[91,92] Since we only included studies published in May and June 2012, seven years after the ICMJE's registration policy was launched, we expect this number to be negligible. Although response rates to our email survey were relatively good, 42% of the study authors did not reply. We thoroughly searched several registries to identify a corresponding registration for these studies but may have missed some, especially since searching in most registries proves to be difficult, as extended search options are lacking. We included studies independent of their study design and type of data collection. We decided to do so because we wanted our study cohort to give a fair presentation of all types of test accuracy studies, and because of the inherent difficulties in categorizing test accuracy studies, due to scarce and substandard reporting. For example, many test accuracy studies do not report whether the study is prospective or retrospective.[91,92]

Why are these findings disappointing and promising at the same time? The results of our study indicate that, at this point, study registration for test accuracy studies does not provide many advantages. The number of registered studies is low, published primary outcomes are often not adequately registered, not registered in an informative way, and many registered studies are not registered before initiation. In addition, registration numbers are often not reported in the final publication, making it hard to find out if a study has been registered. References to the published study are often not reported in the registry, which does not facilitate finding out if a registered study has been published. We acknowledge that prospective registration of test accuracy studies is currently not officially required by the ICMJE. The fact that a considerable number of authors of these studies already seem to endorse the necessity of study registration is promising.

Study registration facilitates the identification of underexplored research areas, and the prevention of unnecessary duplication of research efforts and the corresponding waste of research funds. Full disclosure of all study material, including the protocol, is widely considered as an ethical obligation, especially to human study participants. Study registration also allows interested parties, such as reviewers, editors, physicians, policymakers, members of ethical committees, patients, and colleagues, to identify ongoing, unpublished, and selectively published studies. Non-publication and selective reporting jeopardize evidence-based medicine mainly through skewed literature syntheses. Unpublished research results are not easy to find and include in a systematic review, and this may lead to faulty conclusions based on an incomplete evidence base. Selective reporting may generate bias, offering a too optimistic presentation of test performance. Both are widely recognized problems, especially among randomized controlled trials. Evidence of cohorts of studies registered in ClinicalTrials.gov suggests that only between 46% and 63% gets published.[42,43] Studies with positive or favorable results are more likely to be published than those with negative or disappointing ones.[15] Although formal evidence is scarce, these phenomena are also suspected to be prevalent among test accuracy studies.[18,41]

In 2010, *Lancet* and *BMJ* announced that they would, from then on, encourage researchers to register observational studies in a manner similar to what has become a requirement for clinical trials.[93,94] This caused some disapproving reactions.[87,95] Criticism especially focused on the fact that observational studies vary widely in their design, and that prospective registration is not as useful for one type of study as it is for the other.[96] Several of these issues also apply to test accuracy studies. Study data can be collected prospectively or retrospectively, and study aims, hypotheses, and protocols can be formulated before or after the analysis of the data. Some test accuracy studies are exploratory in nature. Such studies often do not have a predefined protocol or hypothesis, and existing datasets are used to explore potentially interesting findings. The benefits of study registration are not as clear for such studies. Although non-publication and selective reporting are likely to be more prevalent among exploratory studies, it would be impossible to find out whether they weren registered before the post hoc hypothesis was formulated. The bureaucratic load of prospectively registering every post hoc analysis would be enormous and probably outweigh the benefits.

More in general, all the reasons for registering clinical trials seem to equally apply to interventional test accuracy studies, and probably also to all protocol-driven test accuracy studies with a priori defined aims, irrespective of whether data collection was prospective or retrospective. Therefore, we strongly recommend that authors of such studies register their protocol before initiation, and that journal editors start to think about expanding required registration to this type of research.

Acknowledgments

We thank René Spijker for assisting with the searches of this project.

Chapter 5

Endorsement of ICMJE's clinical trial registration policy: a survey among journal editors

Lotty Hooft
Daniël A. Korevaar
Nina Molenaar
Patrick M. Bossuyt
Rob J. Scholten

Netherlands Journal of Medicine 2014;72(7):349-55

Abstract

Background

Since 2005, the International Committee of Medical Journal Editors (ICMJE) requires researchers to prospectively register their clinical trials in a publicly accessible trial registry. The Consolidated Standards of Reporting Trials (CONSORT) statement has supported this policy since 2010. We aimed to evaluate to what extent biomedical journals have incorporated ICMJE's clinical trial registration policy into their editorial and peer review process.

Methods

We searched journals' instructions to authors and performed an internet survey among all journals publishing reports of randomized controlled trials that follow ICMJE's Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals (n=695), and/or that endorse the CONSORT statement (n=404) accessed in January 2011. Survey invitations were sent to the email addresses of the editorial offices and/or editors-in-chief of included journals in June 2011.

Results

For 757 ICMJE and/or CONSORT journals, we identified that they published RCT reports. We could assess the instructions to authors of 747 of these; 384 (51%) included a statement of requiring trial registration, and 33 (4%) recommended this. We invited 692 editorial offices for our survey; 253 (37%) responded, of which 50% indicated that trial registration was required; 18% cross-checked submitted papers against registered records to identify discrepancies; 67% would consider retrospectively registered studies for publication. Survey responses and specifications in instructions to authors were often discordant.

Conclusions

At least half of the responding journals did not adhere to ICMJE's trial registration policy. Registration should be further promoted among authors, editors, and peer reviewers.

Introduction

Clinical trials provide essential evidence on the effectiveness and safety of healthcare interventions. Unfortunately, many studies remain unpublished and results are often presented selectively in trial reports.[13] Since positive and favorable results are more likely to get published than negative and inconclusive ones,[15] the medical literature and systematic reviews are at risk of bias, with an overrepresentation of promising results and an underrepresentation of adverse effects.[37,97,98]

In response to accumulating evidence of selective publication and reporting in the biomedical literature, the International Committee of Medical Journal Editors (ICMJE) introduced a policy in 2005 that requires researchers to register their clinical trial in a publicly accessible trial registry before the enrolment of the first patient, in order to be considered for publication.[24,83] Trial registration improves access to clinical trial data, allows the easy identification of unpublished studies by clinicians, researchers, and reviewers,[42,43,99,100] and provides journal editors and peer reviewers with the opportunity to discover and prevent selective reporting of results. Since 2010, the Consolidated Standards of Reporting Trials (CONSORT) Statement, a reporting guideline for randomized trials, also recommends authors to report a trial registration number in the study report.[101,102]

Although the number of registered trials has grown explosively since 2005,[21] it is unknown how well journals currently adhere to ICMJE's registration policy and whether they consider publication of unregistered or retrospectively registered trials, cross-check submitted papers against registered data, and manage discrepancies between the two. We aimed to evaluate to what extent journals that announced to follow ICMJE's Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals (available at http://www.icmje.org/urm_main.html) and journals that endorse the CONSORT statement, have incorporated trial registration into their editorial and peer review process. For this aim we examined their instructions to authors and performed a survey distributed to the editorial offices of these journals.

Methods

Identification of journals

In January 2011, all journals following ICMJE's recommendations (ICMJE journals; member list obtained at http://icmje.org/journals.html) and/or endorsing the CONSORT statement (CONSORT journals; list of adopting journals obtained at http://www.consort-statement.org/about-consort/consort-endorsement/consortendorsers---journals/) were identified, along with their webpages, and the email addresses of their editorial offices and editors-in-chief. If the latter information was not provided, we tried to identify it through the Google search engine.

To find out whether these journals publish reports of randomized controlled trials (RCTs), one author scanned their webpages and published issues. Journals that did not publish RCTs and journals for which we were unable to obtain this information were excluded. The RCT publication status of each journal was confirmed by a second reviewer, with discrepancies being resolved through discussion. If necessary, a third party made the final decision. Included journals were subdivided into general and specialty journals.

Instructions to authors

Between January and September 2011, one author extracted data from the instructions to authors of included journals (Table 1). Here we excluded journals without a webpage and journals that only provided instructions to authors in languages other than English. All extracted data were confirmed by a second reviewer. Here, also, discrepancies were resolved through discussion, if necessary with a third party.

We assessed whether the journal made a statement about endorsement of ICMJE's or CONSORT's recommendations, and whether a link to these guidelines was provided. We categorized such links as 'webpages' (providing an internet-link to a web address containing the recommendations of either two), 'suitable references' (providing a reference to an article describing ICMJE's criteria published in or after 2004, or to an article describing CONSORT's criteria published in or after 2001), or 'obsolete references' (providing a reference to an ICMJE article published before 2004, or a CONSORT article published before 2001). In addition, we checked whether the instructions to authors contained a statement about the journal's policy regarding trial registration and, if so, whether registration was required or recommended, and whether specific trial registries were suggested.

Survey among editors

For the survey among editors, we excluded journals for which we were unable to identify an email address. Some editorial offices manage more than one journal. When the contact information of such journals overlapped, we considered these journals as a single potential survey responder.

In July 2011, included journals were invited to participate in our online survey through an email to the editorial office. When this email address was not available or not working, we sent the invitation to the journal's editor-in-chief. Two reminders were sent out, each a month apart. We used SurveyMonkey© software to collect responses, which was open until November 2011.

The survey consisted of eight multiple choice questions, some with an option to further clarify chosen answers. One question addressed the respondent's function within the journal's editorial staff; the other questions addressed the journal's policy regarding trial registration and to what extent this policy was incorporated into the editorial and peer review process (Table 2).

Statistical analysis

Data are reported as frequencies and percentages. Incomplete surveys were included in the analysis, for which all available responses were used. χ^2 tests were used to evaluate differences between ICMJE journals and CONSORT journals, between general and specialty journals, and between higher and lower impact journals. For this last analysis, we categorized journal impact factors into quartiles. When a journal had no impact factor, it was categorized in the lowest quartile. When a single person responded on behalf of several journals, we took the average of the impact factors for these journals.

P-values ≤0.05 were considered statistically significant. Data were analyzed using SPSS version 22 (IBM, Armonk, NY, USA).

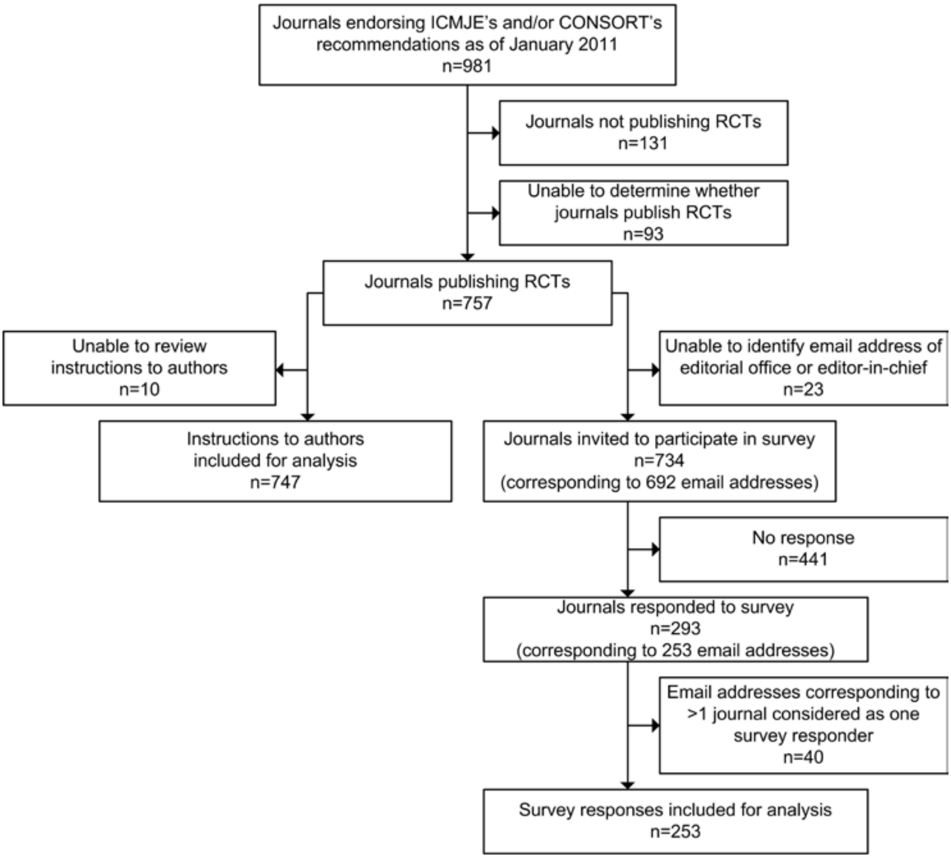
Results

In January 2011, there were 695 ICMJE journals and 404 CONSORT journals. Of these, 118 journals were on both lists. We excluded 224 journals because they did not publish RCTs (n=131), or because we were unable to obtain this information (n=93) (Figure 1). The final study sample consisted of 757 journals: 69 (9%) were general journals, and 688 (91%) were specialty journals.

Results from examination of instructions to authors

Since we were unable to assess the instructions to authors of 10 journals, due to language restrictions (n=6) or because a website was lacking (n=4), we included 747 journals in this analysis (Figure 1). Data extracted from the instructions to authors are provided in Table 1. Of the ICMJE journals, 345 (73%) made a statement about following ICMJE's recommendations. Of these, 291 provided a link to ICMJE's webpage, 15 provided a 'suitable reference' (published after 2004) containing ICMJE's recommendations, and 26 provided an 'obsolete reference'. Of the CONSORT journals, 313 (82%) made a statement about endorsement of the CONSORT statement. Of these, 280 provided a link to CONSORT's webpage, eight provided a 'suitable reference' (published after 2001) containing CONSORT's recommendations, and 10 provided an 'obsolete reference'.

Figure 1. Flowchart of ICMJE and CONSORT journals through the study.


Journals endorsing ICMJE's and/or CONSORT's recommendations as of January 2011 n=981
Journals not publishing RCTs n=131
Unable to determine whether journals publish RCTs n=93
Journals publishing RCTs n=757
Unable to review instructions to authors n=10
Unable to identify email address of editorial office or editor-in-chief n=23
Instructions to authors included for analysis n=747
Journals invited to participate in survey n=734 (corresponding to 692 email addresses)
No response n=441
Journals responded to survey n=293 (corresponding to 253 email addresses)
Email addresses corresponding to >1 journal considered as one survey responder n=40
Survey responses included for analysis n=253

ICMJE journals that had not adopted CONSORT stated significantly less often on their webpage that they required trial registration (37%) than ICMJE journals that had adopted CONSORT (60%), and than non-ICMJE journals that had adopted CONSORT (67%; p<0.0001). No significant difference was found between the proportion of general journals mentioning that trial registration was required (42%), compared with specialty journals (52%; p=0.12).

Specific trial registries that were recommended by journals making a statement about requiring or recommending trial registration were most often ClinicalTrials.gov (n=116), International Standard Randomized Controlled Trial Number register (n=81), the Australian New Zealand Clinical Trial Register (n=59), or the Netherlands Trial Register (n=55).

Results from survey

We were unable to identify an email address of the editorial office and/or editor-in-chief for 23 of the 757 included journals (Figure 1). Some email addresses corresponded to two journals (n=2), three journals (n=1), or 39 journals (n=1). We sent the invitation to 692 email addresses and between June and November 2011, 253 (37%) of these responded, including 51 partially completed surveys.

Table 1. Information provided in the instructions to authors of ICMJE and CONSORT journals.

	All Journals	Journals on ICMJE-list only	Journals on CONSORT-list only	Journals on both lists
	n	n	n	n
Total	747	366	271	110
Statement about following ICMJE's recommendations	542 (73%)	253 (69%)	197 (73%)	92 (84%)
Statement about following CONSORT's recommendations	408 (55%)	95 (26%)	230 (85%)	83 (76%)
Statement about policy regarding trial registration	417 (56%)	153 (42%)	191 (71%)	73 (66%)
Trial registration: required	384 (51%)	137 (37%)	181 (67%)	66 (60%)
Trial registration: recommended	33 (4%)	16 (4%)	10 (4%)	7 (6%)
Reference to specific trial registry provided	261 (35%)	62 (17%)	149 (55%)	50 (46%)

The following persons participated in the survey: 140 (55%) editors-in-chief, 52 (21%) managing editors, 24 (10%) editors or associate editors, 18 (7%) administrators, and 19 (8%) other types of employees. We found no evidence of selective response: 35% of the journals that made no notification on trial registration in their instructions to authors responded to the survey, compared with 38% of the journals that required registration, and 40% of the journals that recommended registration, but this difference was not significant (p=0.67).

Answers to specific questions are provided in Table 2. Only 50% (95%CI 45% to 56%) of the respondents indicated that their journal required trial registration. Significantly more journals with an impact factor in the upper quartile (above 3.5) required registration (76%) than those in the lower three quartiles (42%, 38%, and 46%, respectively; p<0.0001). There were no significant differences in trial registration requirement between ICMJE journals, CONSORT journals, and journals that had adopted both (50%, 44%, and 65%, respectively; p=0.14), nor between general and specialty journals (55% and 50%, respectively; p=0.60). Less than one-fifth of the respondents, and 22% of the journals requiring trial registration, cross-checked the reported data in the manuscript against the registered data. Journals that cross-checked the data did not always act in case of discrepancies.

Two-thirds of all the responding journals, and 56% of the journals that indicated to require trial registration, also considered study reports for publication when the underlying trial was registered after enrolment of the first patient.

Discrepancies between instructions to authors and survey responses

Journals' trial registration policies as indicated in the survey and specifications in the instructions to authors were often not concordant (Table 3). For a quarter of the journals that responded that trial registration was required, we were unable to find a corresponding statement on registration in the instructions to authors. We were also unable to find a statement on trial registration in the instructions to authors of 25% of the journals that indicated that such a statement was available. In contrast, we found a statement on trial registration for 28% of the journals that had responded that such a policy was not included in their instructions to authors. Such discrepancies were found in 37% of the journals with an impact factor in the lowest quartile, compared with 29%, 20%, and 19%, respectively, in those in the higher three quartiles (p=0.11).

Discussion

Although the ICMJE has required prospective trial registration since 2005 and CONSORT has recommended the reporting of registration numbers since 2010, at least half of the journals following ICMJE's Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals and/or endorsing the CONSORT statement do not adhere to this registration policy.

Only half of the journals responding to our survey indicated that they required trial registration. Two-thirds considered trials for publication that were registered after study initiation, against the ICMJE recommendation about prospective registration. These findings are in line with the results of previous studies, which have shown that about half of the currently published RCTs are registered after study completion, or are not registered at all.[44,47,48,103,104]

Table 2. Summary of survey responses among ICMJE and CONSORT journals.

	All responding journals **n**	**Journals on ICMJE-list only** **n**	**Journals on CONSORT-list only** **n**	**Journals on both lists** **n**
Question: What is your journal's policy regarding registration of clinical trials?				
Total respondents	232	119	79	34
Registration required	117 (50%)	60 (50%)	35 (44%)	22 (65%)
Registration recommended	57 (25%)	26 (22%)	24 (30%)	7 (21%)
Not (yet) implemented	58 (25%)	33 (28%)	20 (25%)	5 (15%)
Question: What is your journal's policy regarding registration of observational studies?				
Total respondents	232	119	79	34
Registration required	19 (8%)	13 (11%)	4 (5%)	2 (6%)
Registration recommended	76 (33%)	37 (31%)	21 (27%)	18 (53%)
Registration not necessary	137 (59%)	69 (58%)	54 (68%)	14 (41%)
Question: Is the ICMJE's clinical trial registration policy included in your journal's 'Instructions to Authors' section?				
Total respondents	226	115	77	34
Yes	142 (63%)	72 (63%)	44 (57%)	26 (77%)
No	84 (37%)	43 (37%)	33 (43%)	8 (24%)
Question: Is the ICMJE's clinical trial registration policy incorporated into your editorial and peer review processes?				
Total respondents	216	110	73	33
Yes	99 (46%)	41 (37%)	35 (48%)	23 (70%)
No	117 (54%)	69 (63%)	38 (52%)	10 (30%)
Question: For submitted manuscripts, does your journal cross-check the reported data in the manuscript against the prospectively registered data?				
Total respondents	206	103	70	33
Yes	37 (18%)	16 (16%)	12 (17%)	9 (27%)
No	169 (82%)	87 (85%)	58 (83%)	24 (73%)
Question: What do you do when discrepancies are found between the reported data in the manuscript and the prospectively registered data?				
Total respondents[a]	34	16	9	9
We do not act on that	5 (15%)	2 (13%)	1 (11%)	2 (22%)
Discrepancies are resolved between authors and editors	29 (85%)	14 (88%)	8 (89%)	7 (78%)
Question: Does your journal consider manuscripts for publication when the underlying trial has been registered after enrolment of the first patient?				
Total respondents	202	101	69	32
Yes	103 (51%)	54 (54%)	34 (49%)	15 (47%)
Yes, under certain conditions	33 (16%)	13 (13%)	11 (16%)	9 (28%)
No	66 (33%)	34 (34%)	24 (35%)	8 (25%)

[a] Only journals that had answered "Yes" to the previous question (indicating that they cross-checked reported and registered data) were included in the analysis of this question.

Table 3. Concordance between journals' registration policies as defined in the instructions to authors and according to survey responders.

Registration policy as found in instructions to authors:	**Registration policy according to survey responder:**		
	Required (n=115)	Recommended (n=57)	Not implemented (n=57)
Required (n=118)	87 (76%)	17 (30%)	14 (25%)
Recommended (n=12)	3 (3%)	7 (12%)	2 (4%)
No notification on registration policy (n=99)	25 (22%)	33 (58%)	41 (72%)

Four-fifths of the responding journals in our analysis did not cross-check submitted papers against registered records, even when requiring trial registration. This provides authors with the opportunity to publish their results selectively. A number of studies have shown that this happens frequently. Discrepancies between registered and published outcomes have been found in up to half of published trial reports.[42,47,48,53,103] A survey among peer reviewers showed that only one-third of them compared submitted manuscripts with registered trial information and reported any discrepancies to the journal editors.[105] These results indicate that it is still fairly easy for authors to get around the ICMJE's trial registration requirement and to publish unregistered and improperly registered studies.

We found that half of the journals indicated in their instructions to authors that trial registration was required. Another recent evaluation scrutinized the instructions to authors for a random selection of 200 biomedical journals publishing clinical trial reports. The authors concluded, based on information on journals' webpages, that only 28% required registration.[65]

In our study, journals' registration policies were frequently absent from webpages, and information provided in the survey sometimes differed from the instructions to authors. It seems that survey responders were not always aware of the content of the instructions to authors of their own journals; this applied to a quarter of the journals indicating that they required trial registration, and to a quarter of the journals without a registration policy. Citations referring to ICMJE's or CONSORT's recommendations were often lacking or obsolete in adopting journals. Similar deficiencies in instructions to authors have been found in previous studies. An evaluation of author guidelines of 167 medical journals in 2003 showed that a quarter of those mentioning CONSORT, and more than half of those mentioning ICMJE, provided obsolete references.[106] In another analysis, a survey was sent to journal editors about endorsement of the CONSORT statement. The study authors observed that a positive response about mentioning CONSORT in instructions to authors could not be confirmed in a quarter of cases.[107]

In 2010, *Lancet* and *BMJ* both published a statement in which they indicated that, from then on, they would strongly recommend authors to also register observational research.[93,94] Although this policy led to some controversy in the biomedical literature,[86,87] our survey indicates that more than a quarter of the ICMJE and/or CONSORT journals currently recommend registration of observational research, and a minority even requires it.

A number of elements in our analysis deserve consideration. The response rate to our survey was only 37%, and we cannot exclude selective participation. Although response rates did not significantly differ between journals that indicated in their instructions to authors that trial registration was required and those that did not, it is conceivable that journals without an active implementation of ICMJE's registration policy felt less motivated to participate. If this is the case, we may have even overestimated adherence to ICMJE's policy. We had to exclude 93 journals because we were uncertain whether they published RCT reports, mostly due to language restrictions. Data extraction, performed by a single author, was confirmed by a second one, but we may have missed information regarding registration policies in instructions to authors.

Our study was performed six years after ICMJE's trial registration policy was introduced, which should have given journals enough time to incorporate the policy into their instructions to authors, and into their editorial and peer review process. Our survey did not address reasons for not complying with ICMJE's policy. Future studies should focus on the question why many ICMJE and CONSORT journals currently do not follow these requirements, and which steps should be taken before they are willing to apply them into their editorial and peer review process. This way, barriers can be identified and potential solutions can be developed.

Selective reporting and non-publication of research findings lead to a waste of valuable research efforts and compromise the reliability of the biomedical literature.[9] There have been many examples in which the effectiveness of healthcare interventions was overestimated when solely based on published results. How can we expect medical practitioners to adequately perform evidence-based medicine when the published literature is strongly biased by positive findings? We observe a tendency towards more transparency in health research, and initiatives such as CONSORT and ICMJE's trial registration policy represent important examples. These initiatives have led to undisputable improvements: the quality of reporting has visibly increased,[108] and the number of registered trials and national trial registries has grown substantially over the past decade. Unfortunately, adoption tends to go slowly. There is still a long way to go before the scientific community can fully profit from the potential benefits of trial registration. Journal editors and peer reviewers - especially those supporting ICMJE's and/or CONSORT's recommendations - should be further encouraged to require prospective registration from each clinical trial that is presented to or reported in their journal.

Acknowledgments

We thank Sanne van der Haar for her contribution to the data extraction. We are also very grateful to all those who responded to the online survey.

Part C

Informative reporting of study reports

Chapter 6

Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD

Daniël A. Korevaar
W. Annefloor van Enst
René Spijker
Patrick M. Bossuyt
Lotty Hooft

Evidence-Based Medicine 2014;19(2):47-54

Appendices and supplemental material are available online at:
http://ebm.bmj.com/content/19/2.toc

Abstract

Background

Poor reporting of diagnostic accuracy studies impedes an objective appraisal of the clinical performance of diagnostic tests. The Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement, first published in 2003, aims to improve the reporting quality of such studies. Our objective was to investigate to which extent published artlcles of diagnostic accuracy studies adhere to the 25-item STARD checklist, whether the reporting quality has improved after STARD's launch, and whether there are any factors associated with adherence.

Methods

We performed a systematic review and searched Medline, Embase, and the Methodology Register of the Cochrane Library for studies that primarily aimed to examine the reporting quality of articles of diagnostic accuracy studies in humans by evaluating adherence to STARD. Study selection was performed in duplicate; data were extracted by one author and verified by the second author.

Results

We included 16 studies, analyzing 1,496 articles in total. Three studies investigated adherence in a general sample of articles of diagnostic accuracy studies; the others did so in a specific field of research. The overall mean number of items reported varied from 9.1 to 14.3 between 13 evaluations that evaluated all 25 STARD items. Six studies quantitatively compared post-STARD with pre-STARD articles. Combining these results in a random-effects meta-analysis revealed a modest but significant increase in adherence after STARD's introduction (mean difference 1.41 items (95%CI 0.65 to 2.18)).

Conclusions

The reporting quality of articles of diagnostic accuracy studies was consistently moderate, at least through halfway the 2000s. Our results suggest a small improvement in the years after the introduction of STARD. Adherence to STARD should be further promoted among researchers, editors, and peer reviewers.

Introduction

In 2003, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement was published in 13 biomedical journals.[29,30] Diagnostic accuracy studies provide estimates of a test's ability to discriminate between patients with and without a predefined condition, by comparing the test results against a clinical reference standard. The STARD initiative was developed in response to accumulating evidence of poor methodological quality and poor reporting among test accuracy studies in the prior years.[109,110] The STARD checklist contains 25 items which invite authors and reviewers to verify that critical information about the study is included in the study report. In addition, a flow chart that specifies the number of included and excluded patients and characterizes the flow of participants through the study is strongly recommended. Since its launch, the STARD checklist has been adopted by over 200 biomedical journals.

Over the past 20 years, reporting guidelines have been developed and evaluated in many different fields of research. Although a modest increase in reporting quality is sometimes noticed in the years following the introduction of such guidelines,[111,112] improvements in adherence tend to be slow.[108] This makes it difficult to make statements about the impact of such guidelines. For STARD, there has been some controversy around its effect.[113] While one study noticed a small increase in reporting quality of articles of diagnostic accuracy studies shortly after the introduction of STARD,[91] another study could not confirm this.[92]

Systematic reviews can provide more precise and more generalizable estimates of effect. A recently published systematic review evaluated adherence to several reporting guidelines in different fields of research, but STARD was not among the evaluated guidelines.[114] To fill this gap, we systematically reviewed all the studies that aimed to investigate diagnostic accuracy studies' adherence to the STARD checklist in any research field.

Our main objective was to find out how articles of diagnostic accuracy studies adhere to (specific items on) the STARD checklist. Our research questions were: (1) How is the current quality of reporting of diagnostic accuracy studies? (2) Has the quality of reporting improved after the introduction of STARD? (3) How do diagnostic accuracy studies score on specific items on the checklist? (4) Are there any factors associated with adherence to the checklist?

Methods

Literature search and study selection

The original protocol of this study can be obtained from the corresponding author. We performed a systematic review and searched Medline and Embase, which, to our knowledge, provide the best sources for methodological reviews. To make sure that all relevant data were captured, we also searched the Methodology Register of the Cochrane Library, of which the content is sourced from Medline and additional manual searches.

We included studies that primarily aimed to examine the quality of reporting of articles of diagnostic accuracy studies in humans in any field of research, by evaluating their adherence to the STARD statement. Details on the search strategies are provided in Web only file 1, available online. The final search was performed on August 13, 2013. The searches were performed without any restrictions for language, year of publication, or study type.

We excluded systematic reviews on the accuracy of a single test that had used the STARD checklist to score the quality of reporting in the included articles, as well as studies that investigated the influence of reporting quality on pooled estimates of test accuracy results. Such articles would be on a too specific topic to be able to make statements on the reporting quality of diagnostic accuracy studies in general. Studies focusing on reports about analytical rather than clinical performance were also excluded. Although the design of these two types of studies show many similarities, STARD was not designed for studies on analytical test performance and several items on the lists do not apply in this setting. We also excluded studies that evaluated less than 10 STARD items and studies that had not presented their results quantitatively (as a mean number of reported items or a score per individual item) because this would make an objective comparison between studies impossible.

Two authors (D.A.K. and W.A.v.E.) independently screened the titles and abstracts of the search results to identify potentially eligible studies. If at least one author identified an abstract as potentially eligible, the full-text of the article was assessed by both authors. Disagreements were resolved through discussion, whenever possible. If agreement could not be reached, the case was discussed with a third author (L.H.). One author (D.A.K.) also reviewed reference lists of included studies for additional relevant papers.

Data extraction and quality assessment

An extraction form was created before the literature search was performed, and piloted on three known eligible studies. After the pilot, the form was slightly modified. One author (D.A.K.) extracted relevant data from the included studies which were verified by the second author (W.A.v.E). Disagreements were resolved through discussion. If necessary, a third author (L.H.) made the final decision.

Of each included article, the first author, country, year of publication, and journal were extracted. We also identified the inclusion and exclusion criteria, clinical research field, primary aims, the number of studies included, and which STARD items were evaluated and how they had been scored. In addition, we retrieved (descriptive) statistics regarding overall and item-specific STARD adherence, and adherence comparisons between articles published post-STARD versus those published pre-STARD. Any additional study characteristics mentioned to be associated with STARD adherence were extracted. We also extracted any statistics on interrater agreement in evaluating STARD items, and conclusions, interpretation, and recommendations of the authors.

We assessed the quality of included studies by using the 11-item AMSTAR (Assessment of Multiple Systematic Reviews) tool.[115] As several items on this list do not apply to the studies included in our review, we omitted four items and only assessed the following elements: item 1 (was an 'a priori' design provided?), item 2 (was there duplicate study selection and data extraction?), item 3 (was a comprehensive literature search performed?), item 4 (were inclusion and exclusion criteria provided?), item 5 (was a list of included and excluded studies provided?), item 6 (were the characteristics of included studies provided?), and item 9 (was the conflict of interest included?).

Statistical analysis: overall adherence to STARD

We calculated κ statistics to assess interreviewer agreement for the two phases of study selection. For each included study, we calculated the overall STARD score, defined as the mean number of items reported by articles included in that study, and the proportion of articles adhering to each specific STARD item. For each STARD item, we calculated the median and range of these proportions.

Some studies also counted how often an item was partially reported. To be able to make comparisons between studies, we counted partially reported items as half in calculating proportions. Some STARD items pertain to the index test and the reference standard. Whenever these were analyzed separately, half a point was allocated per reported item. If a study reported that an item on the STARD checklist was not applicable to all evaluated articles, that study was not included in our overall analysis for that specific item. If a study reported that a STARD item was applied to less than 100% of the evaluated articles, the score was calculated for the number of articles for which the item applied and the calculated proportions were adjusted.

Statistical analysis: adherence to STARD before and after its launch

To obtain a summary estimate and the corresponding 95%CI of the difference in adherence before and after STARD's launch, we used inverse variance random-effects meta-analysis.[116] Only studies specifically reporting pre-STARD and post-STARD results were included in this analysis. We explored statistical heterogeneity using the I^2 test.[117] We performed a subgroup analysis by separately analyzing studies examining a general sample of articles of diagnostic accuracy studies, rather than those investigating adherence in a specific field of research.

One included study only reported SDs for (equally sized) subgroups of STARD-adopting and non-adopting journals.[92] We calculated their overall SD by taking the square root of the pooled variances. SDs of one other study were obtained after contacting the authors.[118]

We used inverse variance random-effects meta-analysis to calculate summary ORs and 95%CIs for item-specific adherence in the pre-STARD versus post-STARD groups. Only studies specifically reporting the proportion of evaluated articles adhering to each individual item for the pre-STARD and post-STARD groups were included in this analysis.

Results

Search and selection

Five hundred and eighteen studies were identified through the search, of which 35 were deemed potentially eligible after screening titles and abstracts (Figure 1). After studying the full-texts, we were able to include 16 studies.[91,92,118-131] Reasons for exclusion of potentially eligible studies are provided in Figure 1. No additional studies were identified through reference lists. Interreviewer agreement was substantial for the screening of titles and abstracts (κ=0.77 (95%CI 0.66 to 0.88)), and was perfect for the subsequent assessment of full-texts (κ=1.0).

Study characteristics

Characteristics of the included studies are provided in Table 1. Three studies investigated adherence to STARD in a general sample of articles of diagnostic accuracy studies, and the other 13 did so in a specific field of research. None of the included studies evaluated a recent sample of articles: one study evaluated articles published through 2010, one study through 2008, two studies through 2007, and four studies through 2006. All other studies only included articles published before 2006. Twelve studies evaluated articles published before and after STARD's launch, one study only evaluated articles published pre-STARD, and three studies only evaluated articles published post-STARD.

The number of evaluated articles varied markedly between the included studies, with a median of 55 (range 16 to 300). Most of the studies (n=13) evaluated all 25 STARD items. However, among three of these, one item was found not applicable to all included articles. The other three studies evaluated 24, 22, and 13 items of the 25 items.

κ values for overall interrater agreement on the STARD items were reported by nine studies: moderate agreement (κ=0.41 to 0.6) was achieved in one study, substantial agreement (κ=0.61 to 0.8) in six studies, and almost perfect agreement (κ=0.81 to 1.0) in two other studies.[132] An overall percentage agreement was reported by seven studies; this varied between 81% and 95%. Four studies did not report on interrater agreement.

Figure 1. Flowchart for selection of studies.

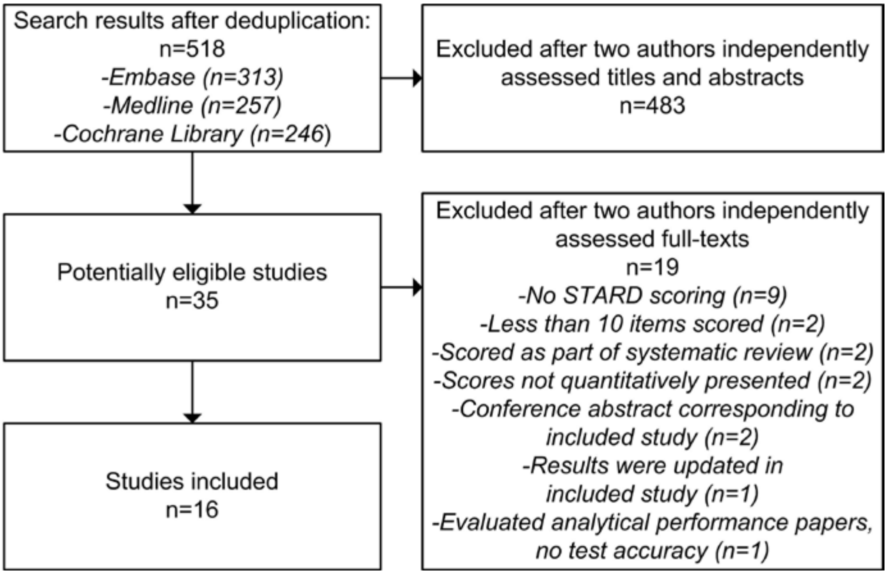

Search results after deduplication:
n=518
-Embase (n=313)
-Medline (n=257)
-Cochrane Library (n=246)
Excluded after two authors independently assessed titles and abstracts
n=483
Potentially eligible studies
n=35
Excluded after two authors independently assessed full-texts
n=19
-No STARD scoring (n=9)
-Less than 10 items scored (n=2)
-Scored as part of systematic review (n=2)
-Scores not quantitatively presented (n=2)
-Conference abstract corresponding to included study (n=2)
-Results were updated in included study (n=1)
-Evaluated analytical performance papers, no test accuracy (n=1)
Studies included
n=16

Table 1. Characteristics of included studies.

Study	Research field	Number of articles included	Timeframe	Number of STARD items evaluated	Mean STARD score (% of items evaluated)	Authors' conclusions on quality of reporting
Areia 2010 Portugal	Endoscopy	110	1998-2008	25	12.9 (52%)	"Recent publications in diagnostic endoscopy achieve only medium quality."
Coppus 2006 The Netherlands	Reproductive medicine	51	1999 versus 2004	25	12.3 (49%)	"The quality of reporting in articles on test accuracy in reproductive medicine is poor to mediocre."
Fontela 2009 Canada	Commercial tests for tuberculosis, HIV, malaria	90	2004-2006	25	13.6 (54%)	"Diagnostic studies on TB, malaria and HIV commercial tests [...] were often poorly reported."
Freeman 2009 UK	Non-invase prenatal diagnostic tests for Rhesus D genotyping	27	1996-2006	25	9.1 (36%)	"Articles have consistent weaknesses in their reporting."
Gómez Sáez 2009 Spain	Any research field, 4 Spanish journals	58	2004-2007	25	12.0 (48%)	"Despite efforts by different groups of research to achieve higher methodological quality in the diagnostics field, on average, they follow less than half of the items proposed by STARD."
Johnson 2007 UK	Optical coherence tomography (OCT) in glaucoma	30	2001-2006	25[a]	13.2 (53%)	"Quality of reporting of the diagnostic accuracy of OCT in glaucoma is suboptimal."
Lumbreras 2006 Spain	Genetic-molecular research	44	2002-2005	24	9.8 (41%)	"The articles on genetic-molecular diagnostic tests [...] fail to satisfy most of the quality requirements assembled in the STARD proposal."
Paranjothy 2007 UK	Scanning laser polarimetry (SLP) for diagnosing glaucoma	20	1997-2000 versus 2004-2005	25[a]	13.5 (54%)	"The quality of reporting of diagnostic accuracy tests for glaucoma with SLP is suboptimal."

Table 1. *Continued.*

Rama 2006 UK	Orthopedics	37	2002-2004	25	14.2 (57%)	"Current standards of reporting of diagnostic accuracy studies in orthopaedic journals are suboptimal."
Selman 2011 UK	Obstetrics and gynaecology	300	1977-2007	25	12.5 (50%)	"The reporting of included studies in this review overall was poor."
Shunmugam 2006 UK	Heidelberg retina tomography (HRT) for glaucoma detection	29	1995-2004	25[a]	14.3 (57%)	"The quality of reporting of diagnostic accuracy tests for glaucoma with HRT is suboptimal."
Siddiqui 2005 UK	Ophthalmology	16	2002	25	11.6 (47%)	"The current standards of reporting of diagnostic accuracy tests are highly variable."
Smidt 2006 The Netherlands	Six general and six disease/discipline-specific journals	265	2000 versus 2004	25	12.8 (51%)	"After publication of STARD, the quality of reporting of diagnostic accuracy studies has slightly improved. There is still room for improvement."
Wilczynski 2008 Canada	Twelve journals on radiology, internal medicine or general medicine	240	2001-2002 versus 2004-2005	13	8.2 (63%)	"We found low rates of adherence to the STARD checklist items."
Zafar 2008 UK	Diabetic retinopathy (DR) screening	76	1995-2006	25	9.9 (40%)	"The quality of diagnostic accuracy reports in DR screening is suboptimal."
Zintzaras 2012 Greece	Anti-CCP2 for the diagnosis of rheumatoid arthritis	103	2003-2010	22	14.0 (64%)	"The overall reporting quality was relatively good but needs further improvement."

[a] One of the 25 evaluated STARD items was not applicable to all the articles included in this study.

Study quality

An a priori study design was provided by only one included study. Seven studies performed the complete study selection in duplicate, while three did so in part. Eleven studies evaluated the reporting quality of all the included studies in duplicate, and three did so for a part of the included studies. All the included studies provided comprehensive data on the literature searches and the inclusion and exclusion criteria. Although more than half (n=9) of the studies provided a list of included studies, only two provided a list of excluded studies. Characteristics of included studies were provided, to some extent, by all studies; all at least provided information on the clinical research field in which included articles were performed. Only three studies gave information on the included studies' design.

Overall adherence to STARD

The overall mean STARD score varied from 9.1 to 14.3 for the 13 studies that evaluated all 25 STARD items, with a median of 12.8 items (Table 1). Fifteen (94%) of the included studies concluded that the adherence to STARD was poor, medium, suboptimal, or needed improvement. One study used more conservative language and concluded that adherence of included articles was highly variable. Seven studies evaluating all 25 items only reported post-STARD results or reported pre-STARD and post-STARD results separately. The overall mean number of items reported in these post-STARD results varied from 12.0 to 15.5, with a median of 13.6. Most of the included studies recommended the use of STARD as a guideline to improve the quality of reporting of diagnostic accuracy studies, and no study discouraged it.

The medians and ranges of the proportions of adherence to individual STARD items reported by included studies are provided in Table 2. There was a large between-study variation in adherence to specific items. Overall, only 12 items had a median proportion exceeding 50%; only three items had a median proportion above 75%. When only evaluating post-STARD results, these median proportions were slightly better: 15 items exceeding 50%, and 6 items exceeding 75%. Six items (8, 9, 10, 11, 13, and 24) concern the index test as well as the reference standard. Reporting of the index test was better than reporting of the reference standard for all of these items.

Several studies reported on factors potentially associated with quality of reporting. One study found that adherence to STARD was significantly better for cohort studies compared with case-control studies,[91] but another study could not confirm this.[127] Other factors reported to be significantly associated with higher STARD scores were sample size (higher scores among larger studies[118]) and research field (obstetric studies scored better than gynaecological studies,[118] and tuberculosis and malaria studies scored better than HIV studies[121]). Factors that did not show a significant difference were geographical area,[118] level of evidence,[127] and pooled sensitivity and specificity,[131] but these findings were not replicated in a subsequent study.

Adherence to STARD before and after its launch

Of the 12 studies that included articles published before and after the publication of STARD, six reported results for the pre-STARD and post-STARD groups separately. Combining these studies in a meta-analysis showed that significantly more items were reported post-STARD, with an estimated difference of 1.41 items (95%CI 0.65 to 2.18) (Figure 2). However, the great majority of the 383 post-STARD articles included in this analysis were published in the two years after the introduction of STARD (2004 and 2005, n=349); only 34 articles were published after 2005. As expected, I^2 test showed evidence of substantial statistical heterogeneity (66%). Subgroup analysis of the two studies that reported on a general sample of articles of diagnostic accuracy studies showed a non-significant increase in the number of reported STARD items (difference of 1.02 items (95%CI −0.08 to 2.12), I^2=80%).[91,92]

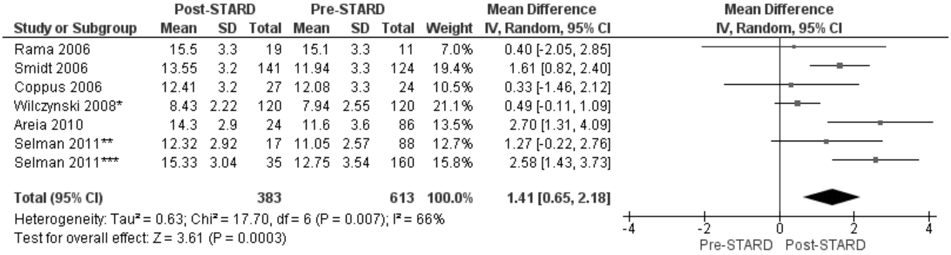
Six other studies, that were not included in the meta-analysis, reported some form of analysis of STARD adherence over time. One of these noticed an upward trend in the number of items reported pre-STARD and post-STARD.[126] Four others could not confirm this: two studies reported that the introduction of STARD did not seem to have improved the quality of reporting of articles included in their analysis,[124,125] one study observed no improvement of quality of reporting over time,[130] and one study noticed a (non-significant) decline in adherence after STARD publication.[123]

The pre-STARD versus post-STARD meta-analyses for individual items are reported in Web only file 2. Six items were significantly more often reported after the publication of STARD: item 4 (describes participant recruitment), item 5 (describes participant sampling), item 6 (describes data collection), item 14 (reports dates of study), item 15 (reports characteristics of study population), and item 23 (reports estimates of variability of accuracy). Although still rare, the number of studies reporting a flow diagram also increased significantly. None of the STARD items showed a significant decrease in frequency of reporting.

Table 2. Proportions of adherence to individual STARD items.

		Overall			Post-STARD results only	
STARD item	**Studies evaluating item**	**Median of proportions**	**Range**	**Studies evaluating item**	**Median of proportions**	**Range**
	n	**%**	**%**	**n**	**%**	**%**
25. Clinical applicability of findings	14	98%	41-100%	5	98%	84-99%
4. Participant recruitment	16	85%	55-100%	7	93%	60-98%
2. Research questions/aims	14	84%	24-100%	5	88%	76-96%
8. Technique of:	16	73%	31-98%	7	74%	40-97%
8a. Index test	5	92%	49-95%	4	84%	58-97%
8b. Reference standard	5	63%	13-86%	4	55%	23-72%
15. Characteristics of study population	16	73%	42-90%	7	70%	60-93%
7. Reference standard and rationale	16	70%	28-98%	7	76%	45-98%
9. Units/cut-offs/categories for:	16	70%	0-98%	7	83%	63-85%
9a. Index test	5	84%	68-95%	4	91%	71-94%
9b. Reference standard	5	73%	55-76%	4	75%	56-80%
3. Study population	16	68%	23-92%	7	63%	21-88%
6. Data collection	16	68%	21-100%	7	83%	43-95%
19. Cross tabulation of results	15	65%	2-99%	6	66%	28-99%
18. Distribution of severity of disease	16	62%	0-97%	7	52%	11-98%
21. Estimates of diagnostic accuracy	15	56%	12-97%	6	56%	22-97%
12. Methods for statistics used	15	49%	8-90%	6	49%	11-90%
14. Dates of study	16	47%	6-73%	7	73%	42-81%
1. Study identified as test accuracy study	13	40%	8-100%	5	24%	18-99%
5. Participant sampling	16	40%	12-89%	7	64%	31-89%
23. Estimates of variability of accuracy	15	37%	0-100%	6	39%	0-100%
17. Time interval between tests	15	34%	0-77%	6	38%	25-74%
11. Blinding of results of:	16	29%	14-54%	7	33%	16-55%
11a. Index test	5	43%	33-72%	4	50%	26-67%
11b. Reference test	5	23%	12-48%	4	25%	15-48%
22. How uninterpretable results were handled	15	28%	8-62%	6	25%	8-57%
10. Persons exectuting:	16	26%	2-73%	7	20%	2-42%
10a. Index test	5	33%	7-46%	4	26%	4-51%
10b. Reference standard	5	20%	0-35%	4	14%	0-33%
16. Eligible patients not undergoing either test	16	24%	5-78%	7	53%	13-70%
16a. Flow diagram	12	5%	0-16%	4	8%	0-22%
13. Methods for test reproducibility for:	15	16%	0-88%	6	18%	0-88%
13a. Index test	4	20%	12-53%	3	35%	6-48%
13b. Reference standard	4	7%	0-12%	3	4%	0-6%
24. Estimates of test reproducibility, for:	15	8%	0-96%	6	8%	0-96%
24a. Index test	4	20%	13-38%	3	22%	6-44%
24b. Reference standard	4	3%	0-8%	3	0%	0-6%
20. Adverse events	12	7%	0-33%	6	11%	1-18%

Figure 2. Forest plot for studies included in meta-analysis comparing adherence post-STARD and pre-STARD.



	Post-STARD			Pre-STARD				Mean Difference
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Random, 95% CI
Rama 2006	15.5	3.3	19	15.1	3.3	11	7.0%	0.40 [-2.05, 2.85]
Smidt 2006	13.55	3.2	141	11.94	3.3	124	19.4%	1.61 [0.82, 2.40]
Coppus 2006	12.41	3.2	27	12.08	3.3	24	10.5%	0.33 [-1.46, 2.12]
Wilczynski 2008*	8.43	2.22	120	7.94	2.55	120	21.1%	0.49 [-0.11, 1.09]
Areia 2010	14.3	2.9	24	11.6	3.6	86	13.5%	2.70 [1.31, 4.09]
Selman 2011**	12.32	2.92	17	11.05	2.57	88	12.7%	1.27 [-0.22, 2.76]
Selman 2011***	15.33	3.04	35	12.75	3.54	160	15.8%	2.58 [1.43, 3.73]
Total (95% CI)			**383**			**613**	**100.0%**	**1.41 [0.65, 2.18]**

Heterogeneity: Tau² = 0.63; Chi² = 17.70, df = 6 (P = 0.007); I² = 66%
Test for overall effect: Z = 3.61 (P = 0.0003)

*Wilczynski et al. evaluated only 13 STARD items;[92] the other studies evaluated 25 STARD items. **Results of the studies in obstetrics. ***Results of the studies in gynecology.

Discussion

In this systematic review, we evaluated adherence to STARD. We were able to include 16 studies, together evaluating 1,496 articles of diagnostic accuracy studies. The overall quality of reporting in these articles, published both in general and in disease-specific journals, was moderate, at least through halfway the 2000s, confirming the necessity of the introduction of STARD. Results of overall adherence were consistent among all included studies, and varied from 9.1 to 14.3 items being reported, of the 25 items on the checklist. Several factors were reported to be associated with STARD adherence by individual studies, but none of these associations was confirmed by a second study.

Although modest, there seemed to be an improvement in reporting quality (1.41 items (95%CI 0.65 to 2.18)) in the first years after STARD's publication in 2003, compared with articles published pre-STARD. Even though the confidence interval is wide, this improvement is significant. The fact that the quality of the seven analyses included in this meta-analysis was acceptable, and that all of them showed an increase in reported items (three of them significant), increases our confidence in the estimates of effect.

Our study has several potential limitations. Most of the studies evaluated articles of diagnostic accuracy studies published before 2006; none evaluated articles published after 2010. Therefore, we cannot comment on how diagnostic accuracy studies currently adhere to STARD. Most of the included studies reported a substantial interrater agreement on individual items, with marked differences between studies in reported frequencies of adherence to specific items (Table 2). There was also considerable heterogeneity in our meta-analysis comparing pre-STARD and post-STARD adherence. It is likely that this can, at least partially, be explained by between-study differences in scoring for specific items. For example, while some studies indicated that for item 3, at least the inclusion and exclusion criteria had to be reported, others only considered this item as fully reported when the setting and locations were also described. Only seven studies specifically reported how often an item was judged not to be applicable to the evaluated articles, while the others did not. Therefore, we were not always able to do a mathematical correction for non-applicable items. It is difficult to say whether between-study differences in scores of specific items were caused by a great diversity in adherence in the respective research fields, by heterogeneity in methods of scoring, or both. We would have liked to compare the differences in compliance between STARD-adopting and non-adopting journals, and between high-impact and low-impact journals, but were unable to do so because this information was almost never available in the included studies.

Although the overall quality of reporting was moderate, several items scored relatively good, with a median proportion of 70% or higher: item 2 (research questions/aims), item 4 (participant recruitment), item 7 (reference standard), item 8 (technique of index test and reference standard), item 9 (units/cut-offs/categories of tests), item 15 (study group characteristics), and item 25 (clinical applicability of findings). Worrisome is the fact that more than half of the 25 STARD items had median proportions of adherence under 50%. Especially, the reporting of study methods and results was suboptimal.

Seven items scored remarkably poor, with a median proportion of 30% or lower: item 10 (persons executing the tests), item 11 (blinding of readers), item 13 (methods for calculating test reproducibility), item 16 (the number of eligible patients not undergoing either test), item 20 (adverse events), item 22 (handling of missing results), and item 24 (estimates of test reproducibility). This is particularly alarming because several of these items can be related to biased results. If no or incomplete information on such items is reported, the potential for bias cannot be determined. Review bias, which can result when readers of a test have knowledge of the outcome of other tests or additional clinical information (item 11),[109] and verification bias, which may occur when a patient is only tested by the reference standard in case of a positive index test (item 16),[50] are likely to give inflated estimates of diagnostic accuracy. Limited test reproducibility (items 13 and 24), an effect of instrumental and/or observer variability, and not including missing responses or outliers (item 22), can also introduce biased or imprecise accuracy estimates.[30]

Interestingly, for all the six items that apply to the index test and reference standard, adherence was better for the index test. Since accuracy estimates of an index test completely depend on the reference standard, authors should be encouraged to provide all the relevant information of both tests. Also flow charts were rarely reported, both pre-STARD and post-STARD. Since these highly facilitate a reader's assessment of study design, their use should be further promoted.

Owing to a constant increase in technological and scientific innovations, the number of available diagnostic tests has been growing exponentially over the past decades. Diagnostic tests are indispensable in patient management because many clinical decisions depend on their results. Implementation and proper usage of a test in any given clinical setting should be based on a thorough consideration of its costs, safety, and clinical performance and utility. High-quality diagnostic accuracy studies are crucial in this consideration. Compared with other forms of research, diagnostic accuracy studies are probably more sensitive to bias.[109,133] The STARD checklist facilitates a complete and transparent reporting of diagnostic accuracy studies and, consequently, allows readers (e.g., clinicians, editors, reviewers, and policy makers) to identify sources of bias that may influence the clinical value and generalizability of a test. Systematic reviews of diagnostic studies often struggle with large heterogeneity across included studies; complete and transparent reporting would facilitate an identification of potential sources of heterogeneity.

Although we have presented evidence that the quality of reporting of diagnostic accuracy studies is slowly increasing, it seems that there is still significant room for improvement. A recent study showed that adherence to reporting guidelines is also suboptimal among other types of studies, such as randomized controlled trials and observational studies.[114] Although the scientific community seems to become more and more aware of the importance of transparent reporting, further enforcement of reporting guidelines among researchers, editors, and peer reviewers is a necessity.

We strongly recommend authors of articles of diagnostic accuracy studies to take STARD into account from the stage of designing the study and onwards. This way, the items can easily be incorporated in the final article. In addition, this may lead to an increased awareness among authors about potential sources of bias, which allows them to take preventive measures and, consequently, also increase the methodological quality of their study. In addition, we recommend that an evaluation of adherence to STARD should be performed on a more recent cohort of diagnostic accuracy studies. A systematic review has recently shown that, after the introduction of the CONSORT (Consolidated Standards of Reporting Trials) statement, adopting journals had a larger increase in reporting quality of randomized controlled trials than non-adopting journals.[108] Such information may be useful in the effort to convince journal editors of the necessity of adopting reporting guidelines. Future evaluations can compare reporting quality of diagnostic accuracy studies between STARD-adopting and non-adopting journals.

This way, an estimation of the impact of adopting STARD on reporting quality can be made.

Chapter 7

Reporting diagnostic accuracy studies: some improvements after 10 years of STARD

Daniël A. Korevaar
Junfeng Wang
W. Annefloor van Enst
Mariska M. Leeflang
Lotty Hooft
Nynke Smidt
Patrick M. Bossuyt

Radiology 2015;274(3):781-9

Appendices and supplemental material are available online at:
http://pubs.rsna.org/toc/radiology/274/3

Abstract

Objective

To evaluate how diagnostic accuracy study reports published in 2012 adhered to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement and whether there were any differences in reporting compared with 2000 and 2004.

Methods

PubMed was searched for studies published in 12 high-impact factor journals in 2012 that evaluated the accuracy of one or more diagnostic tests against a clinical reference standard. Two independent reviewers scored reporting completeness of each article with the 25-item STARD checklist. Mixed-effects modeling was used to analyze differences in reporting with previous evaluations from articles published in 2000 and 2004.

Results

We included 112 articles published in 2012. The overall mean number of STARD items reported in these articles was 15.3 (SD 3.9; range 6.0 to 23.5). There was an improvement of 3.4 items (95%CI 2.6 to 4.3) compared with studies published in 2000, and an improvement of 1.7 items (95%CI 0.9 to 2.5) compared with studies published in 2004. Significantly more items were reported for single-gate studies compared with multiple-gate studies (16.8 versus 12.1, respectively; p<0.001), and for studies that evaluated imaging tests compared with laboratory tests and other types of tests (17.0 versus 14.0 versus 14.5, respectively; p<0.001).

Conclusions

Completeness of reporting improved in the 10 years after the launch of STARD, but remains suboptimal for many articles. Reporting of inclusion criteria and sampling methods for recruiting patients, information about blinding of test readers, and confidence intervals for accuracy estimates are in need of further improvement.

Introduction

The Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement was first published in 2003.[29,30] The aim of STARD is to increase the transparency and completeness of reporting of diagnostic accuracy studies. The statement includes a list of 25 items that should be reported for studies to be scientifically and clinically informative to reviewers and readers.

Diagnostic accuracy studies are used to evaluate the ability of a test to identify patients with a target condition, typically a disease or a form of disease that distinguishes them from those without a target condition. These studies are prone to several types of bias.[27,28,109] Furthermore, the accuracy of a test is not a fixed property; it depends on the clinical setting, the type of patients, and on how the test is performed and interpreted. This information should be provided in the study report, and readers will be able to judge the validity and applicability of the study results when reporting is adequate.

Evaluations of the completeness of reporting for diagnostic accuracy studies in 12 high-impact factor journals in 2000 (before STARD) and 2004 (after STARD) found that, of the 25 items, an average of 1.8 additional items were reported after the introduction of STARD.[26,91] The overall completeness of reporting remained suboptimal; slightly more than half of the items were reported by studies published in 2004.

To our knowledge, it is unknown whether the initial small but statistically significant improvement in reporting quality grew over the years.[77] Our purpose was to evaluate how diagnostic accuracy study reports published in 2012 adhered to the STARD statement and whether there were any differences in reporting compared with 2000 and 2004.

Methods

Literature search and study selection

We made use of the search and selection methods developed for the evaluations of adherence to STARD among studies published in 2000 and 2004.[26,91] On September 17, 2013, we searched PubMed for diagnostic accuracy studies by using a previously validated search filter ("sensitivity AND specificity"[MH] OR specificit*[TW] OR "false negative"[TW] OR accuracy[TW]).[90] The search was limited to studies on human subjects, reported in 2012 in six general medical journals (*Annals of Internal Medicine, Archives of Internal Medicine, BMJ, JAMA, Lancet,* and *New England Journal of Medicine*) and six discipline-specific journals

(*Archives of Neurology, Clinical Chemistry, Circulation, Gut, Neurology,* and *Radiology*). All of these journals had an impact factor higher than 4 in 2000, 2004, and 2012.

We included articles if they reported in detail on a study that evaluated the diagnostic accuracy of one or more tests against a clinical reference standard in human subjects, and reported an estimate of accuracy (sensitivity, specificity, likelihood ratios, predictive values, diagnostic odds ratio, or area under the receiver operating characteristic curve). We excluded studies about the predictive and prognostic accuracy of tests, as well as reviews, letters, viewpoints, and commentaries.

Two authors (D.A.K. and W.A.v.E., with three and five of experience, respectively, in performance of systematic reviews) independently scanned titles, abstracts, and keywords of the search results to identify potentially eligible articles. In line with the previous evaluations of adherence to STARD,[26,91] we assessed only a fourth of the potentially eligible articles published in *Radiology* because of the relatively large number of diagnostic accuracy studies reported in this journal. By using a random number generator (Excel; Microsoft, Redmond, Wash), we built a random list of the potentially eligible articles from this journal and selected at least two articles from each month of the year, starting at the top of the list.

If an article was considered to be potentially eligible by at least one author, the full-text was assessed independently by both authors against the inclusion criteria. Disagreements were resolved through discussion. Whenever necessary, a third author (P.M.B.) made the final decision.

Data extraction

On the basis of the study design, we classified reports of included studies as single-gate studies (or cohort studies, with a single set of inclusion criteria for participants) or multiple-gate studies (or case-control studies, with two or more sets of inclusion criteria).[134] Depending on the index test under investigation, studies were categorized as those that evaluated imaging tests, laboratory tests, or other types of tests (e.g., physical examination). We examined the instructions to authors of the 12 included journals and categorized them as “adopters” if the use of STARD was required or recommended and as “non-adopters” if it was not.

Adherence to STARD

Between November 2013 and February 2014, we examined the extent to which included articles adhered to the 25 items on the STARD list by using a standardized score form previously developed and validated for the evaluation of studies published in 2000 and 2004.[26,91,135] For each included article, we counted the number of reported STARD items.

Six items on the STARD list concern both the index test and the reference standard: item 8 (technical specifications), item 9 (cutoffs and categories), item 10 (number and expertise of readers), item 11 (blinding), item 13 (methods for test reproducibility), and item 24 (results of test reproducibility). These items were evaluated separately for the index test and reference standard. They could be fully reported (for both index test and reference standard), halfway reported (only for index test or for reference standard), or not reported (not for index test and reference standard). If halfway reported, they were counted as one-half. We also assessed whether included articles contained a flowchart, which is strongly recommended by STARD.

Although previous studies have reported good interreviewer agreement regarding the scoring of STARD items,[77,135] the list was originally designed to guide authors, editors, and peer reviewers, not as a tool to assess completeness of reporting. Inevitably, when it is used as a tool to assess completeness of reporting, scoring of some elements is subjective. To assure high interreviewer agreement for each item, a training session was organized. Two included articles were assessed by one author (N.S.) who had also scored all the reports in the 2000 and 2004 evaluations, and by all reviewers involved in the current evaluation. These two articles were discussed in a training session until consensus on all STARD items was reached. In addition, the principal reviewer of the previous evaluations (N.S.) had several meetings with the principal reviewer of the current analysis (D.A.K.), in which they discussed the scoring of STARD items in detail and any ambiguities encountered during the scoring process.

After this, one principal reviewer (D.A.K., with one year of experience in performing literature reviews of diagnostic accuracy studies) and a second reviewer (one of the following: J.W., with one year of experience; W.A.v.E., with three years of experience; or M.M.L., L.H., or P.M.B., each with more than 10 years of experience in performing literature reviews of diagnostic accuracy studies) independently reviewed all included articles. Disagreements were resolved through discussion, but judgment from a third reviewer (P.M.B.) was decisive, if necessary. Reviewers were not blinded to author or journal.

Statistical analysis

For each article that was included, we counted the number of STARD items reported (range 0 to 25 items) and calculated an overall mean, range, and SD for the entire group. We calculated the percentage of agreement to score STARD items for the first, middle, and last article evaluated by each second reviewer (15 studies in total). For each item on the STARD list, the number and percentage of articles that reported the item were calculated.

We used Student t test statistics to compare the total number of STARD items reported between studies that were published in general medical journals and discipline-specific journals, and between single-gate and multiple-gate studies. We used one-way ANOVA to compare studies that evaluated imaging tests, laboratory tests, and other types of tests. These subgroup analyses were also performed with non-parametric test statistics by using Mann-Whitney U and Kruskal-Wallis tests.

To determine whether the reporting of individual items had improved, for each item we compared the proportion of articles that reported the item in 2012 with the corresponding proportions for 2000 and 2004. By using logistic mixed-effects modeling, we accounted for systematic differences in STARD adherence between journals. The mean number of STARD items reported in 2000, 2004, and 2012 was compared by using linear mixed-effects modeling, which again accounted for between-journal differences. We used χ^2 tests to evaluate whether features of included articles differed systematically from those in the 2000 and 2004 evaluations.

Data were analyzed using SPSS version 22 (IBM, Armonk, NY, USA). We performed mixed-effects modeling using the ‘MASS’ package in R version 3.0 (R Foundation for Statistical Computing, Vienna, Austria).

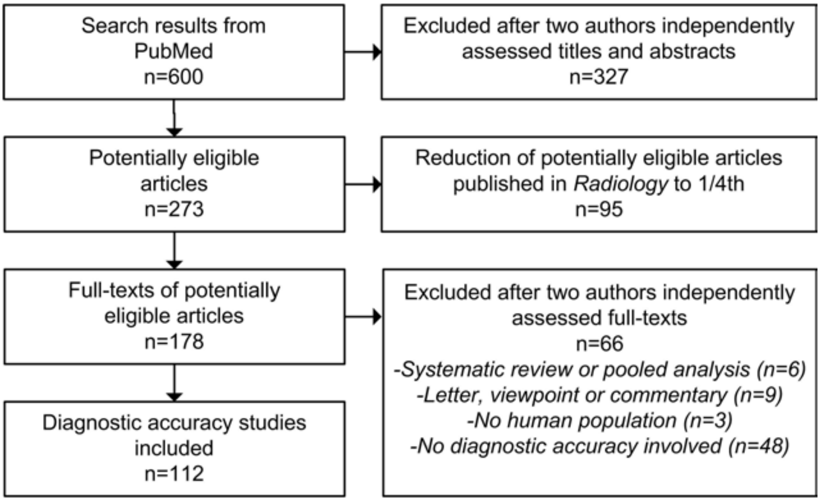
Results

Search and selection

The literature search resulted in 600 publications. On the basis of the title, abstract, and keywords, 273 articles were considered to be potentially eligible (Figure 1). As planned, we randomly excluded three-fourths (95/127) of the potentially eligible articles in *Radiology*. After examining the full-texts of the remaining 178 articles, 112 diagnostic accuracy study reports published in 2012 were considered potentially eligible. Reasons for exclusion of potentially eligible articles are provided in Figure 1. References to the included and excluded studies are available in the Appendix E1, available online.

We considered all but one of the 12 journals to be STARD adopters. Eight of the 12 journals made a clear statement that they required adherence to STARD in their instructions to authors, while three journals only provided a reference to the STARD statement. In 2004, seven journals were considered to be STARD adopters (Table 1).[91]

Figure 1. Flowchart for selection of diagnostic accuracy studies published in 2012.


Search results from PubMed n=600
Excluded after two authors independently assessed titles and abstracts n=327
Potentially eligible articles n=273
Reduction of potentially eligible articles published in *Radiology* to 1/4th n=95
Full-texts of potentially eligible articles n=178
Excluded after two authors independently assessed full-texts n=66
-Systematic review or pooled analysis (n=6)
-Letter, viewpoint or commentary (n=9)
-No human population (n=3)
-No diagnostic accuracy involved (n=48)
Diagnostic accuracy studies included n=112

Study characteristics

The number and characteristics of diagnostic accuracy studies are provided in Table 1. We found that 82.1% (92/112) of the studies were reported in discipline-specific journals versus 17.9% (20/112) in general medical journals; 68.7% (77/112) were single-gate studies and 31.2% (35/112) were multiple-gate studies. These proportions did not differ significantly from those for studies published in 2000 and 2004 (p=0.41 and 0.60, respectively). Imaging tests were evaluated in 40.2% (45/112) of included study reports, laboratory tests in 43.7% (49/112), and other types of tests in 16.1% (18/112). Seven of the 112 included articles (6.3%) explicitly referred to the STARD statement.

Item-specific adherence to STARD

The percentage agreement for scoring STARD items was 82.5% (132/160 items) for the first article evaluated by each reviewer, 88.1% (141/160) for the middle article, and 85.6% (137/160) for the final article. Adherence to individual STARD items is reported in Table 2. There were large differences between items: only one article reported on methods for calculating reproducibility of the reference standard (item 13b), for example, while all but two articles discussed the clinical applicability of the study findings (item 25), although sometimes only in a general way. For all six items that applied to both the index test and reference standard, information that concerned the index test was better reported.

Of 31 features evaluated (six of the 25 items concern both the index test and reference standard), only three were reported in less than one-quarter of the articles. These referred to methods for reproducibility of reference standard (item 13b), adverse events (item 20), and estimates of reproducibility of reference standard (item 24b).

Our analyses showed that the following features were significantly more often reported in 2012 than in 2004: study identification (item 1), study population (item 3), data collection (item 6), blinded readers of index test (item 11a), statistical methods (item 12), time interval between tests (item 17), distribution of severity of disease (item 18), and estimates of diagnostic accuracy with confidence intervals (item 21). A flowchart was reported by 35.7% (40/112) of the studies compared with only 1.6% (2/124) in 2000 and 12.1% (17/141) in 2004 (p<0.001). Compared with 2004, the following features were reported significantly less often: participant sampling (item 5), readers of index test (item 10a), and accuracy across subgroups (item 23).

Overall adherence to STARD

The mean number of STARD items reported was 15.3 (SD 3.9; range 6 to 23.5). Overall, 74.1% (83/112) of the articles reported more than half of the 25 items, while 9.8% (11/112) reported more than 20 items (Figure 2). Significantly more items were reported in studies that were published in general journals than in studies published in discipline-specific journals (17.7 versus 14.8, respectively; p=0.002), for single-gate studies compared with multiple-gate studies (16.8 versus 12.1, respectively; p<0.001), and for studies that evaluated imaging tests compared with laboratory tests and other types of tests (17 versus 14 versus 14.5, respectively; p<0.001) (Figure 3). Repeated analyses with nonparametric instead of parametric testing did not affect conclusions about significance in these three subgroup analyses (p=0.003, <0.001, and 0.001, respectively).

In 2000 and 2004, the mean number of STARD items reported was 11.9 and 13.6, respectively. There was a significant increase in completeness of reporting over the years. Articles in 2012 reported, on average, 3.4 more items (95%CI 2.6 to 4.3) than those published in 2000, and 1.7 (95%CI 0.9 to 2.5) more than those published in 2004. Only 41.1% (51/124) of the articles in 2000 reported more than half of the 25 items and none reported more than 20, compared with 61.7% (87/141) and 2.1% (3/141), respectively, in 2004 (Figure 2).

Figure 2 shows that the increase in reports of completeness of reporting was not gradual across the studies. The proportion of articles that reported less than half of the STARD items (top left-hand part of Figure 2) has barely changed between 2004 and 2012, which indicates that the lowest quarter, with the poorest reporting, has almost made no improvement at all. In the lower right-hand corner of Figure 2, the difference between 2004 and 2012 is more visible, which may indicate that the improvement between 2004 and 2012 is mainly generated by a subset of studies that is substantially more complete in their reporting.

Table 1. Characteristics of included articles by year of publication.

Study characteristics	2000[a] n	2004[a] n	2012 n
Total	124	141	112
Journal			
General medical journals	31 (25%)	30 (21%)	20 (18%)
Annals of Internal Medicine[b]	3 (2%)	6 (4%)	2 (2%)
Archives of Internal Medicine[c]	6 (5%)	4 (3%)	2 (2%)
BMJ[b]	2 (2%)	3 (2%)	8 (7%)
JAMA[b]	4 (3%)	9 (6%)	2 (2%)
Lancet[b]	9 (7%)	5 (4%)	3 (3%)
New England Journal of Medicine[d]	7 (6%)	3 (2%)	3 (3%)
Discipline-specific journals	93 (75%)	111 (79%)	92 (82%)
Archives of Neurology[c]	7 (6%)	7 (5%)	8 (7%)
Circulation[c]	13 (11%)	25 (18%)	2 (2%)
Clinical Chemistry[b]	15 (12%)	24 (17%)	20 (18%)
Gut[c]	13 (11%)	7 (5%)	11 (10%)
Neurology[b]	20 (16%)	21 (15%)	21 (19%)
Radiology[b]	25 (20%)	27 (19%)	30 (27%)
Study design			
Single-gate study	91 (73%)	96 (68%)	77 (69%)
Multiple-gate study	33 (27%)	45 (32%)	35 (31%)
Type of test			
Imaging test	-	-	45 (40%)
Laboratory test	-	-	49 (44%)
Other type of test	-	-	18 (16%)

[a]Results from 2000 and 2004 are from Smidt et al.[26,91] [b]Journal mentioned STARD in its instruction for authors in 2004 and 2012. [c]Journal mentioned STARD in its instruction for authors in 2012, but not in 2004. [d]Journal did not mention STARD in its instruction for authors in 2004 and 2012.

Table 2. Adherence to individual STARD items by year of publication.

STARD item	Number of articles published in 2000[a] n	Number of articles published in 2004[a] n	Number of articles published in 2012 n	*p-value* 2004 versus 2012[b]
Total	124	141	112	
Title/abstract				
1. Identify the article as a study of "diagnostic accuracy"	13 (11%)	26 (18%)	34 (30%)	0.03(↑)
Introduction				
2. State research questions/aims, such as estimating diagnostic accuracy	112 (90%)	136 (97%)	107 (96%)	0.68
Methods				
3. Study population: Inclusion and exclusion criteria, setting and location of data collection	35 (28%)	30 (21%)	73 (65%)	<0.001(↑)
4. Participant recruitment: Based on symptoms, results from previous tests, or the fact participants had received the index test or reference standard?	103 (83%)	130 (92%)	106 (95%)	0.43
5. Participant sampling: Was the study population a consecutive series of participants? If not, specify how participants were further selected	70 (57%)	108 (77%)	62 (55%)	<0.001(↓)
6. Data collection: Prospective or retrospective?	99 (80%)	119 (84%)	104 (93%)	<0.01(↑)
7. The reference standard and its rationale	70 (57%)	64 (45%)	59 (53%)	0.25
8. Technical specifications of materials and methods involved including how and when measurements were taken, or cite references for:				
a. Index test	115 (93%)	137 (97%)	111 (99%)	0.26
b. Reference standard	83 (67%)	101 (72%)	75 (67%)	0.64
9. Definition of, and rationale for, units, cutoffs, and/or categories of results of:				
a. Index test	103 (83%)	132 (94%)	107 (96%)	0.40
b. Reference standard	75 (61%)	102 (72%)	80 (71%)	0.98
10. Number, training, and expertise of persons executing and reading:				
a. Index test	51 (41%)	73 (52%)	51 (46%)	0.02(↓)
b. Reference standard	32 (26%)	46 (33%)	47 (42%)	0.21
11. Blinding to the results of the other test of, and any other clinical information provided to, the readers of:				
a. Index test	46 (37%)	56 (40%)	65 (58%)	0.02(↑)
b. Reference test	23 (19%)	39 (28%)	41 (37%)	0.14
12. Methods for calculating or comparing measures of diagnostic accuracy, and statistical methods used to quantify uncertainty (e.g., 95%CI)	17 (14%)	28 (20%)	51 (46%)	<0.001(↑)
13. Methods for calculating test reproducibility of:				
a. Index test	20 (16%)	49 (35%)	44 (39%)	0.41
b. Reference standard	6 (5%)	9 (6%)	1 (1%)	0.06
Results				
14. When study was performed, including beginning and end dates of recruitment	60 (48%)	89 (63%)	79 (71%)	0.57

Table 2. *Continued.*

15. Clinical and demographic characteristics of the study population (at least information on age, sex and spectrum of presenting symptoms)	65 (52%)	84 (60%)	68 (61%)	0.75
16. Number of participants satisfying inclusion criteria who did not undergo the index test or reference standard, and why they failed to undergo these	75 (61%)	83 (59%)	64 (57%)	0.77
17. Time interval between index test and reference standard, and any treatment administered in between	33 (27%)	35 (25%)	59 (53%)	<0.001(↑)
18. Distribution of severity of disease in those with the target condition, and other diagnoses in those without the target condition	28 (23%)	74 (53%)	95 (85%)	<0.001(↑)
19. Cross tabulation of results of the index test by the results of the reference standard; for continuous test results, distribution by results of the reference standard	104 (84%)	124 (88%)	92 (82%)	0.15
20. Any adverse events from performing the index test or reference standard	21 (17%)	16 (11%)	12 (11%)	0.90
21. Estimates of diagnostic accuracy and 95% confidence intervals	40 (32%)	57 (40%)	74 (66%)	<0.001(↑)
22. How intermediate results, missing data and/or outliers of tests were handled	73 (59%)	80 (57%)	77 (69%)	0.11
23. Estimates of variability of accuracy between subgroups of participants, readers, or centers	48 (39%)	84 (60%)	52 (46%)	0.04(↓)
24. Estimates of test reproducibility, for:				
a. Index test	40 (32%)	62 (44%)	50 (45%)	0.78
b. Reference standard	8 (7%)	8 (6%)	4 (4%)	0.51
Discussion				
25. Discuss the clinical applicability of the study findings	114 (92%)	138 (98%)	110 (98%)	0.85
Recommendations				
Flowchart	2 (2%)	17 (12%)	40 (36%)	<0.001(↑)

[a]Results from 2000 and 2004 are from Smidt et al.[26,91] [b]P-values obtained using mixed-effects logistic modeling, which accounts for journal-level effects, include results from 2000, 2004, and 2012 in the model. Arrows indicate the direction of significant differences.

Discussion

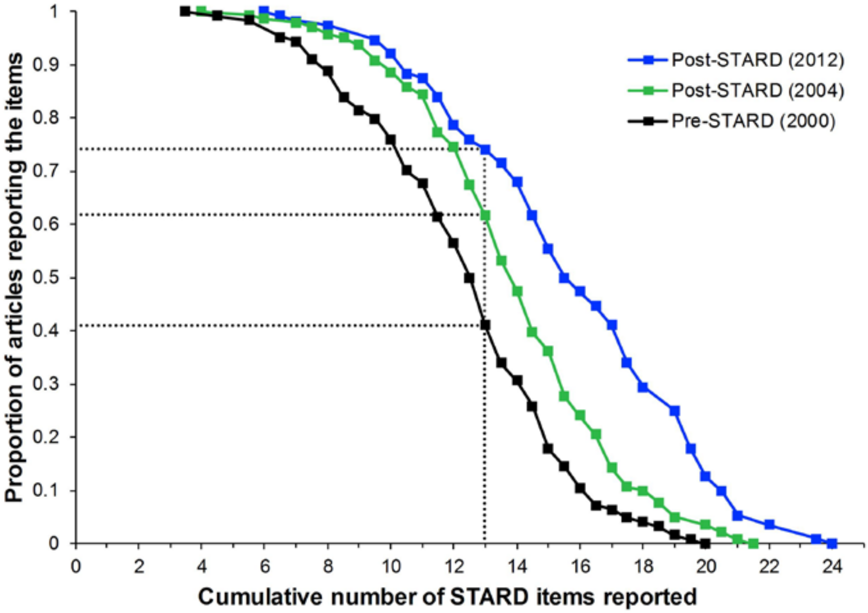
We evaluated the extent to which diagnostic accuracy study reports that were published in 12 high-impact factor journals in 2012 adhered to the STARD list and compared our findings with results from previous, comparable evaluations of articles published in 2000 and 2004.[26,91] We observed that the quality of reporting has slowly but gradually made an improvement, but that completeness of reporting and transparency remain suboptimal in many articles.

This gradual increase in reporting completeness is in line with previous analyses of adherence to STARD. A recent meta-analysis of six of these evaluations showed that studies published after the launch of STARD reported, on average, 1.4 more items.[77] All these evaluations were performed in the first few years after the publication of STARD, which may have been too early to expect large improvements. The results of our analysis indicate that the small initial improvement persisted and grew over the years, but also that it is not as large as may have been anticipated.

Over the years, the reporting of many individual STARD items improved, but there is variability, and some domains definitely need further improvement. A quarter of the evaluated articles reported less than half of the STARD items. Many articles do not adequately report on the patient eligibility criteria, recruitment process, and sampling methods. To allow judgments regarding the applicability of study results, such information is crucial.

Some items associated with bias could also benefit from more complete reporting. It is often unclear whether readers of the tests were blinded to clinical information, which prohibits assessment of the risk of review bias. Many articles do not report how many eligible patients failed to undergo the index test or reference standard, which prohibits a judgment about verification bias. The time interval between the index test and reference standard was unclear in half of the articles. Changes in severity of the target condition, or the initiation or withdrawal of medical interventions could occur between tests and influence accuracy estimates.

Figure 2. Proportion of articles that reported at least the indicated number of STARD items.



Results from 2000 and 2004 are from Smidt et al.[26,91] The dotted lines indicate the proportion of articles that reported more than half of the STARD items.

Although the number of articles that reported confidence intervals around estimates of diagnostic accuracy doubled between 2000 and 2012, it is disappointing that still about one-third failed to do so in 2012. Failure to report measures of precision around estimates of accuracy facilitates an overoptimistic, generous interpretation of study results, a phenomenon that is common in diagnostic accuracy study reports.[49,66]

As of the publication of this article, all but one of the evaluated journals adopted STARD in their instructions to authors, but adherence is suboptimal for many of them. This may indicate that authors, editors, and peer reviewers do not always recognize a diagnostic accuracy study as such, or that journals have not actively implemented the use of STARD in their editorial and peer review process. Journal editors and peer reviewers may be actively trained to identify diagnostic accuracy studies and to evaluate quality of reporting. Reporting experts could be invited to peer review study reports. Previous studies have shown that peer reviewers often fail to identify reporting deficiencies in the methods and results of randomized trials,[136] and that additional reviews on the basis of reporting guidelines increases the quality of articles.[137]

Our study has some potential limitations. We acknowledge that we may have been strict in scoring some items. For example, identification of the study (item 1) was only felt to be satisfactorily handled in a study report when the term diagnostic accuracy was included in the title or abstract. Characteristics of the study population (item 15) were only considered adequately reported when some information (other than age and sex) about presentation of symptoms was also provided. We did this to compare our results with those of analyses of articles published in 2000 and 2004. Other items, especially those with the lowest adherence rates, may not always be applicable. Adverse events (item 20), for example, are not an issue for most imaging tests, and the reproducibility of the reference standard (item 13b and 24b) is often well established. STARD was launched in 2003 and an update is underway. Although there were no major improvements, to our knowledge, of concepts of study design and sources of bias since then, some of the items on the current list may be outdated and redundant, while other relevant items may be absent.

None of the reviewers involved in this evaluation analyzed the articles published in the 2000 and 2004 analyses, but we made considerable efforts to achieve comparability with these previous evaluations. Nevertheless, it is possible that features were interpreted somewhat differently.

Figure 3. Number of STARD items reported by subgroups for articles published in 2012.

Figure 3a. Type of journal.

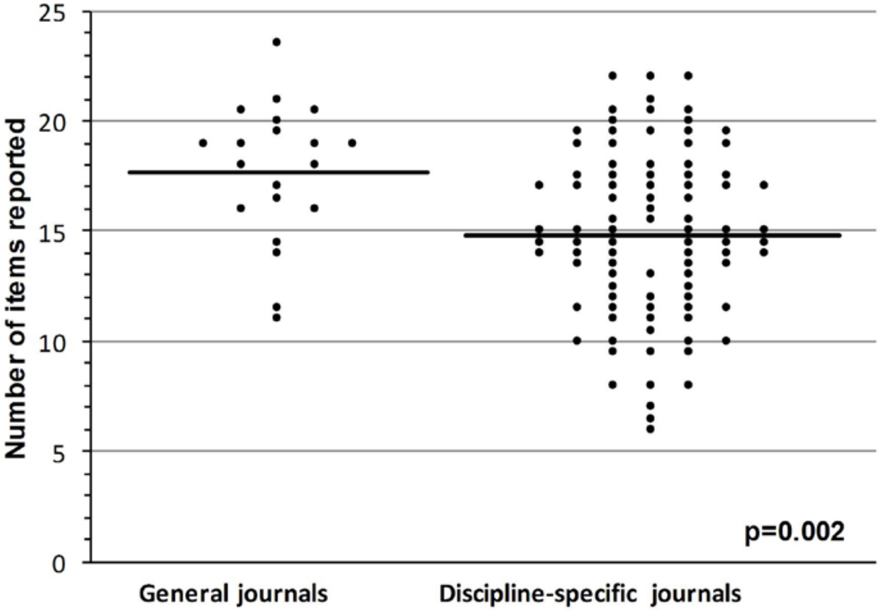


Figure 3b. Study design.

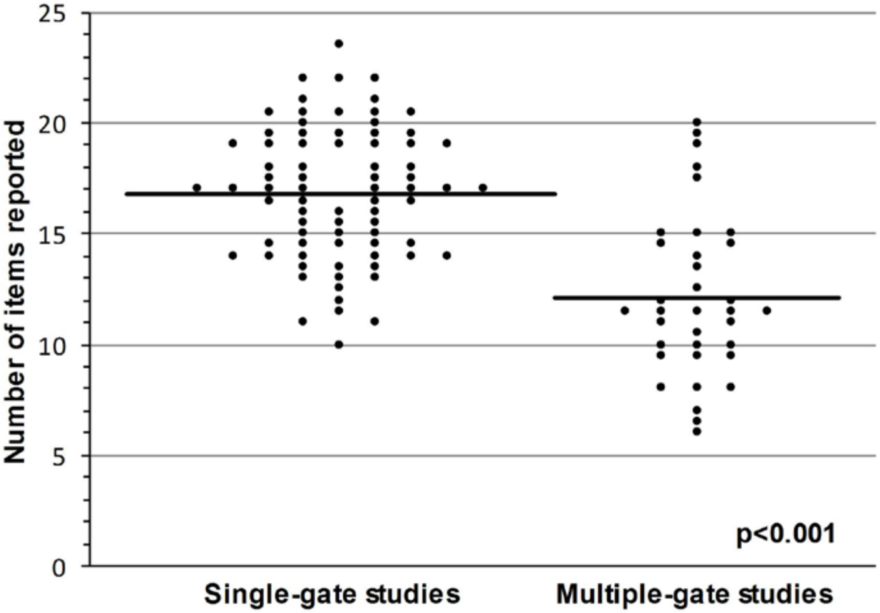
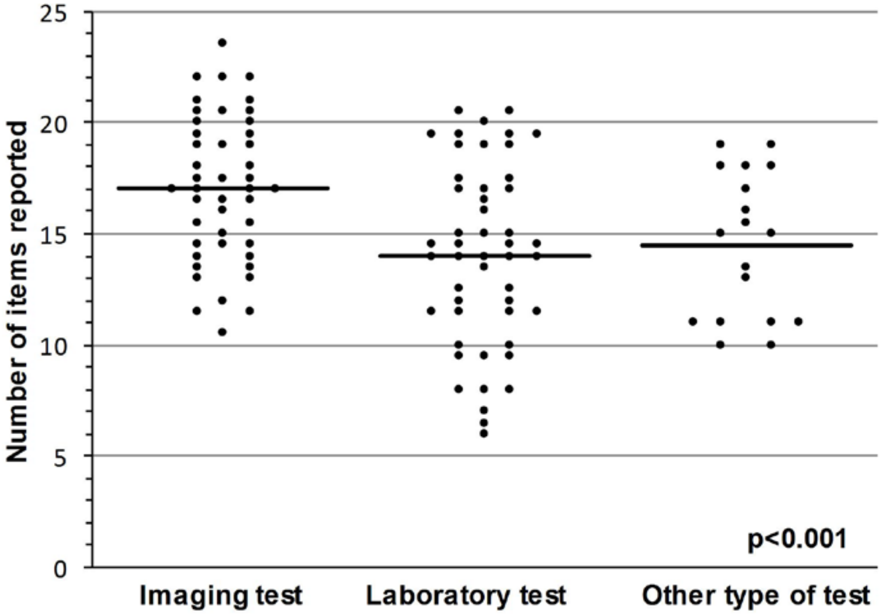


Figure 3. *Continued.*

Figure 3c. Type of test.



Each dot represents one article. The bold horizontal lines represent the mean number of items reported for each subgroup.

We only included studies that were published in journals with an impact factor above 4. In other fields of research, quality of reporting was shown to be lower in journals with lower impact factors.[114] We included studies that evaluated diagnostic accuracy, even if this was not their primary objective. We decided to do this because primary and secondary objectives are often not explicitly reported,[53] and because we believe that any estimate of test accuracy should be accompanied by sufficient information to evaluate its validity and applicability. Because only one of 12 selected journals did not explicitly adopt STARD, we were unable to analyze differences between journals that adopted these standards and journals that did not.

Medical tests are the basis for almost every clinical decision. The tests we rely on are usually not perfect, and patients with the targeted condition may have a negative test result while other patients test positive and do not have the condition. When clinicians order tests and interpret test results, they should consider the likelihood that such errors occur. In modern evidence-based medicine, this should not be on the basis of hearsay or personal experience; it should be informed by the results of diagnostic accuracy studies. However, readers will only be able to identify sources of bias and appreciate limitations regarding the applicability of study results to their own setting when the reporting is honest, transparent, and complete. We strongly encourage authors to use STARD to report their diagnostic accuracy studies, and we encourage editors and peer reviewers to stimulate, encourage, or remind authors to do so as well.

Chapter 8

Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies

Daniël A. Korevaar
Jérémie F. Cohen
Lotty Hooft
Patrick M. Bossuyt

Journal of Clinical Epidemiology 2015;68(6):708-15

Appendices and supplemental material are available online at:
http://www.jclinepi.com/issue/S0895-4356(14)X0007-7

Abstract

Background

Informative journal abstracts are crucial for the identification and initial appraisal of studies. We aimed to evaluate the informativeness of abstracts of diagnostic accuracy studies.

Methods

PubMed was searched for reports of studies that had evaluated the diagnostic accuracy of a test against a clinical reference standard, published in 12 high-impact journals in 2012. Two reviewers independently evaluated the information contained in included abstracts using 21 items deemed important based on published guidance for adequate reporting and study quality assessment.

Results

We included 103 abstracts. Crucial information on study population, setting, patient sampling, and blinding as well as confidence intervals around accuracy estimates were reported in <50% of the abstracts. The mean number of reported items per abstract was 10.1 of 21 (SD 2.2). The mean number of reported items was significantly lower for multiple-gate (case-control type) studies, in reports in specialty journals, and for studies with smaller sample sizes and lower abstract word counts. No significant differences were found between studies evaluating different types of tests.

Conclusions

Many abstracts of diagnostic accuracy study reports in high-impact journals are insufficiently informative. Developing guidelines for such abstracts could help the transparency and completeness of reporting.

Introduction

Evaluating the validity of health research is only possible when study reports are sufficiently informative.[9] In response to increasing evidence of substandard reporting of biomedical studies, collaborative initiatives have led to the development of reporting guidelines in different fields of research, such as the Consolidated Standards of Reporting Trials (CONSORT) statement for randomized controlled trials.[102]

In 2003, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement was first published.[29] Diagnostic accuracy studies evaluate how well a medical test identifies or rules out a target condition, as detected by a clinical reference standard. Study results are typically expressed in measures such as sensitivity and specificity. The STARD statement contains a checklist of 25 items that should be presented in all reports of diagnostic accuracy studies, covering key elements from study design and setting, selection of participants, execution and interpretation of tests, data analysis, and study results.

Unlike some other guidelines, such as those for reporting randomized controlled trials and systematic reviews,[138,139] STARD so far has not provided detailed guidance for writing journal abstracts. Readers, especially those in resource constrained settings where free access to full study reports is limited, might base clinical decision making on the information provided in abstracts only. Clinicians, researchers, systematic reviewers, and policy makers need to assess and critically appraise large amounts of information in short periods of time to keep up to date. Abstracts play a crucial role in this process. Initially introduced in the 1960s,[140] abstracts have especially gained importance in the past three decades because of the development of evidence-based medicine, the almost exponential increase in medical journals and publications, and the increased access to online libraries such as PubMed. To accommodate these changes, the structured abstract was introduced in 1987, and the great majority of biomedical journals has adopted it since then.[141]

Incomplete, partial, or even incorrect information in abstracts makes it difficult for readers to identify research questions, study methods, study results, and the implications of study findings. Despite undisputable improvements,[142,143] the informativeness of many abstracts of randomized trials remains suboptimal.[144-146] Whether similar deficiencies exist in reports of diagnostic accuracy studies is unknown. Two previous studies evaluated the content of abstracts of such studies but only for a small number of items and in specific fields of research.[41,147] We aimed to systematically evaluate the informativeness of abstracts of diagnostic accuracy studies published in 12 high-impact journals in 2012, by scoring whether essential methodological features and study results were reported.

Methods

Literature search and study selection

We searched PubMed using a search filter with high sensitivity for diagnostic accuracy studies ("sensitivity AND specificity"[MH] OR specificit*[TW] OR "false negative"[TW] OR accuracy[TW]).[90] We looked for study reports published in one of six general medical journals (*Annals of Internal Medicine, Archives of Internal Medicine, BMJ, JAMA, Lancet,* and *New England Journal of Medicine*) and six discipline-specific journals (*Archives of Neurology, Clinical Chemistry, Circulation, Gut, Neurology,* and *Radiology*) in 2012. These 12 journals were selected in line with previous evaluations, in which they were found to publish the largest number of diagnostic accuracy study reports among all journals with an impact factor over 4.[26,91] The median impact factor of these journals in 2012 was 12.4 (range 6.3 to 51.7). As of 2012, eight of these journals clearly stated in their instructions to authors that they require adherence to STARD, and three only provided a reference to STARD. The same set of studies has been used previously to evaluate adherence to the STARD reporting guidelines.[148]

Eligible were all articles that reported estimates of the accuracy of medical tests in humans, based on a comparison of index test results against a clinical reference standard. Two reviewers independently examined studies for inclusion; disagreements were solved through discussion. First, all titles and abstracts were screened to identify potentially eligible articles. After this, the full-text of potentially eligible articles was evaluated. In line with previous evaluations of STARD,[26,91] only a randomly selected quarter of the potentially eligible articles published in *Radiology* was evaluated for inclusion because the number of diagnostic accuracy studies published in this journal was relatively large. We prepared a random list of the potentially eligible articles from this journal and, using a random number generator in Excel, selected at least two articles from each month of the year, starting from the top of the list.

For the current evaluation, we secondarily excluded studies if they did not report or mention at least one of these measures of diagnostic accuracy in the abstract: sensitivity, specificity, likelihood ratios, predictive values, diagnostic odds ratio, accuracy, area under the receiver operating characteristic curve, or C index.

Data extraction

We extracted the first author, journal, journal type (general versus discipline-specific), study design [single-gate (cohort type) studies, which used one set of inclusion criteria, versus multiple-gate (case-control type) studies, which used multiple sets of inclusion criteria],[134] and type of test under evaluation (imaging tests versus laboratory tests versus other types of tests). We also extracted the sample size (number of participants or biological specimens) as reported in the abstract and the word count (number of words used) of each included abstract, excluding the title. Two independent reviewers extracted all data; disagreements were solved through discussion.

Informativeness of abstracts

A review team developed a list of items to evaluate the content of abstracts, mostly aiming at key elements related to study validity. The review team consisted of four researchers, all of them part of the STARD group (D.A.K., with two years of experience, J.F.C., with four years of experience, and L.H. and P.M.B., each with more than 10 years of experience in performing literature reviews of diagnostic accuracy studies). First, a longlist of 36 potentially relevant items was generated based on the STARD statement,[29,30] the CONSORT for Abstracts checklist, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for Abstracts checklist,[139] QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies),[28] existing guidance on the structured reporting and the assessment of the quality of journal abstracts in general,[149-152] and previous studies evaluating the content of abstracts of diagnostic accuracy studies (Appendix A, available online).[41,147] After this, each item on the longlist was discussed within the review team, and a subset of items deemed most relevant was selected based on general consensus. The list of items was then piloted and refined by all members of the review team based on an evaluation of 10 included abstracts.

The final list contains 21 items (Appendix B), focusing on study identification, rationale, objectives, methods for recruitment and testing, participant baseline characteristics, missing data, test results and reproducibility, estimates of diagnostic accuracy, and discussion of study findings, implications, and limitations.

Two authors independently evaluated each included abstract and scored each item as reported or not reported. We also established guidance on the interpretation of each item (Appendix B). Any discrepancies were solved through discussion. If consensus could not be reached, the case was discussed with a third author, who made the final decision.

Statistical analysis

We reported general characteristics of included studies as frequencies and percentages or as medians together with IQRs. We counted the total number of reported items for each included abstract (range 0 to 21) and then calculated an overall mean together with SD and range of the number of reported items across studies. For each item on the list, the number and percentage of abstracts reporting the information was calculated. Interreviewer agreement on the scoring of items was assessed by calculating the κ statistic, excluding the 10 abstracts that were used to pilot and refine the list of items.

We used univariate analysis with one-way ANOVA to compare the mean number of items reported between journal types, study designs, test types, and sample sizes and abstract word counts. For the latter two, we used a median split. We also adjusted a multiple-linear regression model that included variables with a $p \leq 0.10$ on univariate analysis, to explore conditional associations with the number of items reported. Statistical analyses were performed using SPSS version 20 (IBM, Armonk, NY, USA).

Results

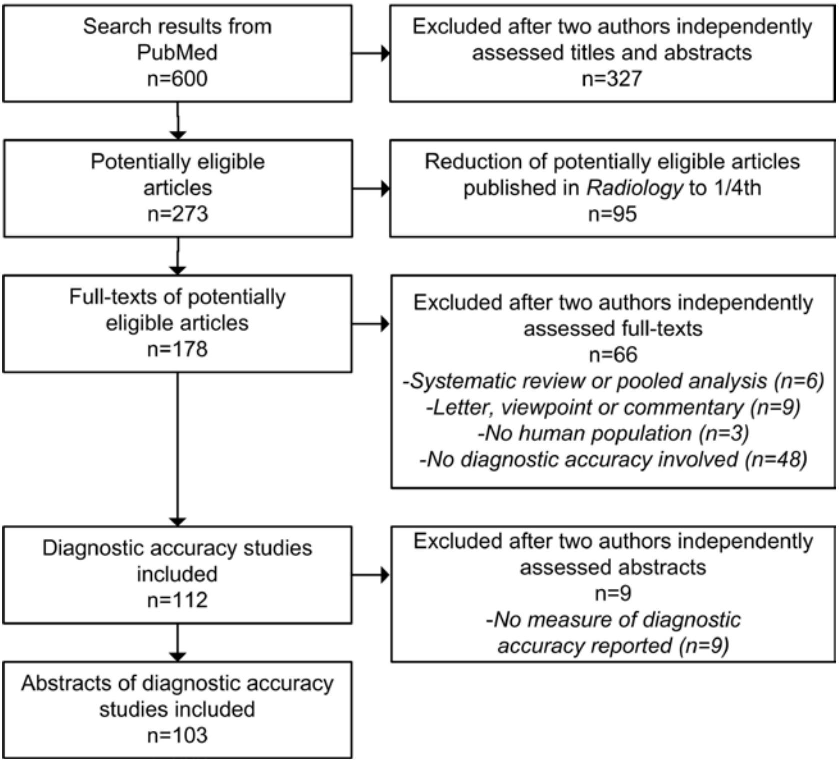
Search and selection

The literature search generated 600 records (Figure 1). Selection based on titles and abstracts resulted in a κ of 0.67 [95%CI 0.62 to 0.73]; this was 0.77 (95%CI 0.68 to 0.88) for full-text selection and 0.63 (95%CI 0.40 to 0.86) for the final abstract selection. We included 103 articles reporting on the evaluation of the diagnostic accuracy of a medical test in their abstract. Characteristics of the included studies are provided in Table 1.

Total number of items reported

The κ statistic in scoring items was 0.85 (95%CI 0.83 to 0.87). The mean number of items reported in the abstracts was 10.1 of 21 (SD 2.2; range 6 to 15). All abstracts reported more than five items on the list, 38% of the abstracts reported 11 items or more. No abstract reported more than 15 items (Figure 2).

Figure 1. Flowchart for selection of studies.


Search results from PubMed n=600
Excluded after two authors independently assessed titles and abstracts n=327
Potentially eligible articles n=273
Reduction of potentially eligible articles published in *Radiology* to 1/4th n=95
Full-texts of potentially eligible articles n=178
Excluded after two authors independently assessed full-texts n=66
-Systematic review or pooled analysis (n=6)
-Letter, viewpoint or commentary (n=9)
-No human population (n=3)
-No diagnostic accuracy involved (n=48)
Diagnostic accuracy studies included n=112
Excluded after two authors independently assessed abstracts n=9
-No measure of diagnostic accuracy reported (n=9)
Abstracts of diagnostic accuracy studies included n=103

Factors associated with number of items reported

The mean number of reported items was significantly lower in abstracts published in specialty journals (9.6; SD 2.0) compared with general journals (12.2; SD 1.9; mean difference (MD) 2.6 (95%CI 1.6 to 3.6); p<0.001), in articles reporting on multiple-gate studies (9.0; SD 2.0) compared with single-gate studies (10.6; SD 2.1; MD 1.5 (95%CI 0.7 to 2.4); p=0.001), in abstracts of studies with sample sizes below the median (9.4; SD 1.9) compared with those above (10.8; SD 2.3; MD 1.4 (95%CI 0.6 to 2.3); p=0.001), and in abstracts with a word count below the median (9.5; SD 2.3) compared with those above (10.6; SD 2.0; MD 1.1 (95%CI 0.3 to 2.0); p=0.008) (Figure 3). The number of items did not significantly differ according to the type of test under evaluation: 10.6 (SD 2.2) for imaging tests, 9.6 (SD 2.2) for laboratory tests, and 10.1 (SD 2.1) for other tests (p=0.13).

In multiple-linear regression, the type of journal (adjusted mean difference (AMD) 1.9 (95%CI 0.8 to 3.0); p=0.001), study design (AMD 1.0 (95%CI 0.1 to 1.8); p=0.03), and sample size (AMD 0.9 (95%CI 0.1 to 1.7); p=0.02) were significantly associated with the number of items reported, whereas word count was not (AMD 0.5 (95%CI -0.3 to 1.3); p=0.22).

Table 1. Characteristics of included diagnostic accuracy studies.

	n
Total	103
Journal	
General medical journals	17 (17%)
Annals of Internal Medicine	2 (2%)
Archives of Internal Medicine	1 (1%)
BMJ	8 (8%)
JAMA	2 (2%)
Lancet	2 (2%)
New England Journal of Medicine	2 (2%)
Discipline-specific journals	86 (84%)
Neurology	17 (17%)
Archives of Neurology	7 (7%)
Circulation	2 (2%)
Clinical Chemistry	20 (19%)
Gut	11 (11%)
Radiology	29 (28%)
Study design	
Single-gate	71 (69%)
Multiple-gate	32 (31%)
Type of test	
Imaging	43 (42%)
Laboratory	47 (46%)
Other	13 (13%)
Sample size[a]	
Median (IQR)	164 (77.5-471.5)
Word count[b]	
Median (IQR)	269 (248-300)

[a]Unclear for 2 of 103 abstracts. [b]Number of words used in the abstract, excluding the title.

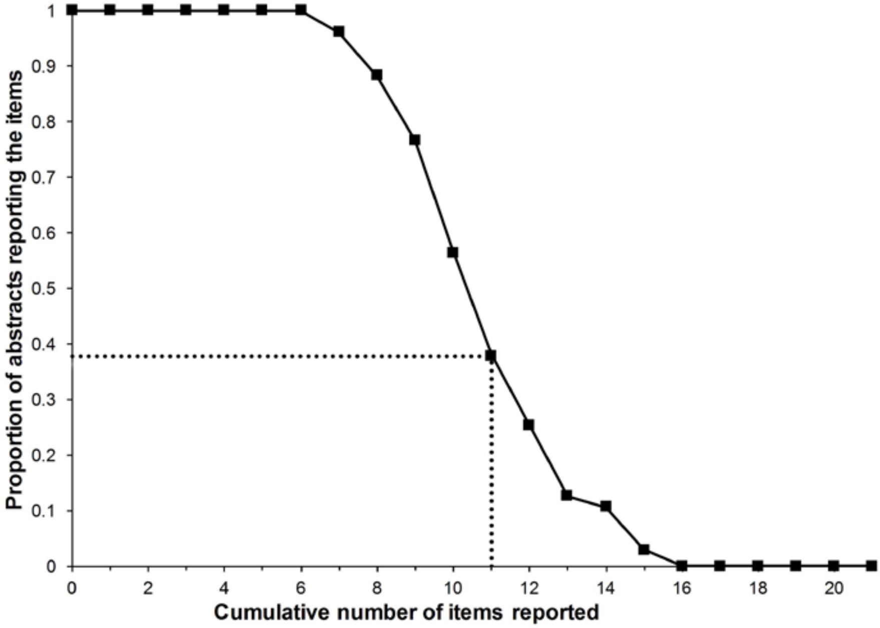
Item-specific reporting

The reporting of individual items on the list was highly variable (Table 2). Twelve of the 21 items were reported in less than half of the evaluated abstracts; only five items were reported in more than three-quarters of the abstracts.

Title, background, and aims

Fifty percent of the abstracts announced the evaluation of a diagnostic test in the title, and 46% provided a rationale for this evaluation in the abstract's introduction. Research objectives, aims, or questions were lacking in 15% of the abstracts.

Figure 2. Proportion of journal abstracts of diagnostic accuracy studies that reported at least the indicated number of items on the 21-item list.



The dotted line indicates the proportion of abstracts that reported more than half of the evaluated items.

Methods

There was large variability in reporting for various aspects of the study methods. Key items that should inform the reader about which participants were eligible and how, where, and when they were recruited were rarely reported: the inclusion criteria (15%), study setting (32%), number of centers (32%), study location (17%), recruitment dates (18%), and patient sampling (11%) were all reported in less than one-third of the abstracts.

Reporting of elements related to the design of the study was better: 51% of the abstracts reported whether data were collected prospectively or retrospectively, and it was clear in 94% of the abstracts whether the article reported on a single-gate or a multiple-gate study.

The reference standard was described in 57% of the abstracts, but all reported the index test. Information on the index test often included some technical specifications (70%), but rarely included details on cutoffs and categories for test positivity (35%), and information on whether readers were blinded to the results of the reference standard or other clinical data (13%).

Figure 3. Number of items reported by subgroups.

Figure 3a. Type of journal.

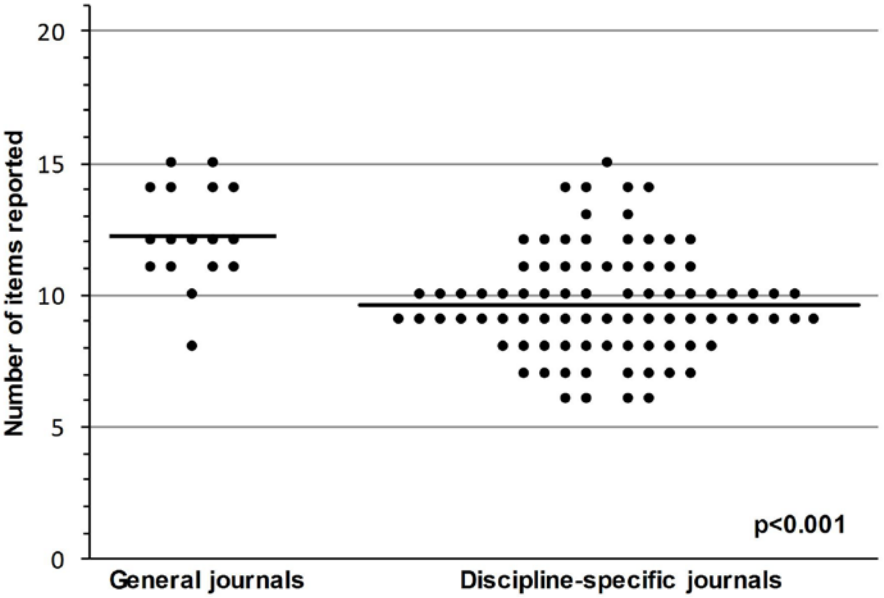


Figure 3b. Study design.

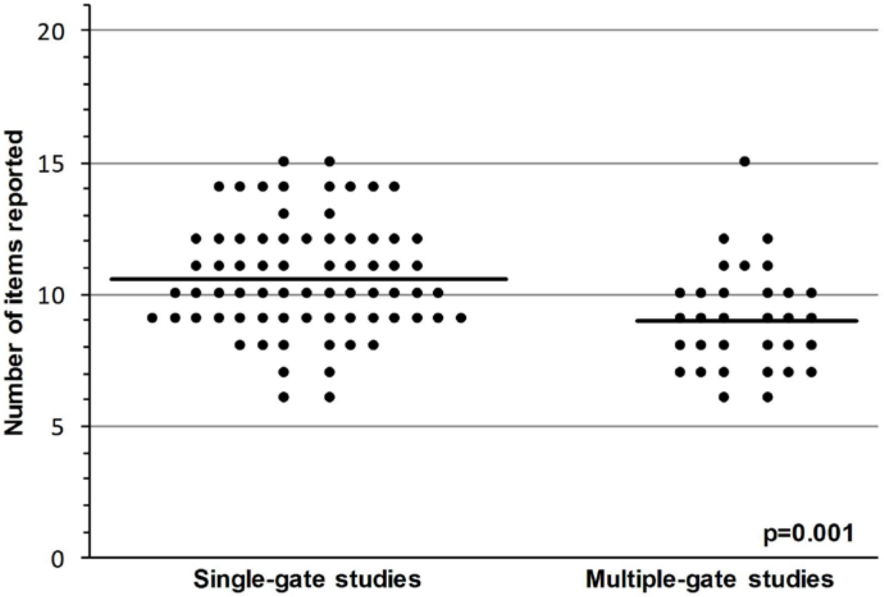


Figure 3. *Continued.*

Figure 3c. Type of test.

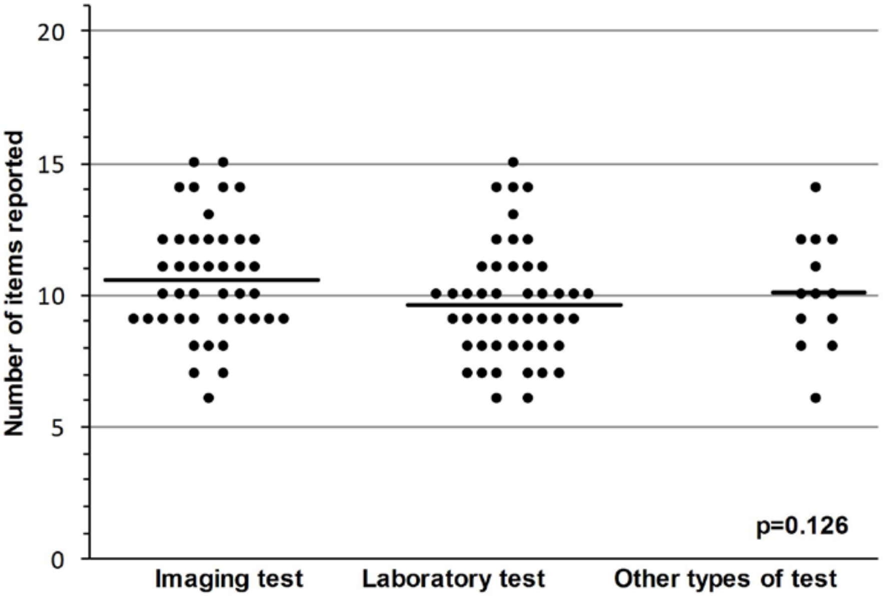


Figure 3d. Sample size.

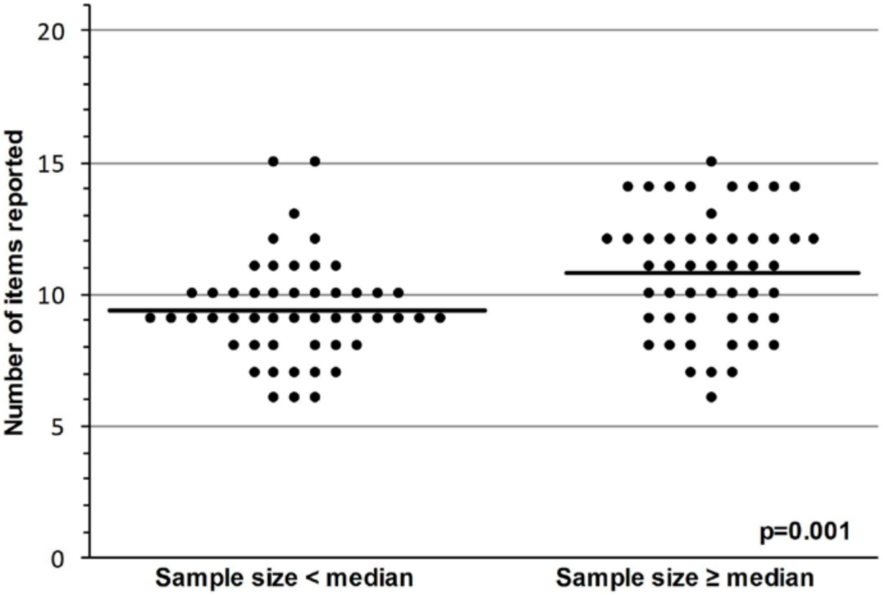
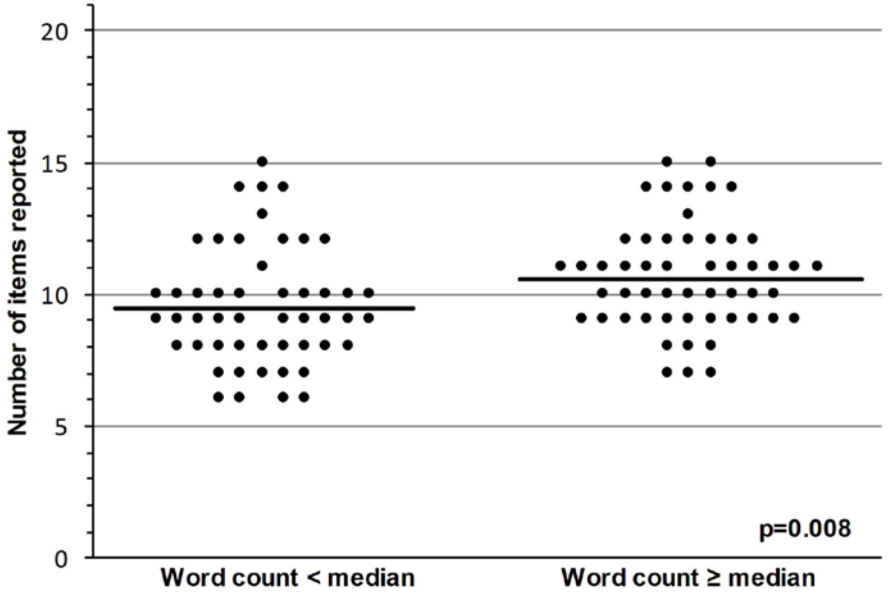


Figure 3. *Continued.*

Figure 3e. Abstract word count.



Each dot represents one article. The bold horizontal lines represent the mean number of items reported for each subgroup.

Results

All but five abstracts (95%) reported the number of participants included, but more specific information regarding demographic characteristics of participants, such as age and gender, was seldom provided (11% and 23%, respectively). Information on disease prevalence was reported by 72% of the abstracts, but the number of indeterminate or missing test results (6%), data to construct 2x2 tables (21%), and results on the reproducibility of the index test (e.g., by means of κ values; 17%) was rarely reported. Estimates of diagnostic accuracy, most often sensitivity and specificity, were available in 93% of the abstracts, but only 26% provided confidence intervals.

Discussion

All but five abstracts (95%) discussed the diagnostic accuracy of the index test under evaluation, but clear implications for future research (9%) and study limitations (3%) were rare.

Table 2. Items reported in journal abstracts of diagnostic accuracy studies.

Item	n
Total	103
Title	
Identify the article as a study of diagnostic accuracy in title	51 (50%)
Background and aims	
Rationale for study/background	47 (46%)
Research question/aims/objectives	86 (84%)
Methods	
Study population (at least one of following)	46 (45%)
a - inclusion/exclusion criteria	15 (15%)
b - study setting	33 (32%)
c - number of centers	33 (32%)
d - study location	17 (17%)
Recruitment dates	18 (18%)
Patient sampling (consecutive versus random sample)	11 (11%)
Data collection (prospective versus retrospective)	52 (51%)
Study design (multiple-gate versus single-gate)	97 (94%)
Reference standard	59 (57%)
Information on the index test (at least one of following)	103 (100%)
a - index test	103 (100%)
b - technical specifications and/or commercial name	72 (70%)
c - cut-offs, categories of results of index test	36 (35%)
Whether test readers were masked (at least one of following)	17 (17%)
a - when interpreting the index test	13 (13%)
b - when interpreting the reference standard	6 (6%)
Results	
Study participants (at least one of following)	98 (95%)
a - number of participants	98 (95%)
b - age of participants	11 (11%)
c - gender of participants	24 (23%)
Information on indeterminate results/missing values	6 (6%)
Disease prevalence	74 (72%)
2×2 tables (number of true and false positive and negative test results)	22 (21%)
Estimates of diagnostic accuracy (at least one of following)	96 (93%)
a - sensitivity and/or specificity	67 (65%)
b - negative and/or positive predictive value	20 (19%)
c - negative and/or positive likelihood ratio	2 (2%)
d - area under the ROC curve/C-statistic	36 (35%)
e - diagnostic odds ratio	0 (0%)
f - accuracy	13 (13%)
95% Confidence intervals around estimates of diagnostic accuracy	27 (26%)
Reproducibility of the results of the index test	17 (17%)
Discussion/Conclusion	
Diagnostic accuracy is discussed	98 (95%)
Implications for future research	9 (9%)
Limitations of study	3 (3%)

Discussion

We systematically evaluated the informativeness of abstracts of diagnostic accuracy studies published in 12 high-impact journals in 2012 and observed important weaknesses in the information provided. Key features of study design and a useful description of study results are often lacking, making proper identification and initial critical appraisal of studies difficult, if not impossible.

We only evaluated studies published in high-impact journals. This selection may have produced an overestimate of the number of items typically reported, as it is conceivable that the quality of diagnostic accuracy abstracts is poorer in low-impact journals. Evaluations of full-text articles in other fields of health research have shown poorer reporting quality in low-impact journals,[114] although this does not necessarily apply to abstracts. A minority of studies in our sample reported multiple results, not just diagnostic accuracy, in which case the abstract had to include information about these other study aims as well, within the journal's word limits. In the absence of proper prospective registration, it is difficult to identify the primary aims of these studies.[53,69] This was an exploratory analysis; the sample size was not calculated to detect differences between subgroups, and the results of subgroup analyses should be interpreted with caution.

Only a few previous studies have evaluated abstracts of diagnostic accuracy studies and only for specific tests or disciplines. Estrada et al. examined 33 abstracts of studies evaluating diagnostic tests for trichomoniasis published between 1976 and 1998, with regard to patient selection and spectrum, verification of index test results, and blinding.[147] None of the abstracts reported more than two of these four methodological criteria. Brazzelli et al. examined determinants of later full publication of 160 abstracts of diagnostic accuracy studies presented at two international stroke conferences between 1995 and 2004.[41] Although not their primary objective, they found that 65% did not report on type of data collection (prospective versus retrospective), 76% did not report on blinding of test results, and 89% did not state whether interobserver agreement had been assessed, whereas only one study did not report the sample size. This is very similar to our results.

Our analyses focused on whether items were reported in the abstract and not whether the abstract was an honest and balanced presentation of the study and its findings. Another review from our group demonstrated that about one in four abstracts of diagnostic accuracy studies are overoptimistic, with stronger conclusions in the abstract than in the full-text, selective reporting of results and discrepancies between the study aims and abstract conclusions, phenomena often referred to as 'spin'.[49] Lumbreras et al. evaluated 108 diagnostic accuracy studies on molecular research and graded all statements referring to the investigated test's clinical applicability, basing the final weight of this grading on the abstract.[66] Almost all articles (96%) made statements that were definitely favorable or promising and 56% overinterpreted the clinical applicability of their findings. Boutron et al. showed that overoptimistic abstracts are also highly prevalent in reports of randomized trials.[153]

Our list of items was developed to evaluate the informativeness of abstracts; it should not be considered as a proposal for a reporting guideline. We acknowledge that it may not be possible to report all 21 items within the word limits of a journal abstract. Guidelines for reporting of abstracts of randomized trials and systematic reviews have proposed 17 and 12 items, respectively.[138,139]

We also acknowledge that some of the 21 items may be more important than others. Providing essential items of study design is crucial for abstracts of diagnostic accuracy studies because diagnostic accuracy is not a fixed test property but reflects the behavior of a test in a particular clinical context and setting. Diagnostic accuracy studies are also prone to multiple sources of bias, and the abstract can inform the reader whether these biases were avoided. If not, the reader may want to skip the article and look further for information on the test's accuracy.

Inclusion criteria, study setting, and participant sampling, insufficiently reported in most abstracts we evaluated, are essential to the reader because disease severity and patient spectrum are well-established sources of variation of diagnostic accuracy.[27] Disease prevalence, one of the most often reported items in our evaluation but still lacking in more than a quarter of abstracts, is a major determinant of the applicability of study findings to another clinical situation because, contrarily to what clinicians usually think, diagnostic accuracy varies with disease prevalence.[27,154] Knowledge of the reference standard, not reported or unclear in almost half of the evaluated abstracts, is also crucial because the use of an inappropriate reference standard may lead to biased conclusions. Not providing confidence intervals around estimates of accuracy, as three-quarters of the evaluated abstracts did, could seriously mislead readers as the uncertainty of the estimates cannot be judged.

Poor reporting represents a waste of time and research resources.[9] Future scientific efforts could include the development of guidelines to facilitate writing sufficiently informative and transparent abstracts of diagnostic accuracy studies, as has been done for randomized trials and for systematic reviews.[138,139] Authors and editors could be actively stimulated to adopt and adhere to such guidelines.[155] Evaluations of the impact of CONSORT for Abstracts have shown an improvement in reporting quality after its launch,[156,157] especially among journals with an active implementation policy.[146] Yet developing guidelines is likely not enough. Guidelines should also be properly disseminated, accompanied by measures to facilitate their use.[155] There is evidence that journal endorsement of reporting guidelines improves completeness of reporting.[108] We believe initiatives to improve reporting quality must be multistaged and multitarget. Increasing awareness about the need for informative, complete, and balanced reporting is one such element, and this applies to study authors, reviewers, editors, and readers. Titles and abstracts are not promotional material but form an essential part of honest reporting, facilitating the timely identification and initial appraisal of studies for those in need of evidence to guide clinical decisions.

Acknowledgments

We thank W. Annefloor van Enst for her contributions to the literature selection as part of the full-text evaluation of adherence to STARD.

Chapter 9

Reporting weaknesses in conference abstracts of diagnostic accuracy studies in ophthalmology

Daniël A. Korevaar
Jérémie F. Cohen
Maurice W. de Ronde
Gianni Virgili
Kay Dickersin
Patrick M. Bossuyt

JAMA Ophthalmology 2015;133(12):1464-7

Appendices and supplemental material are available online at:
http://archopht.jamanetwork.com/issue.aspx?journalid=73&issueid=934754

Abstract

Background

Conference abstracts present information that helps clinicians and researchers to decide whether to attend a presentation. They also provide a source of unpublished research that could potentially be included in systematic reviews. We systematically assessed whether conference abstracts of studies that evaluated the accuracy of a diagnostic test were sufficiently informative.

Methods

We identified all abstracts describing work presented at the 2010 Annual Meeting of the Association for Research in Vision and Ophthalmology (ARVO). Abstracts were eligible if they included a measure of diagnostic accuracy, such as sensitivity, specificity, or likelihood ratios. Two independent reviewers evaluated each abstract using a list of 21 items, selected from published guidance for adequate reporting.

Results

A total of 126 of 6,310 abstracts presented were eligible. Only a minority reported inclusion criteria (5%), clinical setting (24%), patient sampling (10%), reference standard (48%), whether test readers were masked (7%), 2×2 tables (16%), and confidence intervals around accuracy estimates (16%). The mean number of items reported was 8.9 of 21 (SD 2.1; range 4 to 17).

Conclusions

Crucial information about study methods and results is often missing in abstracts of diagnostic studies presented at ARVO, making it difficult to assess risk for bias and applicability to specific clinical settings.

Introduction

Diagnostic accuracy studies evaluate how well a test distinguishes diseased from non-diseased individuals by comparing the results of the test under evaluation ("index test"), with the results of a reference (or "gold") standard. Deficiencies in study design can lead to biased accuracy estimates, suggesting a level of performance that can never be reached in clinical practice. In addition, because of variability in disease prevalence, patient characteristics, disease severity, and testing procedures, accuracy estimates may vary across studies evaluating the same test.[27] For example, in one Cochrane review, the sensitivity of optical coherence tomography in detecting clinically significant macular edema in patients with diabetic retinopathy ranged from 0.67 to 0.94 across included studies, and specificity ranged from 0.61 to 0.97.[158]

Given these potential constraints, readers of diagnostic accuracy study reports should be able to judge whether the results could be biased and whether the study findings apply to their specific clinical practice or policy-making situation.[28,29]

Conference abstracts often are short reports of actual studies, presenting information that helps clinicians and researchers to decide whether to attend a presentation. They also provide a source of unpublished research that could potentially be included in systematic reviews.[14] These decisions should be based on an early appraisal of the risk for bias and applicability of the abstracted study. We systematically evaluated the informativeness of abstracts of diagnostic accuracy studies presented at the 2010 Annual Meeting of the Association for Research in Vision and Ophthalmology (ARVO).

Methods

The online abstract proceedings from ARVO were searched for diagnostic accuracy studies presented in 2010 (eTable 1, available online). One reviewer (D.A.K.) assessed identified abstracts for eligibility. Abstracts were included if they reported on the diagnostic accuracy of a test in humans and stated that they calculated one or more of the following accuracy measures: sensitivity, specificity, predictive values, likelihood ratios, area under the receiver operating characteristic curve, or total accuracy.

For each abstract, one reviewer (D.A.K.) extracted the research field, commercial relationships, support, study design, sample size, and word count (Table 1). Extraction was independently verified by a second reviewer (J.F.C. or M.W.dR.).

The informativeness of abstracts was evaluated using a previously published list of 21 items, selected from existing guidelines for adequate reporting (Table 2; eTable 2).[63] The items focus on study identification, rationale, aims, design, methods for participant recruitment and testing, participant characteristics, estimates of accuracy, and discussion of findings. Two reviewers (D.A.K., and J.F.C. or M.W.dR.) independently scored each abstract. Disagreements were solved through discussion.

Table 1. Mean number of items reported among conference abstracts of diagnostic accuracy studies, stratified by study characteristics.

	n	Mean number of items reported (SD)	*p-value*[a]
Total	126	8.9 (2.1)	
Research field			
Glaucoma	51 (41%)	9.1 (1.6)	0.35
Other than glaucoma	75 (59%)	8.8 (2.4)	
Ocular surface and corneal diseases	16 (13%)	-	
Common chorioretinal diseases	15 (12%)	-	
Various types of uveitis	9 (7%)	-	
Optic nerve diseases	7 (6%)	-	
Other	28 (22%)	-	
Commercial relationships			
≥one author	44 (35%)	8.9 (2.0)	0.85
No author	82 (65%)	9.0 (2.2)	
Support			
Industry support	12 (10%)	8.4 (1.4)	0.58
No industry support	114 (90%)	9.0 (2.2)	
Study design[b]			
Cohort study	38 (35%)	10.1 (2.5)	0.001
Case-control study	72 (66%)	8.6 (1.5)	
Number of patients[c], median (IQR)	100 (50-160)		
<100	50 (49%)	9.0 (2.4)	0.26
≥100	53 (51%)	9.5 (1.8)	
Number of eyes[d], median (IQR)	136 (55-219)		
<136	33 (49%)	8.4 (2.0)	0.03
≥136	34 (51%)	9.4 (1.9)	
Word count[e], median (IQR)	301 (255-327)		
<301	61 (49%)	8.8 (2.0)	0.40
≥301	65 (51%)	9.1 (2.3)	

[a]Mean number of items reported across subgroups was compared using the t test. [b]Unclear for 16 abstracts. [c]Unclear for 23 abstracts. [d]Unclear or not applicable for 59 abstracts. [e]Excluding title, affiliations, commercial relationships, support, references, keywords, tables, and figures.

Results

Of 6,310 abstracts accepted at ARVO 2010, we identified 126 as reporting on diagnostic accuracy studies (eReferences). Abstract characteristics are provided in Table 1. The most common target condition was glaucoma (n=51); corresponding studies mostly (n=39) evaluated imaging of the retinal nerve fiber layer, other retina and choroid structures, or optic disc morphology. Ocular surface and corneal disease (keratoconus and dry eye) and common chorioretinal diseases (diabetic retinopathy and age-related macular degeneration) were targeted in 16 and 15 studies, respectively, followed by various types of uveitis and optic nerve diseases in nine and seven studies, respectively.

The reporting of individual items is presented in Table 2; examples of complete reporting per item are provided in eTable 3. Several elements that are crucial when assessing risk for bias or applicability of the study findings were rarely reported: inclusion criteria (5%), clinical setting (24%), patient sampling (10%), reference standard (48%), masking of test readers (7%), 2×2 tables (16%), and confidence intervals around accuracy estimates (16%). None of the abstracts reported all of these items. Reporting was better for other crucial elements: study design (87%), test under evaluation (100%), number of participants (82%), and disease prevalence (80%).

On average, the abstracts reported 8.9 of the 21 items (SD 2.1; range 4 to 17). Twenty-four abstracts (19%) reported more than half of the items (Figure 1). The mean number of reported items was significantly lower in abstracts of case-control studies compared with cohort studies (p=0.001), and in abstracts with sample sizes (number of eyes) below the median (p=0.03) (Table 1).

Discussion

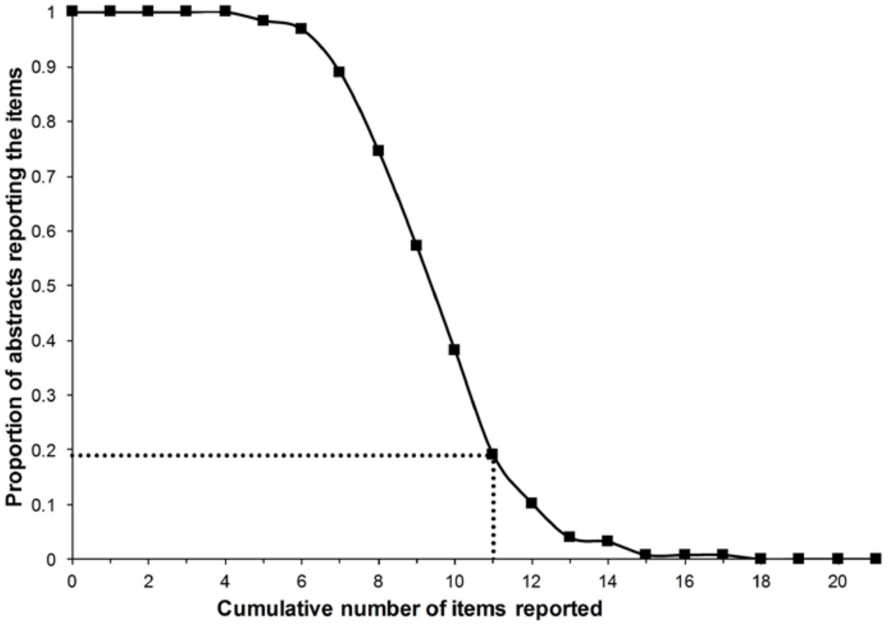
The informativeness of abstracts of diagnostic accuracy studies presented at the 2010 ARVO Annual Meeting was suboptimal. Several key elements of study methods and results were rarely reported, making it difficult for clinicians and researchers to evaluate method quality.

Differences in patient characteristics and disease severity are known sources of variability in accuracy estimates, and non-consecutive sampling of patients can lead to bias.[27,28] Therefore, readers want to know where and how patients were recruited,[29] yet less than a quarter of abstracts reported inclusion criteria, clinical setting, and sampling methods.

Table 2. Items reported in conference abstracts of diagnostic accuracy studies.

Item	n
Total	126
Title	
Identify the article as a study of diagnostic accuracy in title	57 (45%)
Background and aims	
Rationale for study/background	34 (27%)
Research question/aims/objectives	103 (82%)
Methods	
Study population (at least one of following)	36 (29%)
a - inclusion/exclusion criteria	6 (5%)
b - clinical setting	30 (24%)
c - number of centers	25 (20%)
d - study location	18 (14%)
Recruitment dates	15 (12%)
Patient sampling (consecutive versus random sample)	12 (10%)
Data collection (prospective versus retrospective)	27 (21%)
Study design (case-control versus cohort)	110 (87%)
Reference standard	61 (48%)
Information on the index test (at least one of following)	126 (100%)
a - index test	126 (100%)
b - technical specifications and/or commercial name	101 (80%)
c - cutoffs and/or categories of results of index test	40 (32%)
Whether test readers were masked (at least one of following)	9 (7%)
a - when interpreting the index test	6 (5%)
b - when interpreting the reference standard	5 (4%)
Results	
Study participants (at least one of following)	107 (85%)
a - number of participants	103 (82%)
b - age of participants	27 (21%)
c - gender of participants	7 (6%)
Information on indeterminate results/missing values	15 (12%)
Disease prevalence	101 (80%)
2×2 tables (number of true and false positive and negative test results)	20 (16%)
Estimates of diagnostic accuracy (at least one of following)	122 (97%)
a - sensitivity and/or specificity	84 (67%)
b - negative and/or positive predictive value	14 (11%)
c - negative and/or positive likelihood ratio	1 (1%)
d - area under the ROC curve/C-statistic	56 (44%)
e - diagnostic odds ratio	1 (1%)
f - accuracy	6 (5%)
95% Confidence intervals around estimates of diagnostic accuracy	20 (16%)
Reproducibility of the results of the index test	6 (5%)
Discussion/Conclusion	
Discussion of diagnostic accuracy results	120 (95%)
Implications for future research	23 (18%)
Limitations of study	1 (1%)

Figure 1. Proportion of conference abstracts of diagnostic accuracy studies that reported at least the indicated number of items on the 21-item list.



The dotted line indicates the proportion of abstracts the reported more than half of the evaluated items.

Risk for bias and applicability largely depend on the appropriateness of the reference standard.[28] However, the reference standard was not reported in half of the abstracts. Agreement between two tests is likely to increase if the reader of one test is aware of the results of the other test[27,28]; however, information about masking was available in only 7%.

About half of all conference abstracts are never published in full.[14] It is only possible to include the results of a conference abstract in a meta-analysis if the number of true-positive, true-negative, false-positive, and false negative test results are provided; however, 2×2 tables were only available in 16%. Although it is widely recognized that point estimates of diagnostic accuracy should be interpreted with measures of uncertainty, confidence intervals were reported in 16%.

Other crucial elements were more frequently provided. The study design, reported by 87%, is important because case-control studies produce inflated accuracy estimates owing to the extreme contrast between participants with and without the disease.[27,134] Diagnostic accuracy varies with disease prevalence, an important determinant of the applicability of study findings, and reported by 80%.

Suboptimal reporting in conference abstracts is not only a problem for diagnostic accuracy studies.[138] A previous evaluation of the content of abstracts of randomized trials presented at the ARVO Annual Meeting also found important study design information frequently unreported.[99] However, the authors concluded that missing information was often available in the corresponding ClinicalTrials.gov record. Because diagnostic accuracy studies are rarely registered,[69] complete reporting of conference abstracts is even more critical for these studies.

Using the same list of 21 items, we previously evaluated abstracts of diagnostic accuracy studies published in high-impact journals.[63] The overall mean number of items reported there was 10.1; crucial items about design and results were similarly lacking. One previous study assessed elements of reporting in conference abstracts of diagnostic accuracy studies in stroke research.[41] In line with our findings, 35% reported whether the data collection was prospective or retrospective, 24% reported on masking, and 11% reported on test reproducibility. Incomplete reporting is not only a problem for abstracts. Five previous reviews evaluated the reporting quality of full-study reports of ophthalmologic diagnostic accuracy studies, all of them pointing to important shortcomings.[77]

Conclusions

Crucial study information is often missing in abstracts of diagnostic accuracy studies presented at the ARVO Annual Meeting. Suboptimal reporting impedes the identification of high-quality studies from which reliable conclusions can be drawn. This is a major obstacle to evidence synthesis and an important source of avoidable research waste.[9]

Our list of 21 items is not a reporting checklist; we are aware that word count restrictions make it impossible to report all items in an abstract, and some items are more important than others. Reporting guidelines have been developed for abstracts of randomized trials and systematic reviews,[138,139] and a similar initiative is currently under way for diagnostic abstracts.[159] The scientific community should encourage informative reporting, not only for full-study reports, but also for conference abstracts.

Chapter 10

Updating standards for reporting diagnostic accuracy: the development of STARD 2015

Daniël A. Korevaar*
Jérémie F. Cohen*
Johannes B. Reitsma
David E. Bruns
Constantine A. Gatsonis
Paul P. Glasziou
Les Irwig
David Moher
Henrica C. de Vet
Douglas G. Altman
Lotty Hooft
Patrick M. Bossuyt

Research Integrity and Peer Review 2016;1:7

Appendices and supplemental material are available online at:
http://researchintegrityjournal.biomedcentral.com/articles/
10.1186/s41073-016-0014-7

*Equal contributors

Abstract

Background

Although the number of reporting guidelines has grown rapidly, few have gone through an updating process. The STARD statement (Standards for Reporting of Diagnostic Accuracy Studies), published in 2003 to help improve the transparency and completeness of reporting of diagnostic accuracy studies, was recently updated in a systematic way. Here, we describe the steps taken and a justification for the changes made.

Methods and Results

A four-member Project Team coordinated the updating process; a 14-member Steering Committee was regularly solicited by the Project Team when making critical decisions. First, a review of the literature was performed to identify topics and items potentially relevant to the STARD updating process. After this, the 85 members of the STARD Group were invited to participate in two online surveys to identify items that needed to be modified, removed from, or added to the STARD checklist. Based on the results of the literature review process, 33 items were presented to the STARD Group in the online survey: 25 original items and eight new items; 73 STARD Group members (86%) completed the first survey, and 79 STARD Group members (93%) completed the second survey. Then, an in-person consensus meeting was organized among the members of the Project Team and Steering Committee to develop a consensual draft version of STARD 2015. This version was piloted in three rounds among a total of 32 expert and non-expert users. Piloting mostly led to rewording of items. After this, the update was finalized. The updated STARD 2015 list now consists of 30 items. Compared to the previous version of STARD, three original items were each converted into two new items, four original items were incorporated into other items, and seven new items were added.

Conclusions

After a systematic updating process, STARD 2015 provides an updated list of 30 essential items for reporting diagnostic accuracy studies.

Background

The STARD statement (Standards for Reporting of Diagnostic Accuracy Studies) was published in 2003. It was intended to help improve the transparency and completeness of reporting of diagnostic accuracy studies. STARD presented a checklist of 25 items that authors should address when reporting diagnostic accuracy studies.[29,30]

Since its publication, STARD has been adopted by more than 200 biomedical journals.[133] Evaluations of adherence to STARD have revealed statistically significant but modest improvements over time in the reporting of diagnostic accuracy studies.[77,91,148] Unfortunately, reporting remains inadequate for many studies, and journals differ in the extent to which they endorse STARD, recommend it to authors, and use it in the editorial and peer review process.[160-163]

STARD had not been updated in the first 10 years of its existence. In February 2013, the STARD Steering Committee agreed that an update was justified to achieve two main goals (1) to include new items, based on improved understanding of sources of bias and variability, and (2) to facilitate the use of the list, by rearranging and rephrasing existing items, and by improving consistency in wording with other major reporting guidelines such as CONSORT (Consolidated Standards of Reporting Trials).[102]

Although the number of reporting guidelines has grown rapidly, few have gone through an updating process.[155] In this paper, we describe the steps taken to update the original STARD statement, resulting in STARD 2015,[31] and provide a justification for the changes made. The description of our methods may serve as guidance for other groups considering updates of their reporting guidelines.

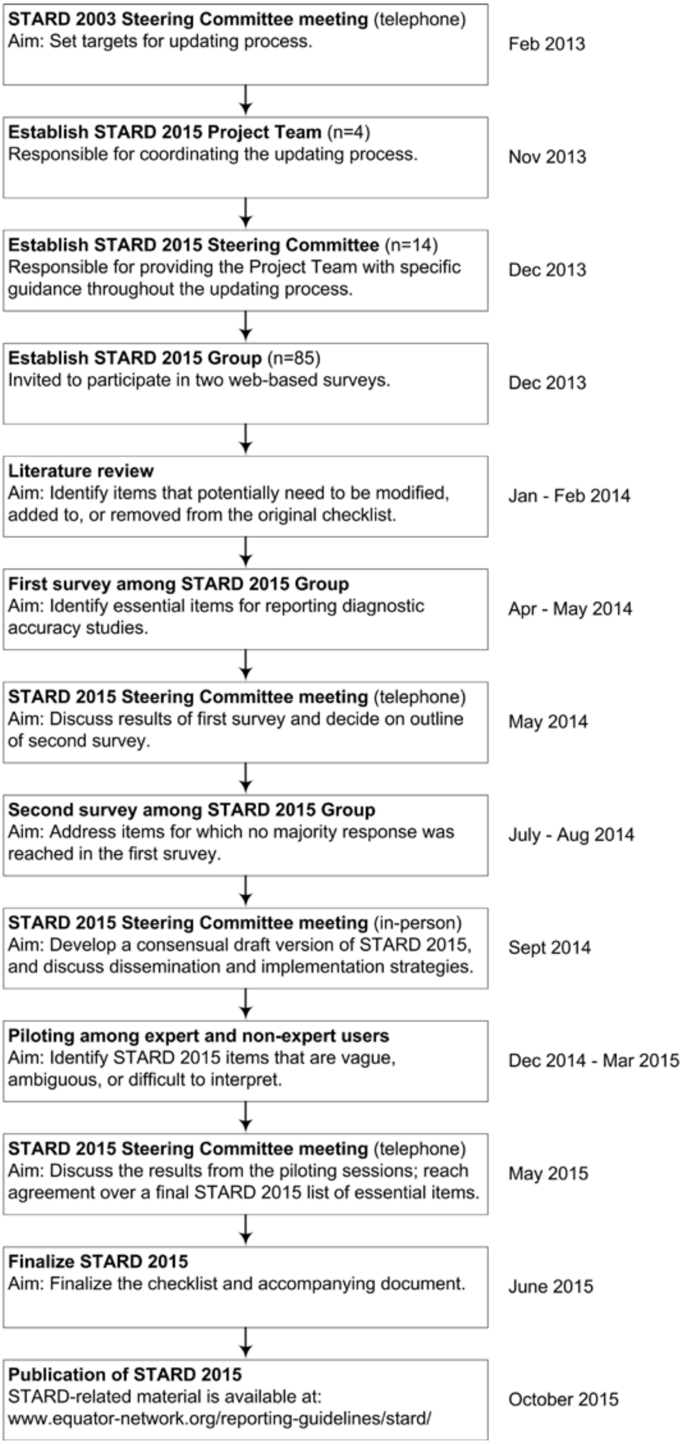
Methods

Figure 1 summarizes our approach for updating STARD and lists critical milestones.

Participants in the development of STARD 2015

The following groups of participants, detailed in Additional file 1, available online, were involved in the STARD updating process.

Figure 1. Milestones in the development of STARD 2015.



STARD 2003 Steering Committee meeting (telephone)
Aim: Set targets for updating process.
Feb 2013

Establish STARD 2015 Project Team (n=4)
Responsible for coordinating the updating process.
Nov 2013

Establish STARD 2015 Steering Committee (n=14)
Responsible for providing the Project Team with specific guidance throughout the updating process.
Dec 2013

Establish STARD 2015 Group (n=85)
Invited to participate in two web-based surveys.
Dec 2013

Literature review
Aim: Identify items that potentially need to be modified, added to, or removed from the original checklist.
Jan - Feb 2014

First survey among STARD 2015 Group
Aim: Identify essential items for reporting diagnostic accuracy studies.
Apr - May 2014

STARD 2015 Steering Committee meeting (telephone)
Aim: Discuss results of first survey and decide on outline of second survey.
May 2014

Second survey among STARD 2015 Group
Aim: Address items for which no majority response was reached in the first sruvey.
July - Aug 2014

STARD 2015 Steering Committee meeting (in-person)
Aim: Develop a consensual draft version of STARD 2015, and discuss dissemination and implementation strategies.
Sept 2014

Piloting among expert and non-expert users
Aim: Identify STARD 2015 items that are vague, ambiguous, or difficult to interpret.
Dec 2014 - Mar 2015

STARD 2015 Steering Committee meeting (telephone)
Aim: Discuss the results from the piloting sessions; reach agreement over a final STARD 2015 list of essential items.
May 2015

Finalize STARD 2015
Aim: Finalize the checklist and accompanying document.
June 2015

Publication of STARD 2015
STARD-related material is available at:
www.equator-network.org/reporting-guidelines/stard/
October 2015

Project Team

A four-member STARD 2015 Project Team was established, which was responsible for coordinating the updating process. This team secured funding, identified and invited potential new members of the STARD Group, reviewed the literature, conducted and analyzed web-based surveys, organized an in-person consensus meeting, drafted the items and accompanying documents, and coordinated piloting of the resulting STARD 2015 list.

Steering Committee

A 14-member STARD 2015 Steering Committee was also established, which was responsible for providing the Project Team with specific guidance throughout the updating process. This committee consisted of all 10 members of the STARD 2003 Steering Committee,[29] along with three journal editors from *Clinical Chemistry*, *JAMA*, and *Radiology*, and the founder of the EQUATOR Network (Enhancing the Quality and Transparency of Health Research), an umbrella organization that promotes complete and transparent reporting.[164]

STARD Group

All 30 members of the original STARD 2003 Group were invited to contribute to the updating process and to suggest potential new members. Other potential new STARD Group members were identified from STARD-related publications during discussions within the Project Team. The resulting STARD 2015 Group now has 85 members; it consists of researchers, journal editors, healthcare professionals, methodologists, a science journalist, statisticians, other stakeholders, and the members of the Project Team and Steering Committee. STARD Group members were invited to participate in two web-based surveys to help identify essential items for reporting diagnostic accuracy studies.

Review of the literature

In January and February 2014, the Project Team undertook a review of the literature to identify items that could be modified, added to, or removed from the original STARD checklist. This literature search focused on eight areas, which are detailed in Additional file 2.

In short, we searched Medline (through PubMed) and the Cochrane Methodology Register, supplemented by non-systematic searches, for topics and items potentially relevant to the STARD updating process in three categories: (1) general considerations about diagnostic accuracy studies and reporting, (2) evidence and statements suggesting modifications to the original STARD checklist or flow diagram, and (3) evidence and statements suggesting new STARD items.

Titles and abstracts were screened by one of two reviewers (D.A.K. or J.F.C.), and potentially eligible publications were retrieved for full-text assessment, again by one of these two reviewers. The electronic search results were augmented by the personal article collections of the Project Team. Based on the results of this search, the Project Team decided which items should be presented for consideration to the STARD Group in an online survey.

Online survey

General structure

We used two web-based surveys to help decide on items that needed to be modified, added to, or removed from the STARD checklist.[155] The surveys were developed by the Project Team in SurveyMonkey© and informally piloted in their institution prior to distribution. All 85 members of the STARD Group were invited by email to participate in each survey. Near the closing dates, non-responders were sent two reminders, one week apart.

Participant responses were summarized by the Project Team and reported back to participants at the end of each survey. The Project Team and Steering Committee had a teleconference in May 2014 to discuss the results of the first survey and to decide on the outline of the second survey. They also set priorities for topics to discuss during the in-person consensus meeting.

First survey

A link to the first survey was sent to the STARD Group on April 16, 2014; the survey was closed on May 31, 2014. The questionnaire consisted of two parts, each containing a set of multiple-choice questions and is provided in Additional file 3.

In the first part of the questionnaire, participants were asked to comment on each of the 25 original STARD items, in order of their appearance in the original checklist. For each item, participants were invited to indicate whether they would prefer to keep the item as it is, to modify the item, or remove the item from the checklist. Each question was accompanied by a suggestion from the Project Team, supported by a brief rationale, based on the literature search results. Each question also contained an open-comment box in which participants could clarify their responses.

In the second part of the questionnaire, participants were asked whether or not they felt that proposed potential new items should be added to the list. The questionnaire also addressed general considerations about the scope of STARD and preferred wording and a box for further suggestions.

Second survey

A link to the second survey was distributed to the STARD Group on July 16, 2014; this survey closed on August 30, 2014. The invitation letter contained a document that summarized the results of the first survey. The questionnaire is provided in Additional file 4.

This second survey focused on items for which less than 75% of the responders agreed on one of the multiple choice options in the first survey. Response options that had been selected by less than 20% of the respondents in the first survey were removed from the questionnaire. Based on the open comments provided by the respondents in the first survey, a brief summary of the main arguments for and against each proposed modification was presented for each item.

Results from the second survey were summarized by the Project Team and used to prepare the first draft version of STARD 2015. Items for which there was no majority response were considered high-priority topics for discussion during the in-person consensus meeting.

In-person consensus meeting

The 14 members from the STARD 2015 Steering Committee were invited to a two-day consensus meeting, held in Amsterdam, the Netherlands, on September 27 and 28, 2014. The meeting was organized, coordinated, and chaired by the Project Team. The primary objective was to develop a consensual draft version of STARD 2015. Secondary objectives were to discuss dissemination and implementation plans for STARD 2015 and additional initiatives around STARD and to discuss how STARD 2015 could be integrated into long-term development strategies of the EQUATOR Network.[164]

After the meeting, Project Team members further revised the consensual draft version of STARD 2015, with collected comments and suggestions, and modified the prototype flow diagram that was provided in the original STARD statement. The updated consensual draft version was circulated by email to the STARD Group for feedback. The Project Team collected comments and suggestions and modified the list accordingly.

Piloting STARD 2015

Three rounds of piloting among expert and non-expert users of STARD were organized. The main aim of these piloting sessions was to identify items on the consensual draft version of STARD 2015 that were vague, ambiguous, difficult to interpret, or missing.

Piloting among radiology residents

STARD 2015 was piloted among radiology residents from the Department of Radiology, Academic Medical Center, University of Amsterdam, the Netherlands. Residents were invited through email to read a diagnostic accuracy study report,[165] and to use the checklist to evaluate completeness of reporting. This was followed by a focus group meeting, which took place on December 15, 2014. During a 90-minute conversation, the moderator (D.A.K.) invited the participants to comment on the wording and on the layout of the list. Thereafter, participants were invited to share how they had evaluated each item in the article provided and their experience with using the checklist.

Piloting among radiology experts

The editor-in-chief of *Radiology* invited editorial board members and reviewers of diagnostic accuracy studies to pilot the consensual draft version of STARD 2015 and to provide comments using an online questionnaire developed by the Project Team (Additional file 5). Responses were collected in SurveyMonkey© between January 9 and April 1, 2015. Invitees were asked to answer eight "yes/no/no opinion" questions about the list, with the option to clarify answers in an open-comment box. Specifically, they were asked whether the aim of STARD 2015 was clear; whether terminology, layout, and outline used were appropriate; and whether any item or information was particularly difficult to understand or missing.

Piloting among laboratory medicine experts

The editor-in-chief of *Clinical Chemistry* invited editors and reviewers of the journal to evaluate the consensual draft version of STARD 2015. Responses were collected between February 26 and March 9, 2015. Collaborators were asked to review the list and to provide feedback on whether they found the language understandable and the items sufficiently clear. They were also asked to indicate if any information deemed essential in evaluating laboratory medicine diagnostic accuracy studies was currently not addressed. This was done by email.

Finalizing STARD 2015

The consensual draft version of the STARD 2015 list was updated following the piloting sessions. The Project Team summarized the feedback obtained from piloting and shared the results with the Steering Committee. In a teleconference on May 7, 2015, the Project Team and the Steering Committee decided on the final STARD 2015 list of essential items.

Initial strategies for disseminating STARD 2015

In August 2015, we non-systematically searched PubMed for editorials and news items that had been published about STARD since its launch in 2003, and 33 were identified, published in 28 different journals. One author (J.F.C.) collected the email addresses of the editors-in-chief or the editorial offices of these publishing journals. On November 26, 2015, these were contacted to inform them about the STARD 2015 update and to invite them to write an editorial or commentary around it.

In August 2015, we also searched PubMed for diagnostic accuracy studies that had been published between January and December 2014, using the following strategy: (sensitivity[tw] AND specificity[tw]) OR diagnostic accuracy[tw] OR predictive value*[tw] OR likelihood ratio*[tw] OR AUC[tw] OR ROC[tw]). We then ordered the search results by journal and established a list of the 100 journals that published most studies. For these journals, one author (D.A.K.) collected the email addresses of the editors-in-chief or the editorial offices, and these were contacted on February 4, 2016, to inform them about the STARD 2015 update, and with the request to consider using and endorsing it.

Results

Review of the literature

A total of 113 full-text articles and reports were reviewed in preparation for the STARD 2015 update. A summary of the results of the literature review is provided in Additional file 6.

Based on the results of this review process, the Project Team decided to present 33 items - the 25 original items and eight new items - for consideration to the STARD Group in the online survey. These eight potential new items were (1) positivity cutoffs for continuous tests when reporting area under the receiver operating characteristic curve, (2) sample size calculation, (3) trial registration number, (4) link to online resources, (5) availability of the study protocol, (6) data sharing policy, (7) conflicts of interest, and (8) sources of funding.

Online survey

First survey

Seventy-three STARD Group members (86%) completed the first survey. Detailed survey results are provided in Additional file 7. For the 25 items in the original STARD checklist, more than three quarters of respondents agreed to keep five items as they were (original STARD items 5/10/17/18/21) and to modify 13 items (original STARD items 2/4/6/8/9/11/12/13/14/16/19/22/24). There was less than 75% agreement on the seven other items (original STARD items 1/3/7/15/20/23/25). Of the eight potential new items proposed, more than 75% of respondents voted in favor of including four: sample-size calculation, availability of the study protocol, conflicts of interest, and sources of funding.

Second survey

Seventy-nine STARD Group members (93%) completed the second survey. Detailed survey results are provided in Additional file 7. The survey addressed eight remaining questions: six items on the original STARD checklist for which less than 75 % of respondents indicated the same answer in the first survey (original STARD items 3/7/15/20/23/25), one potential new item (positivity cutoffs for continuous tests when reporting area under the receiver operating characteristic curve), and one wording issue (continuing to use the term “diagnostic accuracy” rather than moving to “diagnostic performance” as the key concept in reporting comparisons of medical tests with a clinical reference standard). More than 75% voted to keep original STARD item 20 unchanged and to modify item 23 as suggested by the Project Team. No majority response was obtained for the other six questions.

In-person consensus meeting

The Project Team and all but three of the 14 members of the Steering Committee attended the in-person consensus meeting (Additional file 1). On the first day, the items in the draft version of STARD 2015 and items for which no 75% majority response were reached in the survey were discussed until consensus was reached on inclusion and phrasing. Thereafter, discussions focused on dissemination and uptake by journals, research institutions and authors, and strategies for piloting the list. It was also decided that a subgroup should develop a one-page explanatory document that briefly describes the aims of STARD 2015 and the key concepts in it to accompany the 2015 version when distributed.

On the second day, further discussions focused on finalizing a consensual draft version of STARD 2015. After this, additional initiatives around STARD were discussed. The meeting participants agreed that it would be valuable to develop extensions of STARD with more specific guidance for reporting diagnostic accuracy studies in different research fields (e.g., laboratory medicine and radiology) and applications of STARD for specific forms of testing (e.g., physical examination) or specific target conditions (e.g., dementia). The group agreed that STARD should also develop guidance for writing abstracts of diagnostic accuracy studies (STARD for Abstracts; in progress) and for registering protocols of diagnostic accuracy studies in trial registries (STARD for Registration; in progress).

Piloting STARD 2015

Piloting among radiology residents

Four radiology residents (three men, one woman; age range 25 to 35 years; two of them with a PhD) participated in the initial piloting. Three of them declared being aware of the existence of STARD; two had previously used STARD for the critical appraisal of a diagnostic accuracy report they had to present during weekly journal clubs at the Department of Radiology. Comments of the participants were collected. From the interviews, we concluded that a majority of items on the consensual draft version of STARD 2015 were relevant and understandable by non-expert users. Residents suggested minor rewording for some items, adding explanation of key terms (such as “target condition” and “intended use of a test”), and a pointer to STARD for Abstracts currently in development.

Piloting among radiology experts

Twenty editorial board members and peer reviewers from *Radiology* completed the online piloting survey. Seventeen respondents were clinical radiologists, two were journal editors, and one was a biomedical researcher. All but one respondent declared having previously (co-)authored a diagnostic accuracy study. Detailed results are provided in Additional file 8. Most respondents considered the consensual draft version of the STARD 2015 list of essential items and accompanying one-page explanatory document as understandable and complete.

Piloting among laboratory medicine experts

Eight experts in the field of laboratory medicine provided feedback on the consensual draft version of STARD 2015 and the one-page explanation. Three experts indicated that the current draft version may not cover important elements of laboratory test evaluations, such as reproducibility of tests and collection, handling, and storage of samples. These experts highlighted the need for specific extensions or complementary documents dedicated to laboratory tests. Some respondents also suggested minor modifications and edits to the list.

Finalizing STARD 2015

Amended draft versions of STARD 2015 were prepared. Based on the feedback provided during piloting, a new item pointing to STARD for Abstracts was added to the checklist, and a table to clarify key STARD terminology was developed.[31] Additional changes at this stage consisted mostly of minor wording modifications. On May 7, 2015, the Project Team and Steering Committee met in a teleconference during which the results from the piloting sessions were discussed, and STARD 2015 was finalized (Table 1).[31]

STARD 2015 consists of 30 items, with four items having an (a) and (b) part. A detailed rationale for modifying or adding items is provided in Additional file 9, with a summary of the main changes in Table 2. Compared to the original STARD checklist, three original items were each converted into two new items, four original items were incorporated into other items and seven completely new items were added. A modified prototype flow diagram, to illustrate the flow of participants through the study, was incorporated (Figure 2). The remaining items were reworded to make them easier to understand or to bring them in line with phrasing used in other major reporting guidelines, such as CONSORT.[102] STARD 2015 now also has an accompanying one-page explanatory document that can be distributed along with it (Additional file 10). An updated "Explanation and Elaboration" document, which explains each item in detail and gives examples of good reporting,[30] is under development; this document will be submitted for publication.

The STARD 2015 list and the explanatory document have been released under a Creative Commons license that allows for redistribution, and commercial and noncommercial use, as long as it is passed along unchanged and in whole, with credit to the STARD Group. All STARD-related material will be made accessible through the EQUATOR website upon completion (http://www.equatornetwork.org/reporting-guidelines/stard/).

Discussion

Having completed the update of STARD, we would like to share a few observations and reflections. These can be read as limitations that we acknowledge, encouragement for others who are considering an update or an extension of a reporting guideline, and background information for users of reporting guidelines, such as STARD.

Even though STARD intends to cover reports of all studies that provide estimates of a test's diagnostic accuracy, it may not be adequate to serve the special needs of each field. For specific types of tests and specific applications of testing, readers may wish to have more information to help them interpret and appreciate the study findings. The STARD Group encourages the development of extensions of STARD specifically designed for different fields of diagnostic research, and development of STARD applications, explaining how the STARD items should be operationalized for specific forms of testing or target conditions.[166,167] Such extensions should not replace the whole of STARD, but rather modify or extend individual items, or possibly just interpret the items in a particular context. More details on how to develop extensions have been reported elsewhere.[31]

Based on the accumulated experience since the development of STARD in 2003, we now firmly believe that developing a reporting checklist is in itself not sufficient to improve reporting.[168] We now see STARD 2015 as a list of essential items that provides a basis from which additional instruments have to be developed to address the needs of particular audiences. Though based on the STARD 2015 items, these instruments may differ, as they will target different potential users: not only authors of completed studies but also peer reviewers, journal editors, authors of conference abstracts, authors of study protocols, maybe even readers. Such instruments could, for example, be specific templates with standard text for authors, to facilitate complete reporting, or prototype statements for peer reviewers, who can point to reporting failures and explain why they need to be addressed. A writing aid for authors has been shown to be beneficial for improving reporting of randomized trials.[169] Other instruments that can be derived from the STARD 2015 items are guidance for reporting journal and conference abstracts and for registration of protocols of diagnostic accuracy studies in trial registries, initiatives that are currently ongoing.

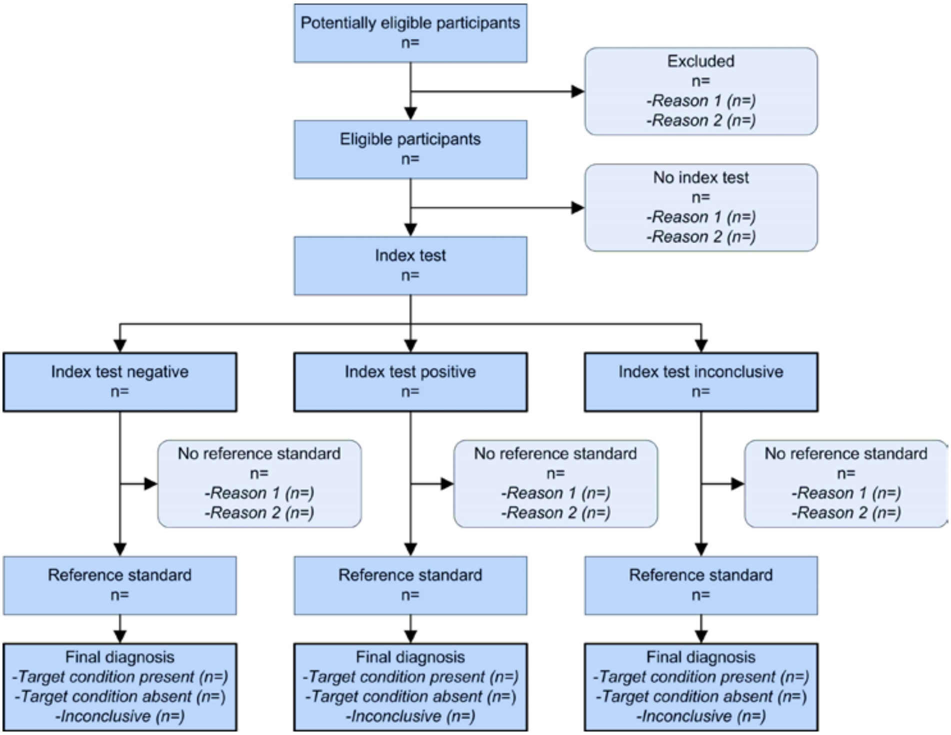
Table 1. The STARD 2015 list.

Section and topic	#	Item
Title or abstract		
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)
Abstract		
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)
Introduction		
	3	Scientific and clinical background, including the intended use and clinical role of the index test
	4	Study objectives and hypotheses
Methods		
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)
Participants	**6**	Eligibility criteria
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)
	8	Where and when potentially eligible participants were identified (setting, location, and dates)
	9	Whether participants formed a consecutive, random, or convenience series
Test methods	**10a**	Index test, in sufficient detail to allow replication
	10b	Reference standard, in sufficient detail to allow replication
	11	Rationale for choosing the reference standard (if alternatives exist)
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory
	13a	Whether clinical information and reference standard results were available to the performers or readers of the index test
	13b	Whether clinical information and index test results were available to the assessors of the reference standard
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy
	15	How indeterminate index test or reference standard results were handled
	16	How missing data on the index test and reference standard were handled
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory
	18	Intended sample size and how it was determined
Results		
Participants	**19**	Flow of participants, using a diagram
	20	Baseline demographic and clinical characteristics of participants
	21a	Distribution of severity of disease in those with the target condition
	21b	Distribution of alternative diagnoses in those without the target condition
	22	Time interval and any clinical interventions between index test and reference standard
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
	25	Any adverse events from performing the index test or the reference standard
Discussion		
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability
	27	Implications for practice, including the intended use and clinical role of the index test
Other information		
	28	Registration number and name of registry
	29	Where the full study protocol can be accessed
	30	Sources of funding and other support; role of funders

Table 2. Summary of main changes in STARD 2015.

Section	Authors are invited to..
Title/Abstract	..report a structured abstract, according to STARD for Abstracts (item 2).
Introduction	..report the intended use and clinical role of the index test under investigations (item 3), along with specific study hypotheses (item 4).
Methods	..report whether test positivity cut-offs or result categories were pre-specified or exploratory (item 12), whether analyses of variability in diagnostic accuracy were pre-specified or exploratory (item 17), and how they determined the intended sample size (item 18).
Results	..always provide a diagram, illustrating the flow of participants through the study (item 19).
Discussion	..discuss potential study limitations (item 26) and the implications for practice of the study findings (item 27).
Other information	..report the registration number (item 28), where the full study protocol can be accessed (item 29), and sources of funding (item 30).

Figure 2. Prototypical STARD diagram to report flow of participants through the study.



Most reporting guidelines have not undergone user testing prior to their release, which may be surprising, given that reporting guidelines are primarily tools designed to help others, and they should be evaluated as such. We therefore decided to pilot STARD 2015 among different groups of potential users. This piloting was still relatively modest, but it helped us to improve the list in several key respects, especially in terms of wording.

Although we substantially extended membership of the STARD Group, the STARD 2015 update process mostly included experienced researchers and authors, and most of them were from USA, UK, or The Netherlands. To judge the formulation and user friendliness of items, the opinion of future users is important as well. The selection of items should be based on strong evidence and sound principles but the development of actual tools and instruments should be guided by repeated, targeted, and methodical user testing.

Conclusions

After a systematic updating process, STARD 2015 provides an updated list of 30 essential items for reporting diagnostic accuracy studies. Incomplete reporting is now considered to be one of the largest sources of avoidable waste in biomedical research.[9] We believe that reporting can be substantially improved, with relatively little effort from multiple parties: from those responsible for training researchers, from the authors themselves, from journal editors, from peer reviewers, and from funders.[25] We invite all stakeholders to help disseminate STARD 2015 and to help the STARD Group in its efforts to promote more complete, more transparent, and more informative reporting of evaluations of medical tests.

Acknowledgments

The authors thank Jeroen G. Lijmer, Drummond Rennie, Herbert Y. Kressel, Nader Rifai, and Robert M. Golub who were also members of the STARD 2015 Steering committee.

Part D

Diagnostic tests in respiratory medicine

Chapter 11

Diagnostic accuracy of minimally invasive markers for detection of airway eosinophilia in asthma: a systematic review and meta-analysis

Daniël A. Korevaar
Guus A. Westerhof
Junfeng Wang
Jérémie F. Cohen
René Spijker
Peter J. Sterk
Elisabeth H. Bel
Patrick M. Bossuyt

Lancet Respiratory Medicine 2015;3(4):290-300

Appendices and supplemental material are available online at:
http://www.thelancet.com/journals/lanres/issue/vol3no4/PIIS2213-2600(15)X0004-X

Abstract

Background

Eosinophilic airway inflammation is associated with increased corticosteroid responsiveness in asthma, but direct airway sampling methods are invasive or laborious. Minimally invasive markers for airway eosinophilia could present an alternative method, but estimates of their accuracy vary.

Methods

We did a systematic review and searched Medline, Embase, and PubMed for studies assessing the diagnostic accuracy of markers against a reference standard of induced sputum, bronchoalveolar lavage, or endobronchial biopsy in patients with (suspected) asthma (inception to August 1, 2014). Unpublished results were obtained by contacting authors of studies that did not report on diagnostic accuracy, but had data from which estimates could be calculated. We assessed risk of bias with QUADAS-2. We used meta-analysis to produce summary estimates.

Results

We included 32 studies: 24 in adults and eight in children. Of these, 26 (81%) showed risk of bias in at least one domain. In adults, three markers had extensively been investigated: fraction of exhaled nitric oxide (FeNO) (17 studies; 3,216 patients; summary area under the receiver operating characteristic curve (AUC) 0.75 (95%CI 0.72 to 0.78)); blood eosinophils (14 studies; 2,405 patients; 0.78 (95%CI 0.74 to 0.82)); total immunoglobulin E (IgE) (seven studies; 942 patients; 0.65 (95%CI 0.61 to 0.69)). In children, only FeNO (six studies; 349 patients; summary AUC 0.81 (95%CI 0.72 to 0.89)) and blood eosinophils (three studies; 192 patients; 0.78 (95%CI 0.71 to 0.85)) had been investigated in more than one study. Induced sputum was most frequently used as the reference standard. Summary estimates of sensitivity and specificity in detecting sputum eosinophils of 3% or more in adults were: 0.66 (95%CI 0.57 to 0.75) and 0.76 (95%CI 0.65 to 0.85) for FeNO; 0.71 (95%CI 0.65 to 0.76) and 0.77 (95%CI 0.70 to 0.83) for blood eosinophils; and 0.64 (95%CI 0.42 to 0.81) and 0.71 (95%CI 0.42 to 0.89) for IgE.

Conclusions

FeNO, blood eosinophils, and IgE have moderate diagnostic accuracy. Their use as a single surrogate marker for airway eosinophilia in patients with asthma will lead to a substantial number of false positives or false negatives.

Introduction

Historically, asthma control has been pursued by means of symptom and lung function monitoring.[32] Although asthma medications are effective in controlling the disease in most patients, a minority deteriorates despite maximum treatment. Non-eosinophilic asthma responds poorly to corticosteroid therapy, the standard treatment for suppressing airway inflammation. About half of patients with asthma seem to be persistently non-eosinophilic.[170]

Bronchoalveolar lavage and endobronchial biopsy are the reference standards for identifying the extent of eosinophilic airway inflammation, but these tests are invasive and expensive. Another option is induced sputum, which has been clinically useful in guiding asthma treatment.[171]

A Cochrane review showed that the frequency of asthma exacerbations is significantly lower in patients in whom inhaled corticosteroids are tailored based on sputum eosinophil levels, compared with those in whom management is based on traditional methods of asthma monitoring.[171] Recent guidelines recommend guiding treatment in severe asthma by sputum eosinophil counts in addition to clinical criteria in centers experienced in using this technique.[32,172] Sputum eosinophilia might also have prognostic value as a marker for persistent airflow limitation,[173] deteriorating asthma over time,[174] and responsiveness to future therapies specifically targeting eosinophilic inflammation, such as mepolizumab.[175]

Unfortunately, sputum induction is time-consuming, needs experienced laboratory personnel, and many patients are unable to produce adequate samples. Several minimally invasive markers of eosinophilic airway inflammation, such as fraction of exhaled nitric oxide (FeNO), blood eosinophils, and serum periostin, could have potential as a surrogate to replace sputum induction, but their accuracy to distinguish between patients with and without airway eosinophilia remains controversial.

We did a systematic review and meta-analysis to obtain summary estimates of the diagnostic accuracy[176] of markers for airway eosinophilia in patients with asthma.

Methods

Literature search and study selection

A medical information specialist (R.S.) developed searches in Medline, Embase, and PubMed without date or language restrictions (Appendix, available online). We included studies if they assessed the diagnostic accuracy of one or more blood, serum, nasal lavage, or exhaled breath markers (index test)[177] in detecting airway eosinophilia (target condition) in patients with asthma or suspected asthma. Direct airway sampling methods (induced sputum, bronchoalveolar lavage, or endobronchial biopsy) were acceptable reference standards, independent of the threshold for positivity used. We excluded review articles. The searches were updated until August 1, 2014. Two independent reviewers (D.A.K., G.A.W.) examined titles and abstracts of all search results. Full reports of studies that were considered potentially eligible by at least one of the reviewers were obtained and independently assessed for inclusion. Disagreements were resolved by consensus. One reviewer (D.A.K.) also scanned reference lists of included articles, and searched trial registries (ClinicalTrials.gov, Current Controlled Trials, Netherlands Trial Register, and Australian New Zealand Clinical Trials Registry) for unpublished or ongoing studies.

To enrich the number of included studies, we also tried to identify unpublished data by contacting authors of published studies that did not report on the diagnostic accuracy of a marker to detect airway eosinophilia, but seemed to have data from which accuracy estimates could be calculated (the "enrichment sample"). Studies were selected if they had done at least one index test and one reference standard, as defined above. Such studies were only eligible if they explicitly distinguished patients with airway eosinophilia from those without, included at least an arbitrary number of 50 patients with asthma, and were published before January 1, 2014. We contacted corresponding authors through email, and asked whether they were willing to calculate and share estimates of accuracy or to send their masked dataset.

Data extraction and quality assessment

One reviewer (D.A.K.) did the data extraction, which was verified by a second reviewer (G.A.W.). We identified the first author, country, journal, year of publication, recruitment setting, sample size, and characteristics of included patients (age, sex, BMI, atopy status, asthma severity, FEV_1 % predicted, smoking status, and corticosteroid treatment status). We also extracted the index test(s), reference standard(s), test positivity thresholds, disease prevalence, accuracy estimates, and data for 2×2 tables presenting index test results by reference standard results for each reported threshold. If 2×2 tables were not reported, we attempted to reconstruct them from summary estimates or by contacting corresponding authors through email. If it appeared from an article that many markers had been assessed, but diagnostic accuracy data were not reported for all of them, we contacted authors to obtain these data. Two authors (D.A.K., G.A.W.) independently assessed risk of bias and applicability concerns using QUADAS-2.[28]

Statistical analysis

Whenever we obtained datasets from studies in the enrichment sample, we assessed diagnostic accuracy as follows. First, we estimated the ability of each index test to discriminate between patients with and without airway eosinophilia by calculating the area under the receiver operating characteristic curve (AUC). Then, we selected the optimal cut-off of sensitivity and specificity on the receiver operating characteristic (ROC) curve using the Youden's index, as has been done by almost all included diagnostic accuracy studies. Depending on the reference standard available, we repeated this analysis for each definition of airway eosinophilia used in the included studies. Patients with missing data on the index test or reference standard were excluded from the analysis for that specific marker. We analyzed datasets using R version 3.0 (R Foundation for Statistical Computing, Vienna, Austria).

We analyzed studies in children and adults separately. To get a view of the overall diagnostic performance of each marker, we did a random effects meta-analysis of AUC estimates,[178] independent of the reference standard or definition of airway eosinophilia that had been used. Whenever a study reported more than one AUC estimate for one marker in the same group of patients, for example because the study relied on many definitions of airway eosinophilia, we included the highest AUC reported. If a study reported an AUC estimate in the total study group and in subgroups, we only included the estimate for the total study group. However, if a study reported on these estimates in subgroups only, and not in the total study group, we included the AUCs of all subgroups. If sufficient data were available (three or more studies), we repeated this meta-analysis for studies that used the same reference standard and airway eosinophilia definition. We assessed statistical heterogeneity using the I^2 statistic.[179]

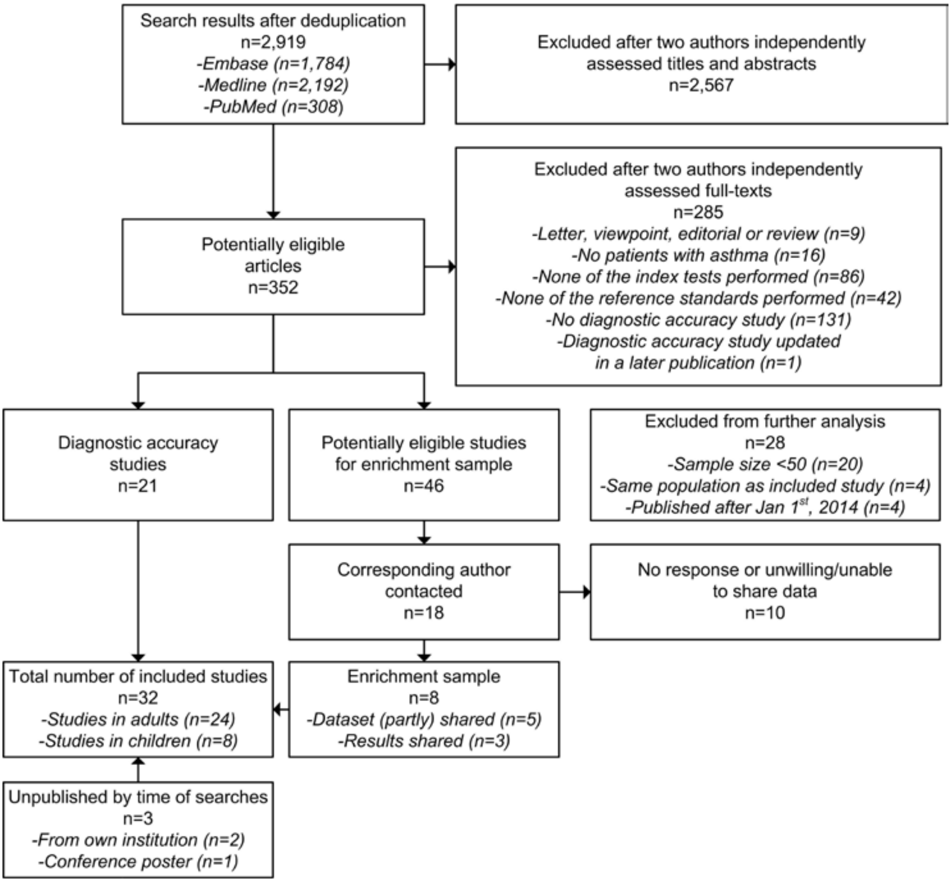
From each collected or reconstructed 2×2 table, we calculated estimates of sensitivity and specificity and 95%CIs. We used a hierarchical random effects model[176] to obtain summary estimates of sensitivity and specificity for studies that used the same reference standard and airway eosinophilia definition. We did so whenever four or more tables were available. If articles provided data on direct, head-to-head comparisons of two or more markers, we assessed whether there were significant differences in accuracy between markers. Such direct comparisons ensure that differences in accuracy are not caused by heterogeneity across study populations. We used Deeks' funnel plot asymmetry test to assess risk of publication bias.[20] We used SAS version 9.2 (SAS Institute, Cary, NC, USA) to fit the models.

Results

Search and selection

The searches retrieved 2,919 unique records, all of them providing titles or abstracts in English language. Among these, we found 21 eligible diagnostic accuracy studies (Figure 1). Another 18 studies fulfilled the eligibility criteria for the enrichment sample. Contacting the authors of these studies led to eight additional inclusions. We also included data from two studies from our own department, and one more identified through a conference poster. No additional studies were identified by scanning reference lists and searching trial registries. Overall, we included 24 studies done in adults,[170,173,180-201] and eight in children.[202-209]

Figure 1. Flowchart for selection of studies.


Search results after deduplication
n=2,919
-Embase (n=1,784)
-Medline (n=2,192)
-PubMed (n=308)
Excluded after two authors independently assessed titles and abstracts
n=2,567
Excluded after two authors independently assessed full-texts
n=285
-Letter, viewpoint, editorial or review (n=9)
-No patients with asthma (n=16)
-None of the index tests performed (n=86)
-None of the reference standards performed (n=42)
-No diagnostic accuracy study (n=131)
-Diagnostic accuracy study updated in a later publication (n=1)
Potentially eligible articles
n=352
Diagnostic accuracy studies
n=21
Potentially eligible studies for enrichment sample
n=46
Excluded from further analysis
n=28
-Sample size <50 (n=20)
-Same population as included study (n=4)
-Published after Jan 1st, 2014 (n=4)
Corresponding author contacted
n=18
No response or unwilling/unable to share data
n=10
Total number of included studies
n=32
-Studies in adults (n=24)
-Studies in children (n=8)
Enrichment sample
n=8
-Dataset (partly) shared (n=5)
-Results shared (n=3)
Unpublished by time of searches
n=3
-From own institution (n=2)
-Conference poster (n=1)

Study characteristics

Detailed characteristics of included studies are provided in the Appendix. All studies used a single set of inclusion criteria (cohort studies) and the number of patients included in the analysis of diagnostic accuracy varied from 24 to 566 in adults, and from 27 to 150 in children. The mean or median age ranged from 27.0 years to 59.8 years in adults, and from 6.8 years to 13.0 years in children. In all cases, study participants were recruited in secondary care or tertiary care facilities and both males and females were included. Studies in adults included patients with asthma of varying severity: mild-moderate (four studies, 17%), mild-severe (four studies, 17%), moderate-severe (four studies, 17%), severe (five studies, 21%), or not reported (seven studies, 29%). In children, asthma severity was mild (one study, 13%), mild-severe (one study, 13%), moderate-severe (one study, 13%), severe (two studies, 25%), or not reported (three studies, 38%). In adults, 12 studies (50%) included current non-smokers only, one study (4%) current smokers only, and 11 studies (46%) included both.

Two studies in adults (8%) assessed corticosteroid (inhaled or oral) untreated patients only, 11 studies (46%) assessed corticosteroid treated patients only, and 11 studies (46%) included both treated and untreated patients. In children, these numbers were one study (13%), three studies (38%), and four studies (50%), respectively. There were large between-study differences in atopy and asthma severity status.

In adults, 21 studies (88%) used only sputum as the reference standard, whereas two studies (8%) used sputum and endobronchial biopsy, and one study (4%) used bronchoalveolar lavage and endobronchial biopsy. In children, sputum was the reference standard in four studies (50%), bronchoalveolar lavage in two studies (25%), bronchoalveolar lavage and endobronchial biopsy in one study (13%), and sputum, bronchoalveolar lavage, and endobronchial biopsy in one study (13%).

Study quality

Detailed results of the QUADAS-2 assessment are provided in the Appendix. All but six studies (81%) showed risk of bias in at least one domain, often because thresholds for index test positivity had not been predefined (21 studies, 66%), or because more than 10% of the patients had been excluded because of missing reference standard results (14 studies, 44%). Additionally, methods for patient sampling (22 studies, 69%) or masking of the index test (20 studies, 63%), or masking of the reference standard (18 studies, 56%) were often unclear.

Meta-analysis: adults

All diagnostic accuracy data for markers and reference standards are summarized in the Appendix. Results of meta-analyses of AUC estimates are reported in Table 1.

In adults, five different definitions of airway eosinophilia were used across studies, most often based on sputum eosinophils of 2% or more, or 3% or more. The prevalence of eosinophilia ranged from 20% to 88%. We obtained diagnostic accuracy data for nine markers, but only FeNO, blood eosinophils, total immunoglobulin E (IgE), serum periostin, serum eosinophil cationic protein, and exhaled breath condensate pH were investigated in more than one study (Table 1). When we pooled data, independent of which reference standard or airway eosinophilia definition had been used, the summary AUC of these markers never exceeded 0.78. We found substantial heterogeneity in most analyses (Appendix).

FeNO (17 studies, 3,216 patients), blood eosinophils (14 studies, 2,405 patients), and IgE (seven studies, 942 patients) were investigated in more than two studies, with pooled AUC estimates of 0.75 (95%CI 0.72 to 0.78), 0.78 (95%CI 0.74 to 0.82), and 0.65 (95%CI 0.61 to 0.69), respectively. We repeated these meta-analyses for studies that used sputum eosinophil values of 3% or more and 2% or more as the definition of airway eosinophilia (Appendix), but the summary AUCs were barely affected: 0.74 (95%CI 0.70 to 0.78) and 0.73 (95%CI 0.68 to 0.77), respectively, for FeNO; 0.78 (95%CI 0.73 to 0.83) and 0.78 (95%CI 0.73 to 0.83) for blood eosinophils; and 0.63 (95%CI 0.57 to 0.69) and 0.66 (95%CI 0.62 to 0.70) for IgE.

Periostin showed promising performance in one study (AUC 0.84),[190] but these results were not replicated in a second study (AUC 0.55).[198] Nasal lavage eosinophils (AUC 0.88) and a model based on exhaled volatile organic compounds (AUC 0.98) showed high accuracy, but were only investigated in single studies (Table 1).

Three studies reported combinations of markers, but none of these showed a significant improvement in the diagnostic accuracy compared with single markers (data not shown).

Comparisons between published and unpublished diagnostic accuracy data for FeNO, blood eosinophils, and IgE are shown in the Appendix. Adding unpublished data led to a substantial increase in precision, but did not affect summary estimates of AUCs.

Table 1. Overall diagnostic performance of markers for detecting any airway eosinophilia.

	Studies in adults[a]				Studies in children[a]			
	Studies assessing marker	AUCs included	Sum of patients	AUC[b]	Studies assessing marker	AUCs included	Sum of patients	AUC[b]
	n	n	n	pooled (95%CI)	n	n	n	pooled (95%CI)
FeNO	17	19	3216	0.75 (0.72-0.78)	6	5	349	0.81 (0.72-0.89)
Blood eosinophils	14	14	2405	0.78 (0.74-0.82)	3	3	192	0.78 (0.71-0.85)
Serum IgE	7	7	942	0.65 (0.61-0.69)	0	-	-	-
Serum periostin	2	3	204	0.65 (0.49-0.81)	0	-	-	-
Serum ECP	2	2	174	0.72 (0.64-0.81)	1	1	77	0.75[c]
EBC pH	2	2	96	0.76 (0.63-0.90)	0	-	-	-
Exhaled VOCs	1	1	18	0.98[c]	0	-	-	-
EBC model	1	1	53	0.69[c]	0	-	-	-
Nasal lavage eosinophils	1	1	130	0.88[c]	0	-	-	-

Detailed information on diagnostic accuracy data for individual studies can be found in the Appendix. [a]Five different definitions of airway eosinophilia were used across studies, based on different thresholds for induced sputum, bronchoalveolar lavage, or endobronchial biopsy. [b]Results based on random effects meta-analysis. [c]Meta-analysis not possible as only one study reported on AUC.

Table 2. Summary estimates of sensitivity and specificity for detecting sputum eosinophilia in adults.

	Sputum eosinophils ≥3%				Sputum eosinophils ≥2%			
Index test	Studies n	Patients n	Sensitivity (95%CI)	Specificity (95%CI)	Studies n	Patients n	Sensitivity (95%CI)	Specificity (95%CI)
FeNO (ppb)	12	1720	0.66 (0.57-0.75)	0.76 (0.65-0.85)	9	1667	0.65 (0.55-0.74)	0.75 (0.62-0.84)
Blood eosinophils (per μL)	12	1967	0.71 (0.65-0.76)	0.77 (0.70-0.83)	6	1180	0.66 (0.56-0.75)	0.83 (0.62-0.94)
Blood eosinophils (%)	5	920	0.76 (0.52-0.90)	0.74 (0.67-0.80)	2	171	-	-
Serum IgE (IU/mL)	6	699	0.64 (0.42-0.81)	0.71 (0.42-0.89)	4	754	0.63 (0.36-0.84)	0.59 (0.37-0.79)

Sufficient data in adults (four or more studies) to do the meta-analysis of sensitivity and specificity for studies that used the same airway eosinophilia definition, were only available for induced sputum as the reference standard. Forest plots of FeNO, blood eosinophils, and IgE for detecting sputum eosinophil values of 3% or more and 2% or more are presented in Figures 2 and 3, with summary ROC curves given in Figure 4. Almost all studies used the optimal cut-off of sensitivity and specificity on the ROC curve to define the positivity threshold of the markers. These thresholds varied widely. For example, the optimal threshold for FeNO to detect sputum eosinophils of 3% or more ranged from 10 to 41 parts per billion (ppb).

Summary estimates of sensitivity and specificity of FeNO, blood eosinophils, and IgE for detecting sputum eosinophils of 3% or more and 2% or more, obtained by meta-analysis, are presented in Table 2. Sensitivity ranged from 0.63 to 0.76, and specificity ranged from 0.59 to 0.83. When pooling direct comparisons, FeNO was significantly more accurate than IgE in detecting sputum eosinophils of 2% or more (four studies; p=0.025), but not in detecting sputum eosinophils of 3% or more (five studies; p=0.34). Pooling of other direct comparisons (FeNO versus blood eosinophils, and IgE versus blood eosinophils) showed no significant differences (data not shown).

Statistical testing for funnel plot asymmetry showed no evidence of publication bias (Appendix). Forest plots of sensitivity and specificity of FeNO, blood eosinophils, and IgE for detecting sputum eosinophilia in subgroups based on smoking, treatment, and asthma severity status are shown in the Appendix. Also in these subgroups, positivity thresholds of the markers varied considerably at the optimal cut-off of sensitivity and specificity.

Meta-analysis: children

In children, five different definitions of airway eosinophilia were used across studies, most often based on sputum eosinophil values of 2.5% or more (Appendix). The prevalence of eosinophilia ranged from 21% to 81%. Diagnostic accuracy was assessed for three markers; two of them in more than one study (Table 1): FeNO (six studies, 349 patients) and blood eosinophils (three studies, 192 patients) had pooled AUC estimates of 0.81 (95%CI 0.72 to 0.89) and 0.78 (95%CI 0.71 to 0.85), respectively.

For children, the summary ROC curve and forest plot of FeNO for detecting sputum eosinophils of 2.5% or more, or 3% or more are presented in Figures 4 and 5. Summary estimates of accuracy based on five studies (318 patients) were 0.72

(95%CI 0.24 to 0.95) for sensitivity and 0.77 (95%CI 0.20 to 0.98) for specificity, again without evidence of publication bias (Appendix).

Discussion

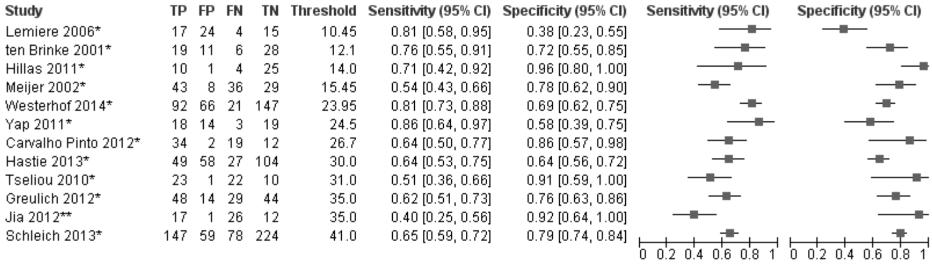
We systematically reviewed studies on the diagnostic accuracy of minimally invasive markers in detecting airway eosinophilia in asthma. In adults, FeNO, blood eosinophils, and total IgE have been extensively investigated, but their ability to distinguish between patients with and without airway eosinophilia is restricted, with summary estimates of AUC, sensitivity, and specificity never exceeding 0.8. Other markers, such as volatile organic compound analysis, were reported to be more accurate in single studies, but these results have not yet been replicated. Studies in children are scarce, but findings for FeNO and blood eosinophils are comparable with those in adults.

Several considerations deserve attention. Almost all studies showed risk of bias. These sources of bias are likely to overestimate diagnostic accuracy,[28] which would mean that the extracted accuracy estimates, although usually moderate, might be even too optimistic. Suboptimal reporting, a common phenomenon for diagnostic accuracy studies,[148] often withheld us from a proper assessment of risk of bias.

Failure to publish is a common phenomenon in diagnostic accuracy studies.[53] We aimed to reduce the risk of publication bias by searching trial registries, and by contacting authors of published studies that seemed to have data from which accuracy estimates could be calculated. This approach was successful. More than one-third of the included results were unpublished at the time of our searches. However, this approach also has its limitations. First, only a minority of diagnostic accuracy studies are registered.[69] Second, most of the included unpublished data came from studies that were, at least partially, reported and had included at least 50 patients. These studies might differ from smaller studies, or those that do not get published at all. Though we did not see any differences between accuracy estimates obtained from published and unpublished data (Appendix) and we recorded no funnel plot asymmetry (Appendix), we cannot completely exclude the possibility of reporting bias. Drivers of non-publication are unknown in diagnostic research, but it is likely that studies with lower accuracy estimates have lower chances of getting published than those with higher accuracy estimates. Should this be the case, this might have led to further overestimations of accuracy.

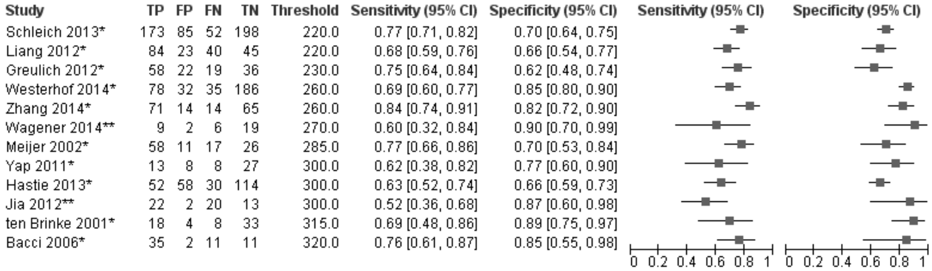
Figure 2. Forest plots for detection of sputum eosinophils of ≥3% in adults.

Figure 2a. FeNO (ppb).



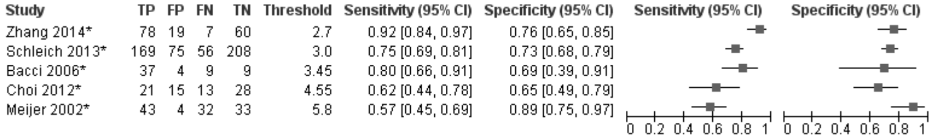
Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Lemiere 2006*	17	24	4	15	10.45	0.81 [0.58, 0.95]	0.38 [0.23, 0.55]
ten Brinke 2001*	19	11	6	28	12.1	0.76 [0.55, 0.91]	0.72 [0.55, 0.85]
Hillas 2011*	10	1	4	25	14.0	0.71 [0.42, 0.92]	0.96 [0.80, 1.00]
Meijer 2002*	43	8	36	29	15.45	0.54 [0.43, 0.66]	0.78 [0.62, 0.90]
Westerhof 2014*	92	66	21	147	23.95	0.81 [0.73, 0.88]	0.69 [0.62, 0.75]
Yap 2011*	18	14	3	19	24.5	0.86 [0.64, 0.97]	0.58 [0.39, 0.75]
Carvalho Pinto 2012*	34	2	19	12	26.7	0.64 [0.50, 0.77]	0.86 [0.57, 0.98]
Hastie 2013*	49	58	27	104	30.0	0.64 [0.53, 0.75]	0.64 [0.56, 0.72]
Tseliou 2010*	23	1	22	10	31.0	0.51 [0.36, 0.66]	0.91 [0.59, 1.00]
Greulich 2012*	48	14	29	44	35.0	0.62 [0.51, 0.73]	0.76 [0.63, 0.86]
Jia 2012**	17	1	26	12	35.0	0.40 [0.25, 0.56]	0.92 [0.64, 1.00]
Schleich 2013*	147	59	78	224	41.0	0.65 [0.59, 0.72]	0.79 [0.74, 0.84]

Figure 2b. Blood eosinophils (per µL).



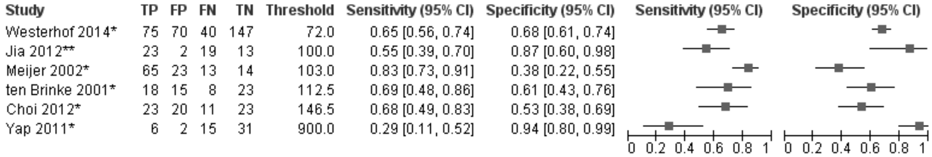
Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Schleich 2013*	173	85	52	198	220.0	0.77 [0.71, 0.82]	0.70 [0.64, 0.75]
Liang 2012*	84	23	40	45	220.0	0.68 [0.59, 0.76]	0.66 [0.54, 0.77]
Greulich 2012*	58	22	19	36	230.0	0.75 [0.64, 0.84]	0.62 [0.48, 0.74]
Westerhof 2014*	78	32	35	186	260.0	0.69 [0.60, 0.77]	0.85 [0.80, 0.90]
Zhang 2014*	71	14	14	65	260.0	0.84 [0.74, 0.91]	0.82 [0.72, 0.90]
Wagener 2014**	9	2	6	19	270.0	0.60 [0.32, 0.84]	0.90 [0.70, 0.99]
Meijer 2002*	58	11	17	26	285.0	0.77 [0.66, 0.86]	0.70 [0.53, 0.84]
Yap 2011*	13	8	8	27	300.0	0.62 [0.38, 0.82]	0.77 [0.60, 0.90]
Hastie 2013*	52	58	30	114	300.0	0.63 [0.52, 0.74]	0.66 [0.59, 0.73]
Jia 2012**	22	2	20	13	300.0	0.52 [0.36, 0.68]	0.87 [0.60, 0.98]
ten Brinke 2001*	18	4	8	33	315.0	0.69 [0.48, 0.86]	0.89 [0.75, 0.97]
Bacci 2006*	35	2	11	11	320.0	0.76 [0.61, 0.87]	0.85 [0.55, 0.98]

Figure 2c. Blood eosinophils (%).



Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Zhang 2014*	78	19	7	60	2.7	0.92 [0.84, 0.97]	0.76 [0.65, 0.85]
Schleich 2013*	169	75	56	208	3.0	0.75 [0.69, 0.81]	0.73 [0.68, 0.79]
Bacci 2006*	37	4	9	9	3.45	0.80 [0.66, 0.91]	0.69 [0.39, 0.91]
Choi 2012*	21	15	13	28	4.55	0.62 [0.44, 0.78]	0.65 [0.49, 0.79]
Meijer 2002*	43	4	32	33	5.8	0.57 [0.45, 0.69]	0.89 [0.75, 0.97]

Figure 2d. IgE (IU/mL).

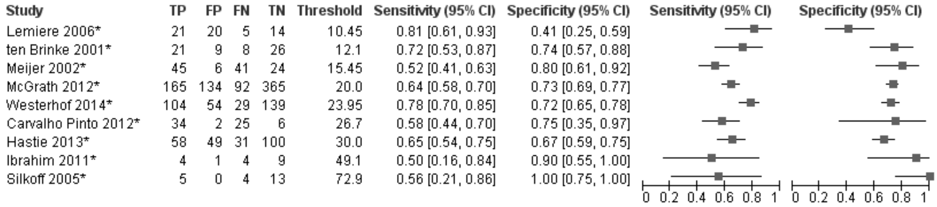


Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Westerhof 2014*	75	70	40	147	72.0	0.65 [0.56, 0.74]	0.68 [0.61, 0.74]
Jia 2012**	23	2	19	13	100.0	0.55 [0.39, 0.70]	0.87 [0.60, 0.98]
Meijer 2002*	65	23	13	14	103.0	0.83 [0.73, 0.91]	0.38 [0.22, 0.55]
ten Brinke 2001*	18	15	8	23	112.5	0.69 [0.48, 0.86]	0.61 [0.43, 0.76]
Choi 2012*	23	20	11	23	146.5	0.68 [0.49, 0.83]	0.53 [0.38, 0.69]
Yap 2011*	6	2	15	31	900.0	0.29 [0.11, 0.52]	0.94 [0.80, 0.99]

Studies are ordered by threshold. TP=true positive; FP=false positive; FN=false negative; TN=true negative. *Threshold based on optimal cut-off between sensitivity and specificity on receiver operating characteristic curve. **Threshold selection arbitrary, based on results from previous studies, or unknown.

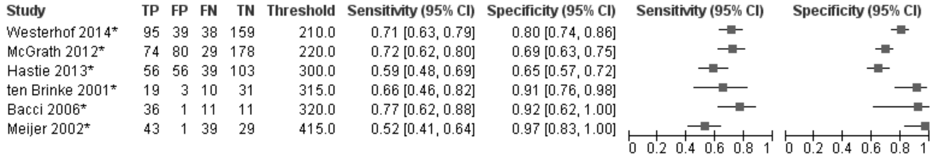
Figure 3. Forest plots for detection of sputum eosinophils of ≥2% in adults.

Figure 3a. FeNO (ppb).



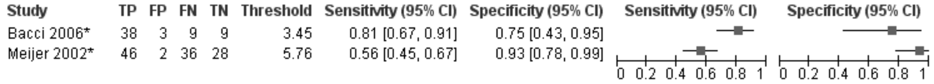
Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Lemiere 2006*	21	20	5	14	10.45	0.81 [0.61, 0.93]	0.41 [0.25, 0.59]
ten Brinke 2001*	21	9	8	26	12.1	0.72 [0.53, 0.87]	0.74 [0.57, 0.88]
Meijer 2002*	45	6	41	24	15.45	0.52 [0.41, 0.63]	0.80 [0.61, 0.92]
McGrath 2012*	165	134	92	365	20.0	0.64 [0.58, 0.70]	0.73 [0.69, 0.77]
Westerhof 2014*	104	54	29	139	23.95	0.78 [0.70, 0.85]	0.72 [0.65, 0.78]
Carvalho Pinto 2012*	34	2	25	6	26.7	0.58 [0.44, 0.70]	0.75 [0.35, 0.97]
Hastie 2013*	58	49	31	100	30.0	0.65 [0.54, 0.75]	0.67 [0.59, 0.75]
Ibrahim 2011*	4	1	4	9	49.1	0.50 [0.16, 0.84]	0.90 [0.55, 1.00]
Silkoff 2005*	5	0	4	13	72.9	0.56 [0.21, 0.86]	1.00 [0.75, 1.00]

Figure 3b. Blood eosinophils (per µL).



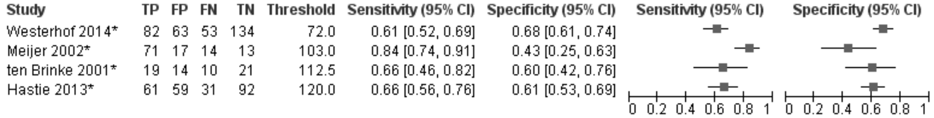
Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Westerhof 2014*	95	39	38	159	210.0	0.71 [0.63, 0.79]	0.80 [0.74, 0.86]
McGrath 2012*	74	80	29	178	220.0	0.72 [0.62, 0.80]	0.69 [0.63, 0.75]
Hastie 2013*	56	56	39	103	300.0	0.59 [0.48, 0.69]	0.65 [0.57, 0.72]
ten Brinke 2001*	19	3	10	31	315.0	0.66 [0.46, 0.82]	0.91 [0.76, 0.98]
Bacci 2006*	36	1	11	11	320.0	0.77 [0.62, 0.88]	0.92 [0.62, 1.00]
Meijer 2002*	43	1	39	29	415.0	0.52 [0.41, 0.64]	0.97 [0.83, 1.00]

Figure 3c. Blood eosinophils (%).



Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Bacci 2006*	38	3	9	9	3.45	0.81 [0.67, 0.91]	0.75 [0.43, 0.95]
Meijer 2002*	46	2	36	28	5.76	0.56 [0.45, 0.67]	0.93 [0.78, 0.99]

Figure 3d. IgE (IU/mL).



Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Westerhof 2014*	82	63	53	134	72.0	0.61 [0.52, 0.69]	0.68 [0.61, 0.74]
Meijer 2002*	71	17	14	13	103.0	0.84 [0.74, 0.91]	0.43 [0.25, 0.63]
ten Brinke 2001*	19	14	10	21	112.5	0.66 [0.46, 0.82]	0.60 [0.42, 0.76]
Hastie 2013*	61	59	31	92	120.0	0.66 [0.56, 0.76]	0.61 [0.53, 0.69]

Studies are ordered by threshold. TP=true positive; FP=false positive; FN=false negative; TN=true negative. *Threshold based on optimal cut-off between sensitivity and specificity on receiver operating characteristic curve.

Figure 4. Summary receiver operating characteristic curve for detection of sputum eosinophils of ≥3% in adults, and ≥2.5-3% in children.

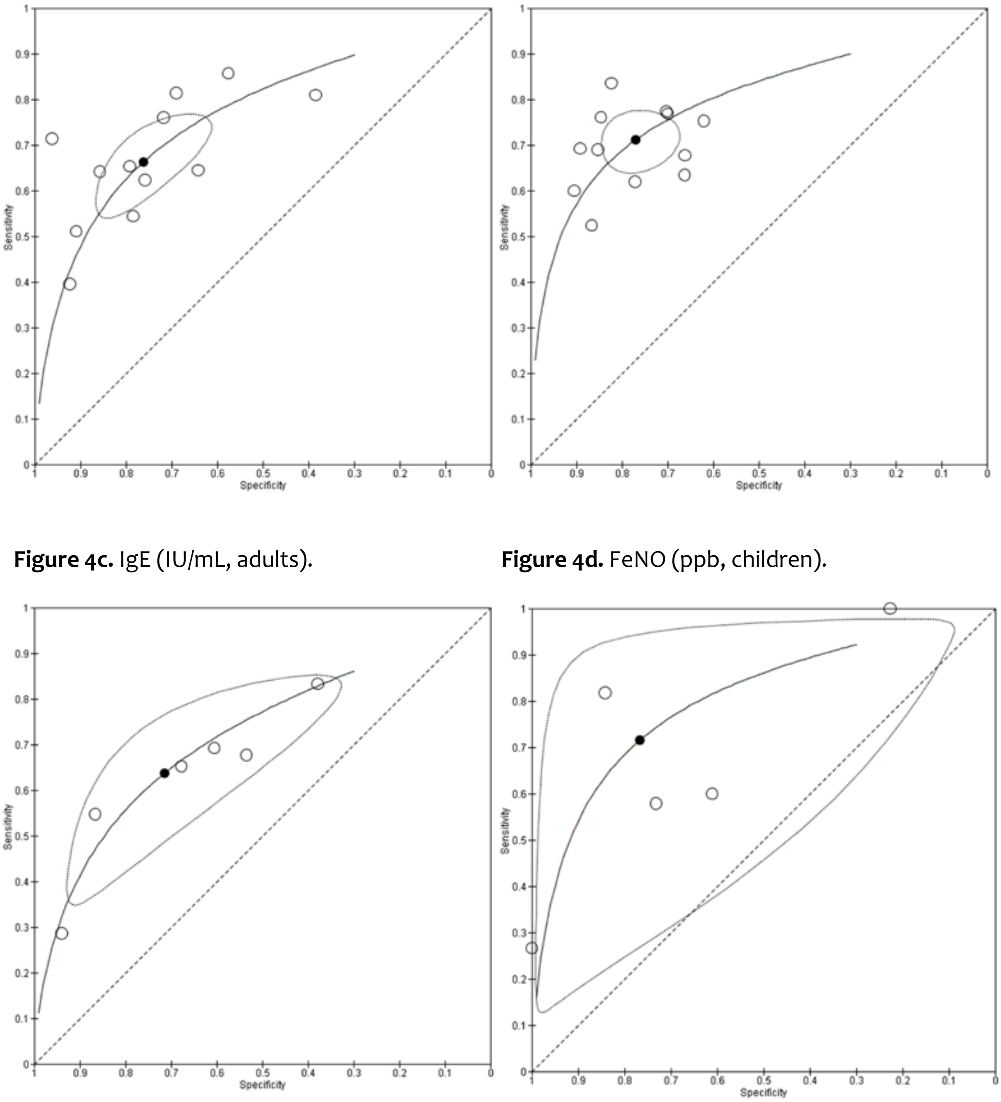


Figure 4a. FeNO (ppb, adults).

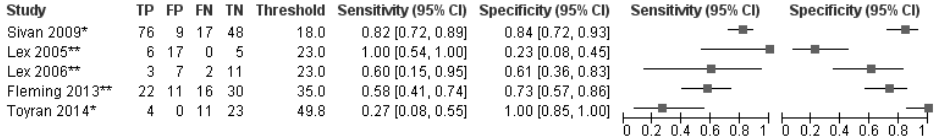
Figure 4b. Blood eosinophils (per μL, adults).

Figure 4c. IgE (IU/mL, adults).

Figure 4d. FeNO (ppb, children).

Each open circle is the result from a single study. Closed circles are summary estimates. Dotted ellipses are 95% confidence regions around summary estimates.

Figure 5. Forest plots of FeNO (ppb) for detection of sputum eosinophils of ≥2.5-3% in children.



Study	TP	FP	FN	TN	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
Sivan 2009*	76	9	17	48	18.0	0.82 [0.72, 0.89]	0.84 [0.72, 0.93]
Lex 2005**	6	17	0	5	23.0	1.00 [0.54, 1.00]	0.23 [0.08, 0.45]
Lex 2006**	3	7	2	11	23.0	0.60 [0.15, 0.95]	0.61 [0.36, 0.83]
Fleming 2013**	22	11	16	30	35.0	0.58 [0.41, 0.74]	0.73 [0.57, 0.86]
Toyran 2014*	4	0	11	23	49.8	0.27 [0.08, 0.55]	1.00 [0.85, 1.00]

Studies are ordered by threshold. TP=true positive; FP=false positive; FN=false negative; TN=true negative. *Threshold based on optimal cut-off between sensitivity and specificity on receiver operating characteristic curve. **Threshold selection arbitrary, based on results from previous studies, or unknown.

Overall, nine different definitions of airway eosinophilia were used across studies, based on different thresholds for eosinophilia in induced sputum, bronchoalveolar lavage, and endobronchial biopsy. These three airway compartments do not show strong correlations with regard to eosinophil counts.[210] Although the diagnostic accuracy of markers can vary across different eosinophilia definitions, we noted that the summary AUCs were stable when comparing studies using any definition of airway eosinophilia, sputum eosinophils of 3% or more, or sputum eosinophils of 2% or more. There was also substantial heterogeneity in the study population and test methods. Some studies only included smokers, for example, whereas others only included non-smokers, and at least four different FeNO measurement devices were used. Many studies analyzed both patients with childhood and adult onset asthma. In the latter group, distinguishing asthma from Chronic Obstructive Pulmonary Disease (COPD) and asthma-COPD overlap syndrome can be problematic. The accuracy of markers can vary across these different subgroups. The prevalence of airway eosinophilia also differed substantially across studies. Diagnostic accuracy typically varies with clinical setting, context, and prevalence. Although the results from the individual studies show substantial heterogeneity, we felt it was safe to draw conclusions because AUCs for FeNO, blood eosinophils, and IgE consistently reflected moderate accuracy.

Combining markers with other clinical features in a prediction model is likely to improve diagnostic accuracy compared with single markers, but this has not been sufficiently investigated yet. All but three studies only reported on accuracy estimates of single markers. Since we did not have individual patient data, we were unable to further analyze the incremental value of combining markers.

The most robust evidence for the clinical value of detecting airway eosinophilia comes from a Cochrane review that showed the frequency of asthma exacerbations can be significantly reduced when tailoring inhaled corticosteroids on sputum eosinophilia.[171] For a marker to be able to replace induced sputum in this context, sensitivity and specificity should probably be at least 90%, so that at most 10% of all patients will be misclassified and, potentially, subjected to inappropriate clinical decisions. Our analysis shows that there are no single markers available with a large enough documented accuracy to fulfil these criteria. However, recent guidelines recommending the use of sputum eosinophil counts in severe asthma, acknowledge that the quality of evidence is very low.[32,172] Additionally, they do not recommend sputum-guided treatment in the general asthma population. Some of the markers assessed in this review on their own might have better potential in managing asthma than sputum eosinophil counts. This fact is shown by a recent study in which volatile organic compound analysis predicted corticosteroid responsiveness with greater accuracy than sputum eosinophils,[211] and by another study that showed good response to mepolizumab in patients with severe eosinophilic asthma as measured by blood eosinophils.[212] The latter study draws attention to the accumulating evidence for the potential role of blood eosinophils as a predictor of responsiveness to novel targeted therapies against eosinophilic airway inflammation.[175]

Moderate accuracy does not necessarily make the investigated markers useless. Markers can also be applied in a triage setting, for example, for ruling-out (high sensitivity required) or ruling-in (high specificity required) airway eosinophilia. In the case of high specificity, for example, those with a positive test result would be considered as eosinophilic, whereas those with a negative test result would need to undergo further testing (e.g., sputum induction) because of a large number of false negatives due to a low sensitivity. Most included studies only reported on the optimal cut-off between sensitivity and specificity, based on the Youden's index. When a marker is not sufficiently accurate to replace the existing test, this optimal cut-off is clinically not very practical because both sensitivity and specificity are typically suboptimal at this cut-off. Therefore, it does not inform the reader about the ability of the marker to rule-in or rule-out airway eosinophilia. Furthermore, data-driven selection of an optimal cut-off leads to overoptimistic estimates of sensitivity and specificity.[28] It could be more informative to report on sensitivity at a fixed high specificity (e.g., 95%), or the other way around.

An American Thoracic Society guideline on the interpretation of FeNO for clinical applications strongly "recommends the use of FeNO in the diagnosis of eosinophilic airway inflammation".[213] It also strongly "recommends that low FeNO less than 25 ppb (20 ppb in children) be used to indicate that eosinophilic inflammation and responsiveness to corticosteroids are less likely", "that FeNO greater than 50 ppb (35 ppb in children) be used to indicate that eosinophilic inflammation and, in symptomatic patients, responsiveness to corticosteroids are likely", and "that FeNO values between 25 ppb and 50 ppb (20-35 ppb in children) should be interpreted cautiously and with reference to the clinical context". Our results challenge this concept. At FeNO thresholds below 25 ppb sensitivity ranges from 0.52 to 0.86 in adults (Appendix). This finding means that of every 100 patients with asthma with airway eosinophilia tested by FeNO, up to 48 would be falsely considered as not having airway eosinophilia, and effective treatment might be withheld from them. In children, sensitivity for FeNO thresholds below 20 ppb ranges from 0.75 to 0.82, showing that up to 25 of every 100 patients with airway eosinophilia would be false negatives. Although these thresholds might be relevant in specific subgroups of asthma, these findings show that FeNO results should be interpreted with much more caution in the general asthma population than recommended by the American Thoracic Society.

It is not surprising that the markers assessed in our review generally were moderately accurate. The underlying biological mechanisms determining airway eosinophil counts are substantially different from those of some of the investigated markers.[214] Several studies also showed significant variability in blood eosinophils[215] and IgE[216] in the same patients with asthma over short periods of time. Some patients with asthma were shown to have persistently raised FeNO concentrations, not suppressed by corticosteroid treatment, and not reflecting raised sputum eosinophils.[217] Corticosteroid treatment significantly affects FeNO, blood eosinophils, IgE, and sputum eosinophils,[218] but the relative magnitude of this effect could vary across markers. Diagnostic accuracy might therefore be affected by treatment status. Also, many other factors, such as age, sex, reflux disease, smoking, and atopy, have been shown to affect FeNO concentrations.[219] This effect might also be the case with other markers and further compromises the identification of an accurate minimally invasive test for airway eosinophilia.

Similar reproducibility problems could apply to the reference standard and target condition. Although some studies showed that a threshold of 3% for sputum eosinophils is reproducible over time,[220] others found the phenotypic classification of asthma to change frequently, both spontaneously and in response to treatment.[221] Longitudinal studies that examined sputum cell counts in successive exacerbations found substantial heterogeneity in the type of inflammation within the same individuals.[222] Consequently, a diagnosis of eosinophilic asthma based on a single sputum sample might be questionable.

Based on our findings, we discourage the use of FeNO, blood eosinophils, or IgE as single surrogate tests for detecting airway eosinophilia in asthma. Our meta-analyses show that, at the optimal cut-off, sensitivities and specificities of these markers for detecting sputum eosinophilia are moderate, and their use would lead to many false positives or false negatives. Future research will mainly need to focus on whether these markers can be applied as rule-in or rule-out tests, whether markers that were poorly investigated or clinical prediction models incorporating many markers together with other clinical data are more accurate, perhaps in specific settings or subgroups, and whether these markers on their own merits have potential in managing asthma.[218] A next step could be an extensive individual patient data project, combining existing datasets from observational asthma studies in which both clinical features, minimally invasive markers, and one or more reference standards for airway eosinophilia were assessed. Thresholds for ruling-in and ruling-out airway eosinophilia based on individual markers can then be reliably defined, and an optimal multivariable clinical prediction model can be developed. The clinical value of these findings can subsequently be investigated in terms of, for example, response to therapy or the reduction of exacerbations.

Acknowledgments

We thank the following collaborators who kindly shared their data or provided requested information: Joseph R. Arron, Elena Bacci, Maarten van den Berge, Anneke ten Brinke, Andrew Bush, Shaoxi Cai, Regina M. De Carvalho-Pinto, Hun Soo Chang, Timm Greulich, John V. Fahy, Ana Luisa G. Fernandes, Louise Fleming, Stephen J. Fowler, Peter G. Gibson, Annette T. Hastie, Baharudin Ibrahim, Huib A.M. Kerstjens, Catherine Lemière, Stelios Loukides, Kelly W. McGrath, Choon Sik Park, Ian D. Pavord, Florence N. Schleich, Antonius Schneider, Michael D. Shields, Philip E. Silkoff , Yakov Sivan, Muge Toyran, Elaine Yap, and Irene S.L. Zeng.

Chapter 12

Biomarkers to identify sputum eosinophilia in different adult asthma phenotypes

Guus A. Westerhof
Daniël A. Korevaar
Marijke Amelink
Selma B. de Nijs
Jantina C. de Groot
Junfeng Wang
Els J. Weersink
Anneke ten Brinke
Patrick M. Bossuyt
Elisabeth H. Bel

European Respiratory Journal 2015;46(3):688-96

Appendices and supplemental material are available online at:
http://erj.ersjournals.com/content/46/3

Abstract

Background

Several biomarkers have been used to assess sputum eosinophilia in asthma. It has been suggested that the diagnostic accuracy of these biomarkers might differ between asthma phenotypes. We investigated the accuracy of biomarkers in detecting sputum eosinophilia (⩾3%) in different adult asthma phenotypes.

Methods

Levels of eosinophils in blood and sputum, fraction of exhaled nitric oxide (FeNO) and total immunoglobulin E (IgE) from 336 consecutive adult patients, enrolled in three prospective observational clinical trials and recruited at five pulmonology outpatient departments, were analyzed. Areas under the receiver operating characteristic curves (AUC) for detecting sputum eosinophilia were calculated and compared between severe and mild, obese and non-obese, atopic and non-atopic, and (ex-)smoking and never-smoking asthma patients.

Results

Sputum eosinophilia was present in 116 patients (35%). In the total group the AUC was 0.83 (95%CI 0.78 to 0.87) for blood eosinophils, 0.82 (95%CI 0.77 to 0.87) for FeNO and 0.69 (95%CI 0.63 to 0.75) for total IgE. AUCs were similar for blood eosinophils and FeNO between different phenotypes. Total IgE was less accurate in detecting sputum eosinophilia in atopic and obese patients than in non-atopic and non-obese patients.

Conclusions

Blood eosinophils and FeNO had comparable diagnostic accuracy (superior to total IgE) in identifying sputum eosinophilia in adult asthma patients, irrespective of asthma phenotype such as severe, non-atopic, obese, and smoking-related asthma.

Introduction

Eosinophilic airway inflammation is an important distinguishing characteristic of specific adult asthma phenotypes.[223] To assess this type of airway inflammation, sputum eosinophil counts are generally considered to be the gold standard.[224] Treatment guided by sputum eosinophils reduces the frequency of asthma exacerbations[171] and patients with sputum eosinophilia have a better response to inhaled corticosteroids with respect to reducing airway hyperresponsiveness, decreasing asthma symptoms, and improving quality of life compared to those without.[225,226] Not surprisingly, the recent European Respiratory Society/American Thoracic Society guidelines on severe asthma recommend sputum eosinophils counts combined with clinical criteria to guide asthma therapy.[172] Unfortunately, sputum induction and differential sputum cell counts are only feasible in specialized clinics, are not always successful, and do not give immediate results.[227]

Several alternative methods of assessing airway eosinophilia have been proposed in the literature, including non-invasive biomarkers such as the fraction of exhaled nitric oxide (FeNO),[181,186,228] peripheral blood eosinophil counts,[186,194] and total immunoglobulin E (IgE),[186] with varying diagnostic accuracy. However, specific patient characteristics that distinguish between different adult asthma phenotypes such as asthma severity,[229] obesity,[230] atopy,[231] and (ex-)smoking status[232] may influence both airway and systemic inflammation. Therefore, the accuracy of biomarkers to assess sputum eosinophilia may vary between these different asthma phenotypes.

The aim of the present study was to evaluate the diagnostic accuracy of FeNO, blood eosinophils, and total IgE for detecting sputum eosinophilia, defined as $\geqslant$3%[191,233] in a large heterogeneous group of adult asthma patients, as well as in patients with different asthma phenotypes.

Methods

Patients

We collected data from 571 patients with adult-onset asthma (onset of asthma after the age of 18 years) who had been included in three separate observational clinical trials (Netherlands Trial Register numbers: NTR2217, NTR1846, and NTR1838)[234,235] between 2009 and 2012. These prospective trials aimed at phenotyping patients with adult-onset asthma based on an extensive set of clinical, functional, and inflammatory parameters. Patients aged $\geqslant$18 years were eligible if they had a confirmed diagnosis of asthma based on international guidelines (history of variable respiratory symptoms and documented variable expiratory airflow limitation).[32] Patients with other pulmonary diseases, unrelated major comorbidities, pregnancy, or a smoking history of >10 pack-years combined with fixed airflow obstruction/reduced diffusion capacity were excluded. Detailed inclusion and exclusion criteria have been reported elsewhere (Amelink et al.,[234] Westerhof et al.,[235] and NTR2217). All eligible patients visiting the pulmonology outpatient department of four secondary and one tertiary referral clinic in the Netherlands were invited to participate. All three trials were reviewed and approved by medical ethical boards before their initiation. All patients gave informed consent. The present additional analysis was registered at the Netherlands Trial Register (NTR4589).

Assessment of specific phenotypic characteristics

Asthma severity

Asthma severity was assessed according to the Innovative Medicines Initiative criteria,[236] based on medication use and degree of asthma control. Severe asthma was defined by the use of $\geqslant$1000 μg·day^{-1} fluticasone equivalent and/or daily oral corticosteroids plus a second controller, combined with an asthma control score >1.5 on the Juniper et al. asthma control questionnaire[237,238] or at least two exacerbations in the past 12 months. Patients who did not fulfil these criteria were considered as having mild-to-moderate asthma.

Obesity

Obesity was defined as a body mass index (BMI) $\geqslant$30 kg·m^{-2}.

Atopy

Specific IgE to common aeroallergens was measured by immunoCAP; atopy was defined as specific IgE >0.35 kU·L^{-1} for at least one allergen.

Smoking status

Smoking status was recorded during history taking. (Ex-)smokers were either current or previous smokers. Non-smokers were patients who had never smoked.

Reference standard: sputum eosinophils

Sputum induction was performed according to internationally accepted standards by trained lung function analysts.[239] All patients inhaled a nebulized saline solution for 5 minutes; if possible this was repeated up to three times. Sputum processing was performed according to full sample method and differential cell counts were analyzed on cytospin preparations. Results for different sputum cell types are presented as percentage of total non-squamous cell count. Laboratory analyses were performed blinded to patient characteristics and index test results.

Index tests: FeNO, blood eosinophils and total IgE

FeNO (index test 1) was measured using a portable rapid-response chemoluminescent analyser (flow rate 50 $mL \cdot s^{-1}$; NIOX System, Aerocrine, Sweden). FeNO results are reported as parts per billion (ppb).[213] Venous blood was collected and differential white blood cells counts were performed. Absolute blood eosinophil numbers (index test 2) are reported as 109 $cells \cdot L^{-1}$. Total IgE (index test 3) was measured using immunoCAP tests and reported as $kU \cdot L^{-1}$. All measurements in blood samples were performed by the general laboratories of the participating hospitals, which were blinded to the outcome of other tests. All data were collected in one or two visits <2 weeks apart.

Statistical analysis

Adequate sputum samples from 336 patients were available (Supplementary Figure E1, available online) and these patients were included in the analyses of diagnostic accuracy. Baseline characteristics between patients with and without adequate sputum were compared. Patients with missing data on blood esoinophils, FeNO, or total IgE were excluded for the analysis of that index test.

Receiver operating characteristic curve (ROC) analysis was used to evaluate the diagnostic accuracy of FeNO, blood eosinophils, total IgE, and their combinations to identify sputum eosinophilia ⩾3%. This was done first in the complete group and thereafter in subgroups with specific phenotypic patient characteristics as described above. Analyses included the following: 1) area under the ROC curve (AUC) (95%CI) for the different biomarkers (FeNO, blood eosinophils, and total IgE); 2) sensitivity (95%CI) and corresponding threshold of each biomarker at a specificity of ⩾95%; and 3) specificity (95%CI) and corresponding threshold of each biomarker at a sensitivity of ⩾95%. McNemar's test was used to compare sensitivities and specificities between biomarkers. DeLong tests were used to compare AUCs between different asthma phenotypes and to evaluate whether combinations of any of the three biomarkers improved the diagnostic accuracy of each single biomarker.

We also developed a multivariate logistic regression model for the prediction of sputum eosinophilia ⩾3% based on phenotypic features and the three markers. First, we evaluated whether patient characteristics (age, sex, BMI, asthma duration, race, smoking status, post-bronchodilator forced expiratory volume in 1 s (FEV_1), post-bronchodilator FEV_1/forced vital capacity (FVC) ratio, atopy status, and medication use (high dose versus low dose)) were significantly associated with sputum eosinophilia in a univariate analysis ($p<0.20$). With the significant characteristics we then built a multivariable logistic model. We then used a stepwise procedure to arrive at a parsimonious model by removing in each step the variable with the smallest Wald statistic, until further removal would lead to a significant loss in goodness-of-fit ($p<0.05$; likelihood-ratio test). Then, the three markers were added to the resulting multivariable model, and the stepwise procedure was repeated.

Data were analyzed using SPSS version 22 (IBM, Armonk, NY, USA) and R version 3.0 (R Foundation for Statistical Computing, Vienna, Austria).

Table 1. Baseline characteristics of patients who provided an adequate sputum sample.

Characteristic	
Total subjects (n)	336
Age (years)	53 (SD 13)
Female (%)	55
BMI (kg·m^{-2})	28 (SD 5)
Age at asthma onset (mean years)	45 (SD 15)
Asthma duration (median years)	3 (IQR 0-10)
Current or ex-smoker (%)	54
Pack-years (median years)	1 (IQR 0-13)
ICS fluticasone equivalent (μg)	500 (IQR 250-500)
ACQ score	1.3±0.8
Atopy (%)	32
Nasal polyposis (%)	19
Post-bronchodilator FEV_1 (% predicted)	97±18
Post-bronchodilator FEV_1/FVC (% predicted)	93±12
FeNO (ppb)	23 (13-43)
Total IgE (kU·L^{-1})	56 (18-216)
Blood neutrophils (×10^9 cells·L^{-1})	4.3±1.7
Blood eosinophils (×10^9 cells$^{\cdot L-1}$)	0.2 (0.1-0.3)
Sputum neutrophils (%)	66.3 (45.4-82.3)
Sputum eosinophils (%)	0.8 (0.1-6.6)

BMI=body mass index; ICS=inhaled corticosteroid; ACQ=asthma control questionnaire; FEV_1=forced expiratory volume in 1 s; FVC=forced vital capacity; ppb=parts per billion.

Results

Table 1 shows the baseline characteristics of the 336 patients who were included in the analyses. Compared to these patients, the excluded patients (n=235) were younger, more often female, and had slightly lower blood eosinophils (Supplementary Table E1). Sputum eosinophilia was present in 116 (35%) patients. FeNO, blood eosinophils, and total IgE were missing in ten, five, and four of the included patients, respectively. Correlations of the three biomarkers with sputum eosinophils are shown in Supplementary Figure E2.

Diagnostic accuracy of biomarkers

In the complete group as well as in the eight subgroups, FeNO and blood eosinophils had similar diagnostic accuracy, whereas the AUC for total IgE was significantly lower (Tables 2 and 3). Combining FeNO and blood eosinophils significantly improved diagnostic accuracy compared to FeNO alone (p=0.001) or blood eosinophils alone (p=0.027) (AUC 0.87 (95%CI 0.83 to 0.91); Tables 2 and 3, Figures 1 and 2). Adding total IgE to the combination of FeNO and blood eosinophils did not significantly improve the AUC (0.87; p=0.732). Total IgE performed significantly better in obese than in non-obese patients, and in non-atopic compared to atopic patients, respectively (Table 3 and Figure 2).

A multivariable logistic model was created and reduced using stepwise backward selection. The final model included age, sex, FEV$_1$/FVC, pulmonary medication (high or low inhaled corticosteroid dose), FeNO, and blood eosinophils (Supplementary Table E2). This model further improved the diagnostic accuracy to a minimal extent compared to FeNO and blood eosinophils combined (AUC 0.89 (95%CI 0.85 to 0.93); p=0.041).

Sensitivity, specificity and biomarker thresholds

Table 2 shows the sensitivity and specificity for each biomarker at either a high specificity or sensitivity and the associated threshold of this marker; Supplementary Figure E3 shows the formula to calculate the probability of sputum eosinophilia for the combined model of FeNO and blood eosinophils.

At a sensitivity of ⩾95% (i.e., low number of false negatives), FeNO, blood eosinophils, and total IgE had a comparable specificity, whereas the specificity of FeNO and blood eosinophils combined was significantly higher. Negative predictive values ranged between 0.92 and 0.94 for biomarker values below the corresponding thresholds.

Table 2. Diagnostic accuracy of the biomarkers in the complete group.

	AUC (95%CI)	Positivity threshold	Sensitivity (95%CI)	Specificity (95%CI)	PPV (95%CI)	NPV (95%CI)	TP	FP	FN	TN
FeNO	0.82 (0.77-0.87)	≥12.2 ppb	0.96 (0.90-0.99)	0.28 (0.22-0.34)	0.41 (0.35-0.47)	0.92 (0.82-0.97)	108	154	5	59
		≥64.5 ppb	0.39 (0.30-0.49)	0.95 (0.92-0.98)	0.81 (0.68-0.90)	0.75 (0.69-0.80)	44	10	69	203
Blood eosinophils	0.83 (0.78-0.87)	≥0.09 $\times 10^9$ cells·L^{-1}	0.96 (0.90-0.99)	0.26 (0.20-0.33)	0.40 (0.34-0.46)	0.92 (0.81-0.97)	108	161	5	57
		≥0.41 $\times 10^9$ cells·L^{-1}	0.36 (0.27-0.46)	0.95 (0.91-0.97)	0.79 (0.65-0.88)	0.74 (0.69-0.79)	41	11	72	207
Total IgE	0.69 (0.63-0.75)	≥13.5 kU·L^{-1}	0.96 (0.90-0.99)	0.28 (0.22-0.34)	0.41 (0.35-0.47)	0.92 (0.82-0.97)	110	157	5	60
		≥763.5 kU·L^{-1}	0.08 (0.04-0.14)	0.95 (0.92-0.98)	0.47 (0.25-0.71)	0.66 (0.61-0.71)	9	10	106	207
FeNO + blood eosinophils	0.87 (0.83-0.91)	≥0.095[a]	0.95 (0.90-0.99)	0.39 (0.32-0.46)	0.45 (0.39-0.52)	0.94 (0.86-0.98)	106	129	5	82
		≥0.70[a]	0.46 (0.36-0.56)	0.95 (0.91-0.98)	0.84 (0.71-0.91)	0.77 (0.71-0.82)	51	10	60	201

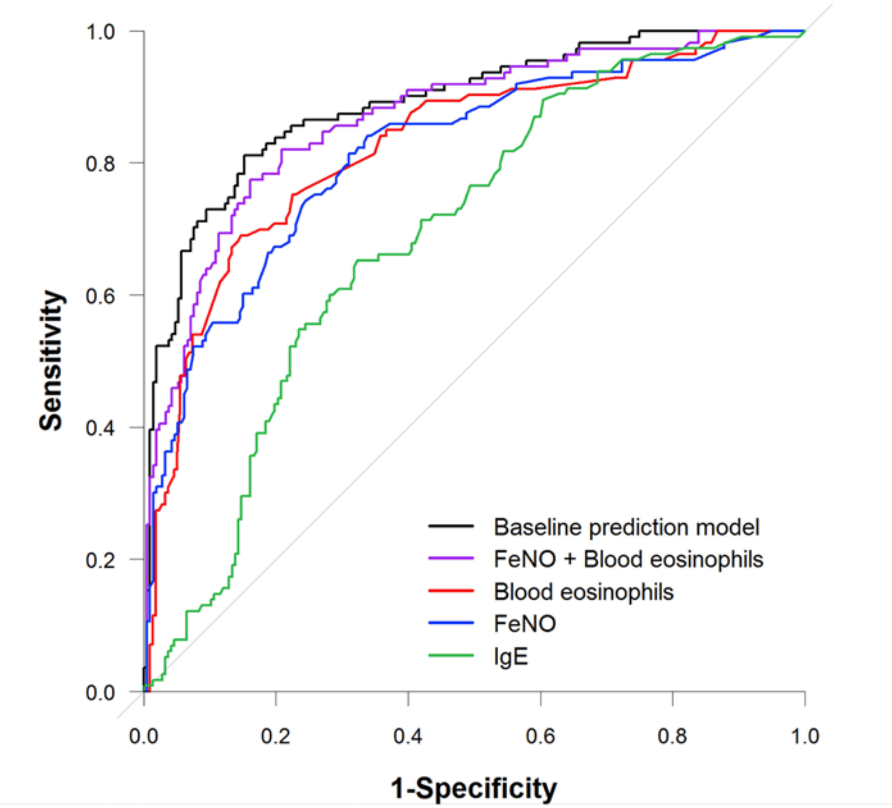
Data are presented as n, unless otherwise stated. n=336. [a]These values correspond to an individual's probability of sputum eosinophilia, as determined by the formula provided in Supplementary Figure E3. TP=true positive; FP=false positive; FN=false negative; TN=true negative; PPV=positive predictive value; NPV=negative predictive value.

Table 3. Diagnostic accuracy in in patients with different asthma phenotypes.

	Obesity			**Atopy**			**Asthma severity**			**Smoking status**		
	Non-obese	**Obese**	***p-value***	**Non-atopic**	**Atopic**	***p-value***	**Mild-moderate**	**Severe**	***p-value***	**Never-smoker**	**Ex- or current smoker**	***p-value***
Eosinophilia (<3%/≥3%)	154/82	66/34	0.90	153/74	67/42	0.28	161/58	58/57	<0.01	103/51	117/65	0.62
FeNO	0.83 (0.77-0.88)	0.78 (0.68-0.89)	0.46	0.83 (0.77-0.89)	0.78 (0.69-0.88)	0.40	0.81 (0.74-0.88)	0.83 (0.75-0.91)	0.67	0.84 (0.77-0.90)	0.81 (0.73-0.88)	0.52
Blood eosinophils	0.83 (0.77-0.89)	0.82 (0.73-0.91)	0.82	0.83 (0.77-0.89)	0.83 (0.74-0.91)	0.99	0.82 (0.76-0.89)	0.80 (0.72-0.89)	0.73	0.86 (0.79-0.93)	0.80 (0.73-0.87)	0.23
IgE	0.73 (0.67-0.80)	0.59 (0.47-0.70)	0.03	0.75 (0.68-0.82)	0.57 (0.46-0.68)	<0.01	0.68 (0.61-0.76)	0.66 (0.56-0.76)	0.70	0.64 (0.55-0.73)	0.74 (0.66-0.81)	0.13
FeNO + Blood eosinophils	0.88 (0.83-0.93)	0.85 (0.76-0.93)	0.55	0.88 (0.82-0.92)	0.85 (0.77-0.93)	0.63	0.86 (0.81-0.92)	0.85 (0.78-0.92)	0.81	0.88 (0.82-0.94)	0.86 (0.80-0.92)	0.64

Data are presented as n or AUC (95%CI), unless otherwise stated. AUC (95%CI) is given per biomarker in every subgroup. The difference between the AUCs for every biomarker is compared within the separate subgroups and the result depicted as p-value.

Figure 1. Receiver operating characteristic curves.



At a specificity of ⩾95% (i.e., low number of false positives), sensitivities for FeNO, blood eosinophils, and their combination did not significantly differ, but the sensitivity of total IgE was significantly lower compared to the other biomarkers. The positive predictive values of FeNO, blood eosinophils, and their combination ranged from 0.79 to 0.84, but was only 0.47 for total IgE.

With these thresholds (Table 2), the biomarkers can be used in up to half of the patients, as they had test results below the lower threshold or above the upper threshold: 47% for FeNO and blood eosinophils combined (150 out of 322 patients), 36% for FeNO (117 out of 326 patients), 34% for blood eosinophils (113 out of 331 patients) and 25% for total IgE (83 out of 332 patients).

Thresholds for the separate biomarkers in different phenotypes are summarized in Table 4; details are shown in Supplementary Tables E3 to E10. Across subgroups, thresholds were relatively stable for FeNO and the FeNO/blood eosinophils combination model, but varied considerably for the upper levels of blood eosinophils and IgE.

Figure 2. Receiver operating characteristic curves for varying asthma phenotypes.

Figure 2a. FeNO.

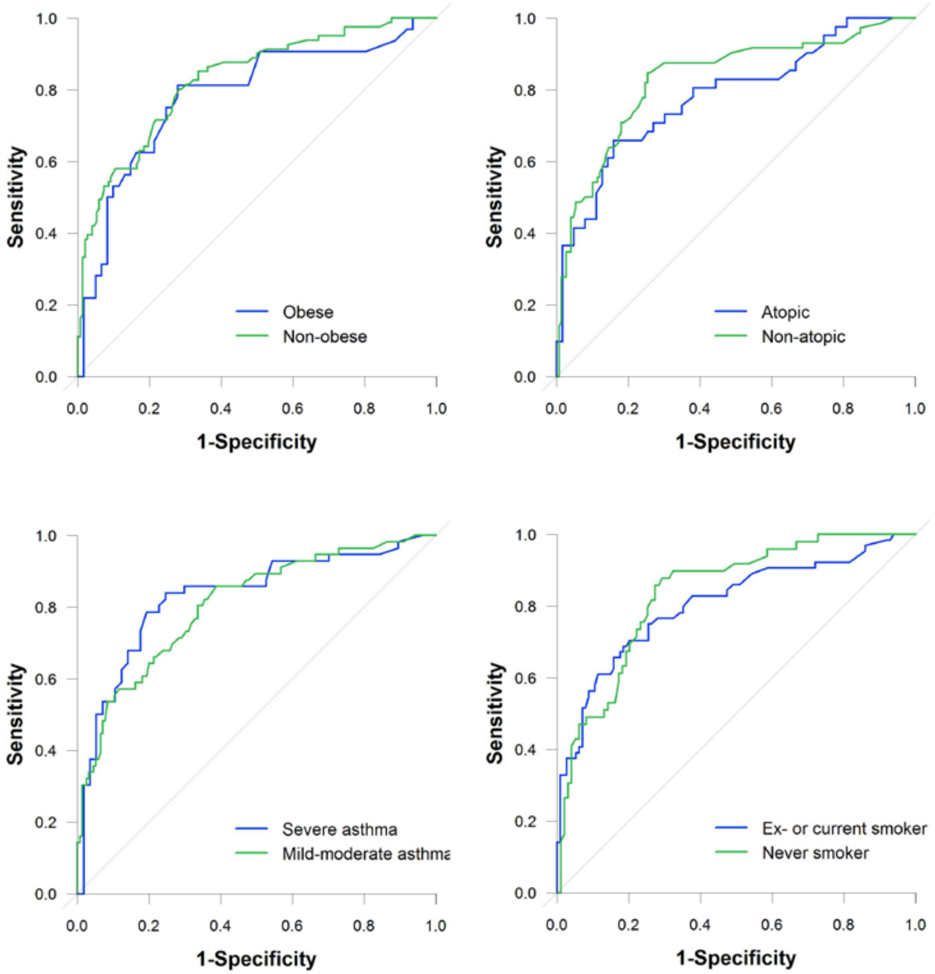


Figure 2. *Continued.*

Figure 2b. Blood eosinophils.

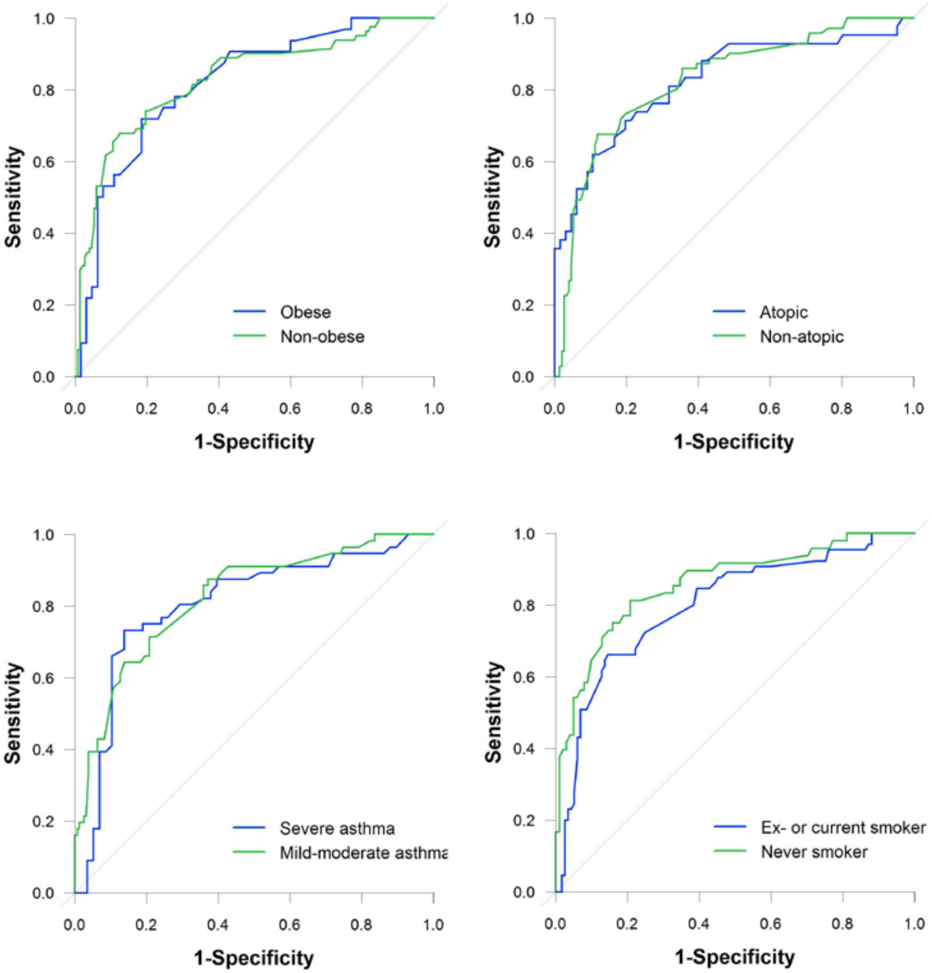


Figure 2. *Continued.*

Figure 2c. Total IgE.

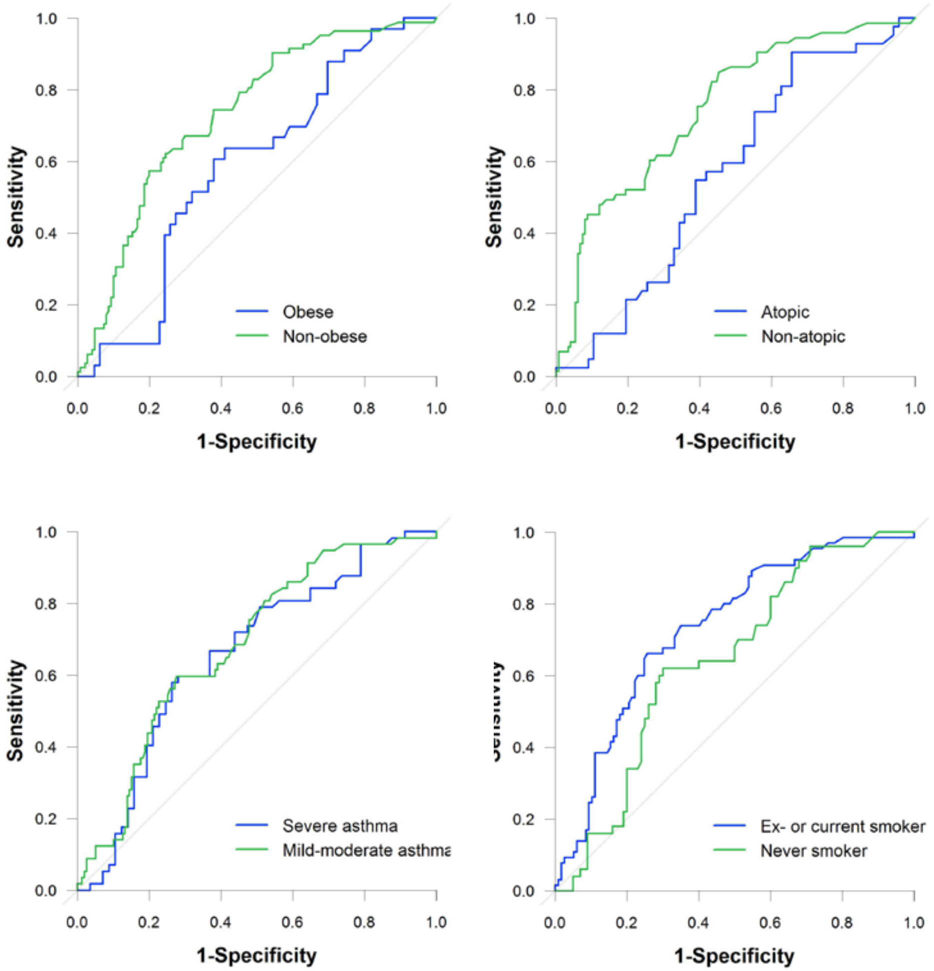


Figure 2. *Continued.*

Figure 2d. Combined model (FeNO and blood eosinophils).

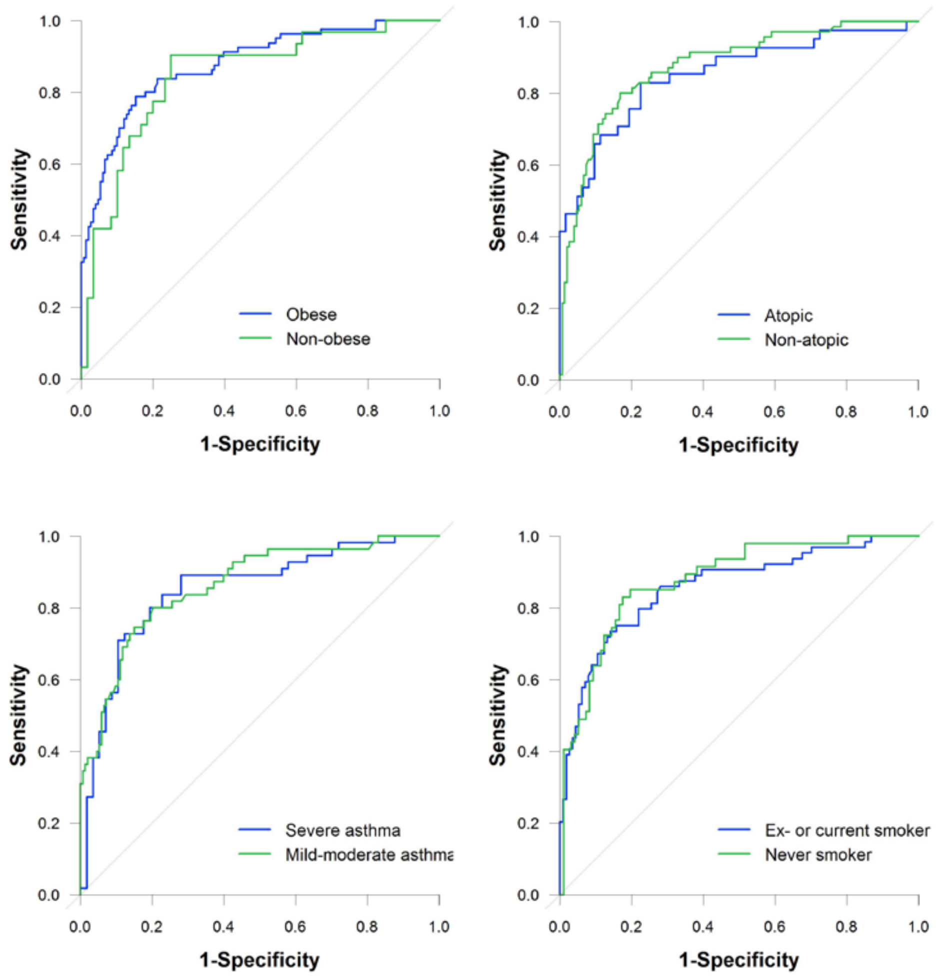


Table 4. Distribution of marker thresholds at 95% sensitivity and specificity in different asthma phenotypes.

	Range lower threshold, sensitivity ≥95%	Range upper threshold, specificity ≥95%
FeNO (ppb)	8.6-15.1	48.5-69.5
Blood eosinophils ($\times 10^9$ cells·L^{-1})	0.06-0.095	0.34-0.73
Total IgE (kU·L^{-1})	8.5-25.5	389-2181
FeNO + blood eosinophils	0.086-0.138[a]	0.656-0.75[a]

[a]These values correspond to an individual's probability of sputum eosinophilia, as determined by the formula provided in Supplementary Figure E3.

Discussion

This study shows that the diagnostic accuracy of FeNO and blood eosinophils to detect sputum eosinophilia did not significantly differ between obese and non-obese, atopic and non-atopic, (ex-)smoking and never-smoking, and severe and mild-to-moderate asthma patients. Total IgE was less accurate in atopic and obese patients than in non-atopic and non-obese patients. In unselected adult-onset asthma patients the diagnostic accuracy of FeNO and blood eosinophils is superior to that of total IgE, while combining FeNO and blood eosinophils into one model improves the overall diagnostic accuracy. The results suggest that FeNO and blood eosinophils (but not total IgE) can be used to confirm or exclude sputum eosinophilia with high certainty in up to half of adult asthma patients, irrespective of asthma phenotype.

The present study is the first to compare the diagnostic accuracy of FeNO, blood eosinophils, total IgE, and their combinations between different adult asthma phenotypes. Previous studies have mainly investigated the diagnostic accuracy of these biomarkers in general asthma populations. Our findings in the total study group of asthma patients on FeNO, blood eosinophils, and total IgE are in line with the results of these previous studies, which we recently summarized in a systematic review,[240] in which we found a pooled AUC of 0.75 for FeNO, 0.78 for blood eosinophils, and 0.65 for total IgE. Our findings on FeNO and blood eosinophils are more promising than those of two other recent reports in which the authors concluded that FeNO and blood eosinophils lack sufficient sensitivity or specificity to be useful as markers of sputum eosinophilia.[170,186] In addition, we developed a combination model of FeNO and blood eosinophils, which increased the diagnostic accuracy significantly compared to the separate markers. Adding four clinical variables to the model further increased the AUC, although only to a very minimal extent. For clinical purposes the use of two variables is obviously more practical.

The diagnostic accuracy of FeNO and blood eosinophils in detecting sputum eosinophilia was similar in the different asthma phenotypes. This may be surprising, since remarkable differences in airway eosinophilia and its associated cytokines and markers have been described in specific asthma subgroups; for example, between obese and non-obese asthma patients.[241] One study showed more eosinophils in the airway submucosa than in the airway lumen of obese patients with asthma, and higher levels of interleukin-5 in bronchoalveolar lavage fluid.[230] Apparently, only a subset of obese asthma patients with eosinophilic airway inflammation shows sputum eosinophilia. In our study, total IgE was relatively more accurate in predicting sputum eosinophilia in non-obese patients compared to obese patients, but had lower diagnostic accuracy than the other two biomarkers. Discordance between different biomarkers for airway eosinophilia has been reported previously.[191,218] More interestingly, discordance between various biomarkers of the effects of anti-inflammatory therapy or ability to predict asthma attacks have also been noted.[218,227,242] These data suggest that discordance between biomarkers in different asthma phenotypes may point towards different underlying mechanisms.

There was no significant difference in diagnostic accuracy of FeNO and blood eosinophils between atopic and non-atopic patients. One previous study showed lower diagnostic accuracy for FeNO in non-atopic patients than in atopic patients.[181] The discrepancy between these results and ours could be due to differences in patient characteristics or the devices used to measure FeNO. The higher diagnostic accuracy of total IgE in non-atopic patients compared to atopic patients might be related to different underlying mechanisms. While eosinophilia in classical atopic asthma is likely to be T-helper (Th)2-cell driven and includes higher basal IgE production, in non-atopic asthma there is accumulating evidence that activation of eosinophils might be mediated by alternative pathways.[231]

Patients with severe asthma often show discrepancies between airway and blood eosinophilia, which is probably explained by their high doses of inhaled or oral corticosteroid treatment. We did not find a difference in the diagnostic accuracy of blood eosinophils or FeNO between mild-to-moderate and severe asthma patients, but previous studies have found conflicting results. One study found an AUC of blood eosinophils of 0.55 in corticosteroid-treated patients and of 0.73 in untreated patients,[192] whereas these numbers were 0.75 and 0.62, respectively, in another study.[186] Three previous studies evaluated the accuracy of FeNO among severe/treated and mild/untreated asthma patients.[181,186,197] None of them found considerable differences in the differences in the AUCs. Remarkably, despite comparable AUCs for FeNO and blood eosinophils in our study, the upper threshold range for blood eosinophils was relatively wide due to the higher threshold in patients with severe asthma compared to the other asthma phenotypes (Tables 4 and Supplementary Table E10). Apparently, a subset of patients with severe asthma shows elevated levels of blood eosinophils without evidence of airway eosinophilia, which confirms previous findings.[229] Circulating eosinophils might serve as a reservoir in these patients, thereby maintaining airway inflammation, which cannot be adequately suppressed by inhaled corticosteroids.

Smoking in asthma has often been associated with neutrophilic airway inflammation[232] and enhancement of Th2 mediated inflammation,[243] and has also been shown to be associated with reduced FeNO levels.[244] Therefore, (ex-)smoking could have had an effect on the diagnostic accuracy of FeNO to detect sputum eosinophilia.[181,228] A previous study found a lower AUC for FeNO among smokers compared to non-smokers (0.63 versus 0.77),[181] but this was not the case in our study. This suggests that even in smokers and ex-smokers FeNO can be used as a biomarker for sputum eosinophilia.

The major strength of our study is the large number and the extensive characterization of the patients, which enabled us to investigate clinical (sub)phenotypes of adult-onset asthma. Another strength is that we reported biomarker thresholds at either high sensitivity or high specificity. These cut-off points are more useful for practicing physicians to confirm or exclude airway eosinophilia with high certainty. However, a limitation of this approach is that this method only gives a clear outcome in up to half of the patients; the remainder still need to undergo sputum induction to confirm or exclude sputum eosinophilia. Another possible limitation of our study is the number of missing sputum samples, in particular in patients with mild-to-moderate asthma. This limits the extrapolation of our results to all patients with adult asthma. However, unsuccessful sputum induction in mild-to-moderate asthma might be indicative of a low level of sputum eosinophils, which fits in with the observed lower level of blood eosinophils in this group.

Our study has clinical implications. First, it shows that in a large subset of adult patients airway eosinophilia can be identified with high certainty by using FeNO and blood eosinophils instead of induced sputum. Second, it shows that the accuracy of these biomarkers is similar in various subtypes and severities of asthma. Currently, FeNO and blood eosinophils are mainly used in clinical trials to identify patients with eosinophilic asthma who are eligible for treatment with novel targeted therapies. For example for mepolizumab, a blood eosinophil cut-off of >0.15×109 cells·L−1 was introduced to detect eosinophilic asthma and predict reduction of asthma exacerbations.[245] Our data show that this is an adequate threshold to detect eosinophilia, since an eosinophil count <0.09×109 cells·L−1 is associated with absence of airway eosinophilia in 92% of patients. Still, consensus about the respective biomarker thresholds is needed, as well as an algorithm and external validation that incorporates a combination of biomarkers.

In conclusion, we showed that FeNO and blood eosinophils have a comparable diagnostic accuracy to identify airway eosinophilia in adult asthma patients irrespective of phenotypic characteristics such as asthma severity, atopy, obesity, and smoking status, and, possibly, irrespective of underlying pathways leading to airway eosinophilia. In future clinical trials and day-to-day practice both markers, preferably in combination, may become the preferred method to assess eosinophilic airway inflammation and to guide targeted treatment in adult asthma patients with different phenotypes.

Chapter 13

Added value of combined endobronchial and esophageal endosonography for mediastinal nodal staging in lung cancer: a systematic review and meta-analysis

Daniël A. Korevaar
Laurence M. Crombag
Jérémie F. Cohen
René Spijker
Patrick M. Bossuyt
Jouke T. Annema

Lancet Respiratory Medicine 2016
In press

Appendices and supplemental material will become available online at:
http://www.thelancet.com/journals/lanres/issues

Abstract

Background

Clinical guidelines recommend endosonography with fine-needle aspiration for mediastinal nodal staging in non-small cell lung cancer, but most do not specify whether this should be through endobronchial endoscopy (EBUS), esophageal endoscopy (EUS), or both. We evaluated the added value and diagnostic accuracy of the combined use of EBUS and EUS.

Methods

We performed a systematic review and searched Medline, Embase, Biosis Previews and Web of Science (January 1, 2000 to February 25, 2016) without language restrictions. We included studies that assessed the accuracy of the combined use of EBUS and EUS in detecting mediastinal nodal metastases (N2/N3 disease) in patients with lung cancer. We performed random-effects meta-analysis.

Results

We included 13 studies (2,395 patients). Median prevalence of N2/N3 disease was 34% (range 23 to 71%). On average, adding EUS to EBUS increased sensitivity by 0.12 (95%CI 0.08 to 0.18). Adding EBUS to EUS increased sensitivity by 0.22 (95%CI 0.16 to 0.29). Average sensitivity of the combined approach was 0.86 (95%CI 0.81 to 0.90), and negative predictive value (NPV) was 0.92 (95%CI 0.89 to 0.93). Average NPV was significantly higher in studies with a prevalence ≤34% compared to studies with a prevalence >34%: 0.93 (95%CI 0.91 to 0.95) versus 0.89 (95%CI 0.85 to 0.91) (p=0.013). There were no significant differences in average sensitivity and NPV between studies that first performed EBUS or first performed EUS, and between studies that used an EBUS-scope to perform EUS or used a regular echo-endoscope.

Conclusions

The combined use of EBUS and EUS substantially improves sensitivity in detecting mediastinal metastases, reducing the need for surgical staging procedures.

Introduction

In non-small cell lung cancer, the stage of disease directly determines the prognosis and treatment options.[246] If distant metastases are absent, surgery with curative intent is the treatment of choice when disease is confined to the lung and hilar lymph nodes (N0/N1 disease).[35] However, when mediastinal lymph nodes are involved (N2/N3 disease), chemoradiotherapy is usually indicated.[35] This makes accurate mediastinal staging crucial.

Computed tomography (CT) or positron emission tomography-computed tomography (PET-CT) are commonly used in the initial characterization of lung tumours and in the search for metastases, but these tests are generally insufficiently accurate for mediastinal staging.[246-248] In case of a small peripheral tumour without radiological evidence of mediastinal involvement, additional preoperative mediastinal staging is not required.[248] However, additional testing with tissue confirmation is recommended in patients with enlarged or PET-positive intrathoracic lymph nodes, as well as in those with a normal mediastinum but at increased risk of mediastinal involvement, for example, because of a primary tumour size of ≥3cm.[248-250]

Mediastinoscopy and thoracoscopy have been used for nodal tissue confirmation, but these surgical procedures are relatively costly and invasive. Endobronchial ultrasound with real-time guided transbronchial needle aspiration (EBUS) and transoesophageal endoscopic ultrasound-guided fine-needle aspiration (EUS) are cheaper and less-invasive alternatives for mediastinal staging.[251] With these endosonographic techniques, biopsies of mediastinal structures can be obtained under real-time ultrasound guidance through the bronchial and oesophageal wall, respectively.[252] Clinical guidelines recommend to perform EBUS and/or EUS before considering additional surgical staging procedures.[248-250]

Although the specificity of EBUS and EUS in detecting mediastinal nodal metastases is considered close to perfect, sensitivity is less optimal.[253,254] False negative results can lead to unnecessary surgical interventions and suboptimal treatment. As EBUS and EUS are complementary techniques, relying on different modes of access to the mediastinum, combining these two procedures is likely to further increase sensitivity.[250] EUS can now also be performed with an EBUS-scope (EUS-B), which further facilitates a combined approach.

The extent to which the combined use of EBUS and EUS(-B) increases sensitivity for mediastinal nodal metastases is controversial, and many clinics only use one of these endosonographic techniques. We performed a systematic review and meta-analysis to obtain summary estimates of the added value and diagnostic accuracy of the combined use of EBUS and EUS(-B) in detecting mediastinal nodal metastases (N2/N3 disease).

Methods

This systematic review was prospectively registered at PROSPERO (CRD42015019249).

Eligibility criteria

Studies were eligible if they evaluated the diagnostic accuracy of the combined use of EBUS and EUS(-B) in detecting mediastinal nodal metastases (N2/N3 disease) in patients with known or suspected potentially resectable lung cancer. Both EUS and EUS-B were eligible procedures in this combined approach.

We excluded studies focusing on diagnosing primary lung tumors, studies focusing on restaging the mediastinum after induction therapy, studies evaluating the staging accuracy of EBUS or EUS(-B) only, and studies that did not aim to perform both EBUS and EUS(-B) in every eligible patient. We also excluded studies that selected patients for inclusion based on the results of the first endoscopic test.

Literature search and study selection

A medical information specialist (R.S.) developed literature searches in Medline, Embase, Biosis Previews and Web of Science. The search strategies consisted of a combination of index terms and free text words related to mediastinal staging in lung cancer (e.g. 'mediastinal staging', 'non small cell lung cancer') and endosonography (e.g. 'endoscopic echography', 'fine-needle aspiration', 'EUS', 'EBUS', 'endobronchial ultrasonography'). The full search strategy and detailed database information is provided in Appendix 1, available online. There were no language restrictions. Studies published before January 1, 2000 were not considered, as EBUS for mediastinal staging has first been described in the early 2000s. The final searches were performed on February 25, 2016.

Two independent investigators (D.A.K., L.M.C.) examined titles and abstracts of all search results. If a study was considered potentially eligible by at least one of them, they both read the full article to decide on inclusion. They resolved disagreements by discussion.

For each included article, one investigator (D.A.K.) also scanned reference lists, the first 20 "related citations" in PubMed, and all articles citing that article according to Google Scholar. The same investigator also searched ClinicalTrials.gov and the WHO International Clinical Trials Registry Platform Search Portal without time limits for unpublished or ongoing studies. These additional searches were performed in March 2016.

Data extraction and quality assessment

Data extraction was performed by one investigator (D.A.K.), and verified by a second investigator (L.M.C.). Disagreements were resolved by consensus. For each included study, we extracted data on the age and gender of participants, inclusion criteria regarding tumor stage on imaging, details of the endoscopic testing protocol, average duration of each endoscopic procedure, average number of lymph nodes sampled, serious adverse events occurring during the endoscopic procedures, and the reference standard.

We built three 2x2 tables for each study, based on the number of true and false positives and true and false negatives for EBUS, for EUS(-B), and for the combined approach. Patients with mediastinal nodal metastases (N2/N3 disease) were considered positives; patients without mediastinal nodal metastases (N0/N1 disease) were considered negatives. Whenever 2x2 tables were not provided in the study report, we attempted to reconstruct them from summary accuracy estimates or by contacting authors.

Two investigators (D.A.K., L.M.C.) independently assessed risk of bias and applicability concerns of the included studies using QUADAS-2.[28] Disagreements were resolved by consensus. We considered the following reference standards to have a low risk of bias for the confirmation of a negative endoscopic test result: pulmonary resection with mediastinal nodal dissection or exploration, transcervical extended mediastinal lymphadenectomy, video-assisted mediastinoscopic lymphadenectomy, and video assisted thoracic surgery with mediastinal nodal dissection. Reference standards that were considered to have a high risk of bias were: mediastinoscopy, bronchoscopy, and clinical or radiological follow-up. If multiple reference standards were used, we considered risk of bias to be high if 10% or more of the patients classified as true negatives were confirmed by a reference standard with a high risk of bias.

Outcomes

In this review, we evaluated the added value (absolute increase in sensitivity and in detection rate) of the combined use of EBUS and EUS(-B) in detecting mediastinal nodal metastases over either test alone, and the diagnostic accuracy (sensitivity and negative predictive value (NPV)) of the combined approach.

The increase in sensitivity is the proportion of additional true positives with the combined approach, compared to EBUS or EUS(-B) alone, relative to the total number of reference standard positives. The increase in detection rate is the proportion of additional true positives with the combined approach, compared to EBUS or EUS(-B) alone, relative to the total number of evaluated patients. The inverse of the increase in detection rate can be considered as the number needed to test: the number of patients in whom the combined approach needs to be applied to detect one additional patient with mediastinal nodal metastases, relative to using only one test. In a post hoc analysis, we also compared the diagnostic accuracy of EBUS and EUS(-B) in isolation.

Statistical analysis

For each study, we used the extracted 2x2 tables to calculate estimates of increase in sensitivity and increase in detection rate (as defined above), and of the overall sensitivity and NPV of EBUS, EUS(-B), and the combined approach. We calculated 95% confidence intervals (CIs) around these proportions using the normal approximation.

We then logit transformed these proportions and performed univariate random effects meta-analysis according to DerSimonian-Laird,[116] to obtain summary estimates of the average increase in sensitivity and in detection rate, and of the average overall sensitivity and NPV of the combined approach. We used the same methods to obtain estimates of the average sensitivity and NPV of EBUS versus EUS(-B). In this analysis, we only included studies that provided a 2x2 table for both tests, which allowed us to make direct comparisons between the accuracy of EBUS and EUS(-B), as each patient underwent both tests.

From the random-effects meta-analyses, we report the means of the respective distributions, as reflections of the average performance statistics, with 95% confidence intervals around these means. We calculated I^2-statistics, to indicate the percentage of variability that is explained by between-study heterogeneity; an I^2-statistic >50% is considered to represent substantial heterogeneity.[179] We calculated 95% prediction intervals (PIs), which takes into account the full uncertainty about the location of the summary estimate, generated by the imprecision in the estimated average and the between-study heterogeneity.[255]

Possible methodological sources of variability in estimates were evaluated by comparing studies that used a reference standard with a low risk of bias with those that did not. Potential clinical sources of heterogeneity were evaluated by comparing subgroups based on prevalence of mediastinal nodal involvement (below versus above the median), order of testing (EBUS first versus EUS(-B) first), and type of oesophageal endoscopy (EUS versus EUS-B).

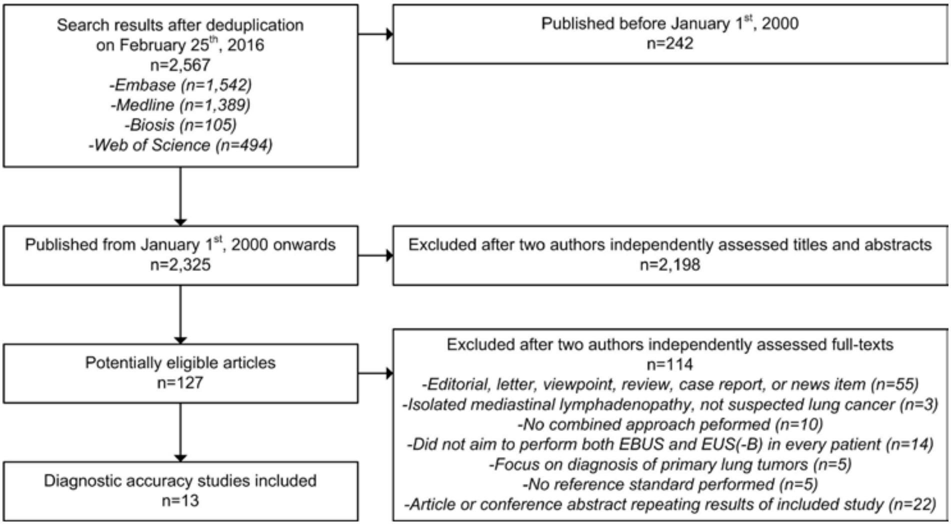
Data analysis was performed using the 'meta' package in R version 3.0 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Search and selection

From 2,567 search results, we included 13 diagnostic accuracy studies (Figure 1).[256-268] One of the included studies was reported in a conference abstract from our own institution (full manuscript in preparation).[267] Searching trial registries revealed three additional studies that could not be included: two because they had been terminated (ClinicalTrial.gov identifiers NCT00970645 and NCT01117714), and one because it was ongoing (University hospital Medical Information Network Clinical Trials Registry identifier UMIN000009752). Screening reference lists, related citations, and citing articles revealed no additional eligible studies.

Figure 1. Flowchart for selection of studies.


Search results after deduplication
on February 25th, 2016
n=2,567
-Embase (n=1,542)
-Medline (n=1,389)
-Biosis (n=105)
-Web of Science (n=494)
Published before January 1st, 2000
n=242
Published from January 1st, 2000 onwards
n=2,325
Excluded after two authors independently assessed titles and abstracts
n=2,198
Potentially eligible articles
n=127
Excluded after two authors independently assessed full-texts
n=114
-Editorial, letter, viewpoint, review, case report, or news item (n=55)
-Isolated mediastinal lymphadenopathy, not suspected lung cancer (n=3)
-No combined approach peformed (n=10)
-Did not aim to perform both EBUS and EUS(-B) in every patient (n=14)
-Focus on diagnosis of primary lung tumors (n=5)
-No reference standard performed (n=5)
-Article or conference abstract repeating results of included study (n=22)
Diagnostic accuracy studies included
n=13

Study characteristics and test procedures

Detailed characteristics of the 13 included studies are provided in Appendix 2 and 3. In summary, the mean or median age of patients ranged from 58 to 69 years (median 65), and the proportion of males from 48% to 80% (median 69%). Most studies (n=9).

Seven studies first performed EBUS before performing EUS(-B), whereas three studies first performed EUS(-B). One study assigned patients to first receive EBUS or EUS-B based on random allocation, and one study assigned patients to first receive EBUS or EUS based on a non-random allocation mechanism. Information about the order of testing was unclear in one study. In six studies, one endoscopist performed both procedures, whereas in two studies, EBUS and EUS(-B) were performed by separate endoscopists. The number of endoscopists was unclear in five studies.

Seven studies performed EUS and five studies performed EUS-B; in one study, patients were non-randomly assigned to receive EUS or EUS-B. The mean or median duration of the procedures ranged from 14 to 26 minutes (median 19) for EBUS, from 10 to 21 minutes (median 16) for EUS, and from four to 16 minutes (median 10) for EUS-B, but not all studies reported this information.

Study quality

Detailed results of the study quality assessment are provided in Appendix 4. About half of the studies (n=7; 54%) showed a high risk of bias, but usually only in one domain, almost always because they (partly) relied on imperfect reference standards (n=5; 38%).

Meta-analysis: added value

Across the 13 included studies, the number of patients ranged from 28 to 696, and prevalence of mediastinal nodal metastases (N2/N3 disease) ranged from 23% to 71% (median 34%). In total, 2,395 patients were included in this review who all underwent both EBUS and EUS(-B), of whom 887 (37%) had mediastinal nodal metastases. Estimates of diagnostic accuracy for EBUS, EUS(-B), and the combined approach for each study are reported in Table 1, with 2x2 tables in Appendix 5.

Adding EUS(-B) to EBUS led to an average increase in sensitivity of 0.12 (95%CI 0.08 to 0.18; I^2=61.8%; 95%PI 0.03 to 0.39) (Figure 2a), and to an average increase in detection rate of 0.04 (95%CI 0.03 to 0.06; I^2=36.0%, 95%PI 0.02 to 0.10), which implies a number needed to test of 25 (95%CI 17 to 33).

Adding EBUS to EUS(-B) led to an average increase in sensitivity of 0.22 (95%CI 0.16 to 0.29; I^2=40.3%; 95%PI 0.10 to 0.42) (Figure 2b), and an average increase in detection rate of 0.07 (95%CI 0.05 to 0.09; I^2=23.7%; 95%PI 0.04 to 0.13), corresponding to a number needed to test of 14 (95%CI 11 to 20). Estimates of added value in subgroups of studies are provided in Appendix 6.

Meta-analysis: diagnostic accuracy

Average sensitivity of the combined approach was 0.86 (95%CI 0.81 to 0.90; I^2=62.3%; 95%PI 0.66 to 0.95), with an average NPV of 0.92 (95%CI 0.89 to 0.93; I^2=39.2%; 95%PI 0.84 to 0.96) (Figure 3). Accuracy was lower in studies that relied on a reference standard with a low risk of bias compared to those with a high risk of bias, with a sensitivity of 0.83 (95%CI 0.77 to 0.87) versus 0.91 (95%CI 0.86 to 0.95) (p=0.015), and an NPV of 0.91 (95%CI 0.88 to 0.93) versus 0.94 (95%CI 0.90 to 0.97) (p=0.10), respectively.

Estimates of diagnostic accuracy in subgroups of studies are provided in Table 2. NPV was significantly higher in studies with a prevalence <34% compared to studies with a prevalence ≥34%: 0.93 (95%CI 0.91 to 0.95) versus 0.89 (95%CI 0.85 to 0.91) (p=0.013), respectively. This was also the case in the subgroup of studies that used a reference standard with a low risk of bias: 0.92 (95%CI 0.89 to 0.94) versus 0.87 (95%CI 0.81 to 0.90) (p=0.020), respectively. There were no significant differences in average sensitivity and NPV between studies that first performed EBUS compared to those that first performed EUS(-B), and between studies that used EUS compared to those that used EUS-B.

Among the seven studies (847 patients) for which 2x2 tables were available for both EBUS and EUS(-B), average sensitivity was 0.72 (95%CI 0.58 to 0.82) for EBUS, and 0.67 (95%CI 0.54 to 0.78) for EUS (Appendix 7). Average NPV was 0.88 (95%CI 0.85 to 0.90) for EBUS, and 0.86 (95%CI 0.84 to 0.89) for EUS(-B) (Appendix 7).

Serious adverse events and complications

Serious adverse events or complications were reported in seven of the 2,171 patients (0.32%) for whom this information was available. These were all due to EBUS: severe cough (n=2), left-sided mainstem bronchus laceration (n=1), massive hemoptysis (n=1), lymph node abscess (n=1), pneumothorax (n=1), and pneumomediastiun (n=1) (Appendix 2).

Table 1. Accuracy in detecting mediastinal nodal metastases across included studies.

Study	Order of testing		Patients		EBUS		EUS(-B)		Combined approach	
	Test 1	Test 2	n	% N2/N3	Sensitivity (95%CI)	NPV (95%CI)	Sensitivity (95%CI)	NPV (95%CI)	Sensitivity (95%CI)	NPV (95%CI)
Vilmann 2005[a]	EUS	EBUS	28	71%	0.85 (0.62-0.95)	0.73 (0.41-0.91)	0.80 (0.57-0.92)	0.67 (0.38-0.87)	1.00 (0.71-1.00)	1.00 (0.50-1.00)
Wallace 2008[a]	EBUS	EUS	138	30%	0.69 (0.54-0.81)	0.88 (0.81-0.93)	0.69 (0.54-0.81)	0.88 (0.81-0.93)	0.93 (0.80-0.98)	0.97 (0.91-0.99)
Annema 2010[b]	EUS	EBUS	123	54%	-	-	-	-	0.85 (0.74-0.92)	0.85 (0.74-0.92)
Herth 2010[a]	EBUS	EUS-B	139	51%	0.92 (0.82-0.96)	0.92 (0.83-0.96)	0.89 (0.79-0.94)	0.89 (0.80-0.95)	0.96 (0.88-0.99)	0.96 (0.88-0.99)
Hwangbo 2010[b]	EBUS	EUS-B	143	31%	0.84 (0.71-0.92)	0.93 (0.87-0.97)	-	-	0.91 (0.79-0.97)	0.96 (0.90-0.99)
Szlubowski 2010[b]	EUS	EBUS	120	23%	0.46 (0.29-0.65)	0.86 (0.78-0.91)	0.50 (0.32-0.68)	0.87 (0.79-0.92)	0.68 (0.49-0.82)	0.91 (0.83-0.95)
Ohnishi 2011[b]	EBUS	EUS	110	28%	0.74 (0.56-0.87)	0.91 (0.83-0.95)	0.65 (0.47-0.79)	0.88 (0.79-0.93)	0.84 (0.67-0.93)	0.94 (0.86-0.98)
Szlubowski 2012[b] *Group 1*	EUS	EBUS	110	54%	-	-	-	-	0.92 (0.81-0.96)	0.91 (0.80-0.96)
Szlubowski 2012[b] *Group 2*	EBUS	EUS-B	104	55%	-	-	-	-	0.84 (0.72-0.92)	0.83 (0.70-0.91)
Kang 2014[b] *Group 1*	EBUS	EUS-B	74	46%	0.82 (0.66-0.92)	0.87 (0.74-0.94)	-	-	0.85 (0.69-0.94)	0.89 (0.76-0.95)
Kang 2014[b] *Group 2*	EUS-B	EBUS	74	34%	-	-	0.60 (0.40-0.77)	0.83 (0.71-0.91)	0.92 (0.73-0.98)	0.96 (0.86-0.99)
Liberman 2014[a]	EBUS	EUS	166	32%	0.72 (0.58-0.82)	0.88 (0.81-0.93)	0.62 (0.49-0.74)	0.85 (0.78-0.90)	0.91 (0.79-0.96)	0.96 (0.90-0.98)
Oki 2014[b]	EBUS	EUS-B	146	23%	0.52 (0.35-0.68)	0.88 (0.81-0.92)	0.45 (0.30-0.62)	0.86 (0.79-0.91)	0.73 (0.55-0.85)	0.93 (0.86-0.96)
Crombag 2015[a]	EBUS	EUS-B	224	47%	0.82 (0.73-0.88)	0.86 (0.79-0.91)	-	-	0.87 (0.79-0.92)	0.89 (0.83-0.94)
Hauer 2015[b]	NR	NR	696	31%	-	-	-	-	0.75 (0.69-0.81)	0.90 (0.87-0.92)

[a]Reference standard with a high risk of bias. [b]Reference standard with a high risk of bias. NR=not reported.

Table 2. Accuracy of combined EBUS and EUS(-B) in detecting mediastinal nodal metastases.

		All studies				Studies with a reference standard with a low risk of bias only			
		Studies n[a]	Patients n	Sensitivity (95%CI)	NPV (95%CI)	Studies n[a]	Patients n	Sensitivity (95%CI)	NPV (95%CI)
Overall		15	2395	0.86 (0.81-0.90)	0.92 (0.89-0.93)	10	1700	0.83 (0.77-0.87)	0.91 (0.88-0.93)
Prevalence N2/3:	**≤34%**	8	1593	0.83 (0.76-0.89)	0.93 (0.91-0.95)[b]	6	1289	0.80 (0.71-0.86)	0.92 (0.89-0.94)[b]
	>34%	7	802	0.88 (0.84-0.91)	0.89 (0.85-0.91)	4	411	0.86 (0.81-0.90)	0.87 (0.81-0.90)
Order of testing:[c]	**EBUS first**	9	1244	0.87 (0.83-0.91)	0.93 (0.90-0.95)	5	577	0.83 (0.77-0.88)	0.92 (0.86-0.95)
	EUS(-B) first	5	455	0.87 (0.75-0.93)	0.90 (0.86-0.93)	4	427	0.85 (0.73-0.92)	0.90 (0.85-0.94)
Type of esophageal endoscopy:	**EUS**	8	1491	0.85 (0.78-0.90)	0.92 (0.89-0.94)	5	1159	0.81 (0.72-0.88)	0.90 (0.88-0.92)
	EUS-B	7	904	0.87 (0.81-0.91)	0.92 (0.88-0.95)	5	541	0.85 (0.77-0.90)	0.92 (0.86-0.95)

[a]Two studies each consisted of two groups with separate 2x2 tables; these were counted as separate studies. [b]NPV was significantly higher in studies with a prevalence ≤34% compared to studies with a prevalence >34% in 'all studies' (p=0.013), as well as in 'studies with a reference standard with a low risk of bias only' (p=0.020). No statistically significant differences were found in the other subgroup analyses. [c]Order of testing unclear for one study.

Discussion

Accurate mediastinal staging of lung cancer is crucial, as the stage of disease directly determines the prognosis and guides treatment options. In this systematic review we found that the combined use of EBUS and EUS(-B) leads to a significant increase in sensitivity and detection rate compared to either test alone.

Every patient included in this review underwent both EBUS and EUS(-B). This allowed us to perform direct, fully paired, within-patient comparisons of the accuracy of these single tests and of the combined approach. This strategy is considered methodologically superior to relying on indirect comparisons, where differences in accuracy may also be caused by differences in study group characteristics.[269] We observed substantial heterogeneity in our meta-analyses, which is not unusual for systematic reviews of diagnostic accuracy studies, and often caused by variability in patient and study characteristics.[270] Disease prevalence, for example, is a well-documented source of variation in diagnostic accuracy,[154] and varied considerably across included studies.

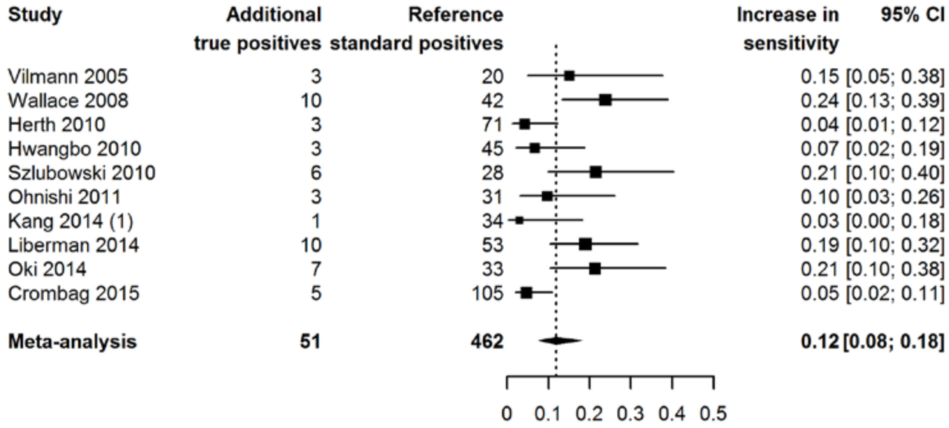
We found that the sensitivity of the combined approach was significantly higher in studies that used a reference standard with a high risk of bias. With inferior reference standards, fewer false negatives will be detected, and the sensitivity of the investigated test will be overestimated. The results from diagnostic accuracy studies that use imperfect reference standards should therefore be interpreted with caution.

Clinical guidelines recommend the use of endosonography for mediastinal staging in lung cancer, but most do not specify whether this should be EBUS, EUS(-B), or a combination.[248,249] Based on our findings, we recommend that the combined approach is considered, especially in settings where currently only EUS(-B) is performed. Adding EUS(-B) to EBUS increases sensitivity by 0.12, and adding EBUS to EUS(-B) increases sensitivity by 0.22. EUS(-B) needs to be added to EBUS in 25 patients, and EBUS to EUS(-B) in 14 patients, to detect one additional patient with mediastinal nodal metastases that would not have been detected if only one test had been performed. This could imply that unnecessary surgical interventions can be prevented by performing the combined approach.

Clinical guidelines also recommend that additional surgical staging should be considered to rule out mediastinal nodal involvement in patients with a negative endosonography.[248-250] Our findings support this recommendation, especially in settings with a high prevalence of mediastinal metastases, for example, in patients with a suspected mediastinum by imaging. Metastatic disease is not ruled out after a negative test result, even with the combined approach. Based on the results in the subgroup of studies that used a reference standard with a low risk of bias, average NPV was 0.92 in low prevalence settings, implying that 8% of patients with a negative test would still have mediastinal nodal metastases. This post-test probability after a negative test was 13% in high-prevalence settings.

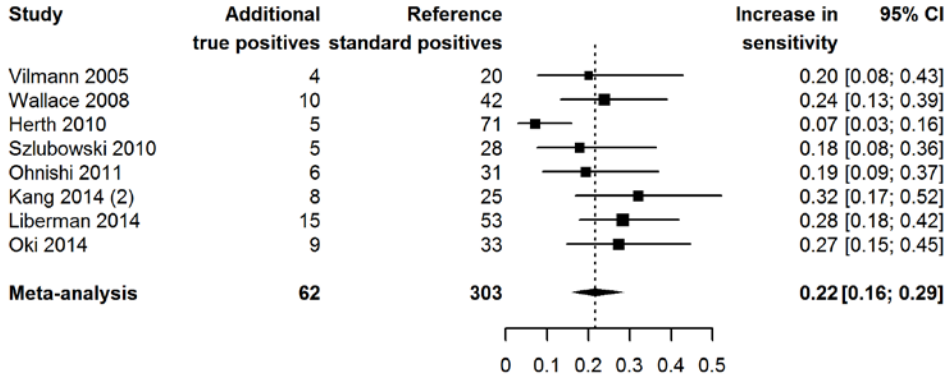
Figure 2. Added value of combined EBUS and EUS(-B) in detecting mediastinal nodal metastases.

Figure 2a. Increase in sensitivity of the combined approach compared to EBUS alone.



Study	Additional true positives	Reference standard positives	Increase in sensitivity	95% CI
Vilmann 2005	3	20	0.15	[0.05; 0.38]
Wallace 2008	10	42	0.24	[0.13; 0.39]
Herth 2010	3	71	0.04	[0.01; 0.12]
Hwangbo 2010	3	45	0.07	[0.02; 0.19]
Szlubowski 2010	6	28	0.21	[0.10; 0.40]
Ohnishi 2011	3	31	0.10	[0.03; 0.26]
Kang 2014 (1)	1	34	0.03	[0.00; 0.18]
Liberman 2014	10	53	0.19	[0.10; 0.32]
Oki 2014	7	33	0.21	[0.10; 0.38]
Crombag 2015	5	105	0.05	[0.02; 0.11]
Meta-analysis	**51**	**462**	**0.12**	**[0.08; 0.18]**

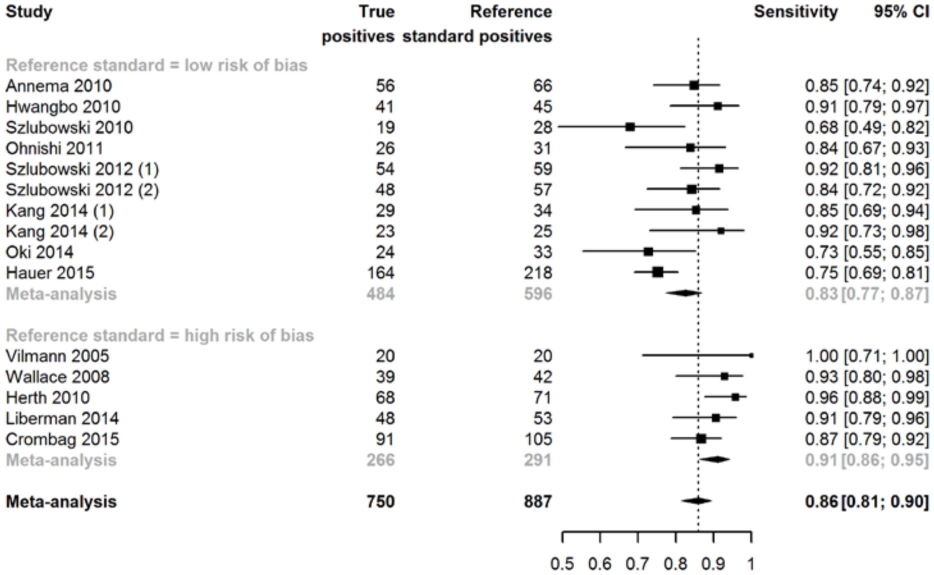
Figure 2b. Increase in sensitivity of the combined approach compared to EUS(-B) alone.



Study	Additional true positives	Reference standard positives	Increase in sensitivity	95% CI
Vilmann 2005	4	20	0.20	[0.08; 0.43]
Wallace 2008	10	42	0.24	[0.13; 0.39]
Herth 2010	5	71	0.07	[0.03; 0.16]
Szlubowski 2010	5	28	0.18	[0.08; 0.36]
Ohnishi 2011	6	31	0.19	[0.09; 0.37]
Kang 2014 (2)	8	25	0.32	[0.17; 0.52]
Liberman 2014	15	53	0.28	[0.18; 0.42]
Oki 2014	9	33	0.27	[0.15; 0.45]
Meta-analysis	**62**	**303**	**0.22**	**[0.16; 0.29]**

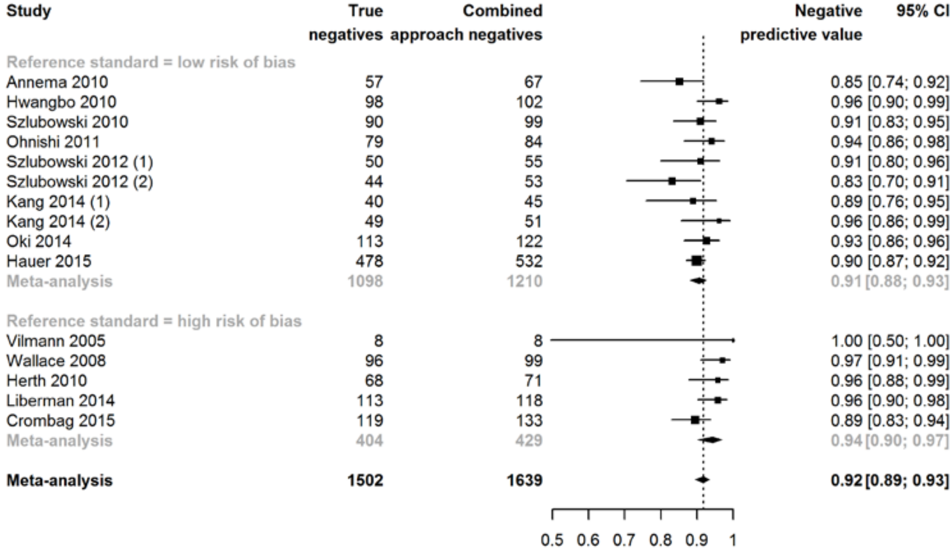
Figure 3. Accuracy of combined EBUS and EUS(-B) in detecting mediastinal nodal metastases.

Figure 3a. Sensitivity.



Study	True positives	Reference standard positives	Sensitivity	95% CI
Reference standard = low risk of bias				
Annema 2010	56	66	0.85	[0.74; 0.92]
Hwangbo 2010	41	45	0.91	[0.79; 0.97]
Szlubowski 2010	19	28	0.68	[0.49; 0.82]
Ohnishi 2011	26	31	0.84	[0.67; 0.93]
Szlubowski 2012 (1)	54	59	0.92	[0.81; 0.96]
Szlubowski 2012 (2)	48	57	0.84	[0.72; 0.92]
Kang 2014 (1)	29	34	0.85	[0.69; 0.94]
Kang 2014 (2)	23	25	0.92	[0.73; 0.98]
Oki 2014	24	33	0.73	[0.55; 0.85]
Hauer 2015	164	218	0.75	[0.69; 0.81]
Meta-analysis	484	596	0.83	[0.77; 0.87]
Reference standard = high risk of bias				
Vilmann 2005	20	20	1.00	[0.71; 1.00]
Wallace 2008	39	42	0.93	[0.80; 0.98]
Herth 2010	68	71	0.96	[0.88; 0.99]
Liberman 2014	48	53	0.91	[0.79; 0.96]
Crombag 2015	91	105	0.87	[0.79; 0.92]
Meta-analysis	266	291	0.91	[0.86; 0.95]
Meta-analysis	**750**	**887**	**0.86**	**[0.81; 0.90]**

Figure 3b. Negative predictive value.



Study	True negatives	Combined approach negatives	Negative predictive value	95% CI
Reference standard = low risk of bias				
Annema 2010	57	67	0.85	[0.74; 0.92]
Hwangbo 2010	98	102	0.96	[0.90; 0.99]
Szlubowski 2010	90	99	0.91	[0.83; 0.95]
Ohnishi 2011	79	84	0.94	[0.86; 0.98]
Szlubowski 2012 (1)	50	55	0.91	[0.80; 0.96]
Szlubowski 2012 (2)	44	53	0.83	[0.70; 0.91]
Kang 2014 (1)	40	45	0.89	[0.76; 0.95]
Kang 2014 (2)	49	51	0.96	[0.86; 0.99]
Oki 2014	113	122	0.93	[0.86; 0.96]
Hauer 2015	478	532	0.90	[0.87; 0.92]
Meta-analysis	1098	1210	0.91	[0.88; 0.93]
Reference standard = high risk of bias				
Vilmann 2005	8	8	1.00	[0.50; 1.00]
Wallace 2008	96	99	0.97	[0.91; 0.99]
Herth 2010	68	71	0.96	[0.88; 0.99]
Liberman 2014	113	118	0.96	[0.90; 0.98]
Crombag 2015	119	133	0.89	[0.83; 0.94]
Meta-analysis	404	429	0.94	[0.90; 0.97]
Meta-analysis	**1502**	**1639**	**0.92**	**[0.89; 0.93]**

Additional surgical staging in endosonography negative patients will further improve the detection of mediastinal nodal metastases. In the ASTER trial, for example, NPV of the combined approach was 0.85, which increased to 0.93 after performing mediastinoscopy in all endosonography negative patients.[258] In the ASTER-2 trial, similar numbers were found; NPV was 0.81 after endosonography and increased to 0.91 when adding mediastinoscopy.[271] However, in the latter study, the majority of patients underwent EBUS only, whereas only 25% had both EBUS and EUS(-B) because they had nodes that were inaccessible by EBUS alone, which may be an explanation for the low NPV.

EUS and EUS(-B) are relatively safe procedures. Serious adverse events of the combined approach occurred in 0.32%. This is in line with a previous systematic review that assessed rates of serious adverse events related to endosonography in the analysis of mediastinal and hilar lymph nodes and central intrapulmonary lesions, and reported a rate of 0.30% for EUS, and of 0.05% for EBUS.[272]

Lack of availability and expertise currently hamper the widespread implementation of the combined approach for mediastinal staging, and performing EBUS and EUS with separate scopes can be considered as time-consuming and inconvenient for patients. However, although not yet widely adopted, EBUS and EUS-B can now be performed in conjunction in a single session, both with an EBUS-scope, by a single endoscopist.[250] This strategy highly facilitates the combined approach, as it is quicker and more comfortable to patients. We also observed that the sensitivity and NPV of a combined approach that includes EUS-B are similar compared to one that includes EUS, indicating that endoscopists should be trained in performing both EBUS and EUS-B in a single procedure.[273]

Future research should focus on the optimal strategy for performing endosonography. A remaining question is whether both a complete EBUS and EUS-B should always be performed, or whether it is sufficient to only sample nodes with the second test if they were out of reach of the first test, or only if the first test was negative, assuming that rapid onsite cytology is available. If one of the two the latter strategies is preferred, a follow-up query would be whether the strategy should start with EBUS or EUS(-B). In our review, no significant differences in sensitivity and NPV were observed between studies that first performed EBUS and those that first performed EUS(-B). However, since adding EUS(-B) to EBUS led to a smaller increase in sensitivity than adding EBUS to EUS(-B), we recommend an EBUS-centered strategy as the preferred approach. A randomized trial, the results of which were included in this review, came to the same conclusion after comparing such an EBUS-centered strategy with an EUS-centered one.[264] In line with our findings, no significant difference in sensitivity between the two strategies was observed, but adding EBUS to EUS-B increased sensitivity by 0.32, whereas adding EUS-B to EBUS increased sensitivity by only 0.03.

This review shows that the combined use of EBUS and EUS(-B) leads to a significant gain in the detection of patients with mediastinal nodal metastases. Eventually, the maximal number of patients one is willing to submit to a combined approach to detect one additional case of mediastinal nodal metastases will guide clinical recommendations about the use of EBUS and EUS(-B).[274] At present no consensus on this criterion exists. Defining one would come down to weighing the downsides of combined testing, instead of relying on EBUS or EUS(-B) only, against the benefits of decreasing the number of false negative endoscopy results, thereby reducing futile surgical interventions in lung cancer patients.

Acknowledgments

We would like to thank Noelia Cubero de Frutos, Felix J. Herth, Bin Hwangbo, and Ichiro Yasuda, who kindly shared requested data.

Chapter 14

Five-year survival after endosonography versus mediastinoscopy for mediastinal nodal staging of lung cancer

Jolanda C. Kuijvenhoven
Daniël A. Korevaar
Kurt G. Tournoy
Thomas L. Malfait
Christophe Dooms
Robert C. Rintoul
Jouke T. Annema

JAMA 2016;316(10):1110-2

Introduction

Lung cancer accounts for the highest cancer-related mortality rate worldwide.[33] Accurate mediastinal nodal staging is crucial in the management of non-small cell lung cancer (NSCLC) as it directs therapy and has prognostic value.[248,250]

ASTER (Assessment of Surgical Staging vs Endosonographic Ultrasound in Lung Cancer: a Randomized Clinical Trial) compared a surgical staging strategy (mediastinoscopy) with an endosonographic staging strategy (combined use of endobronchial and transesophageal ultrasound, followed by mediastinoscopy if negative).[258] The endosonographic strategy was significantly more sensitive for diagnosing mediastinal nodal metastases than surgical staging (94% versus 79%).

If mediastinal staging is improved, more patients should receive optimal treatment and might survive longer. The current post hoc analysis evaluated survival in ASTER.

Methods

ASTER was registered at ClinicalTrials.gov (identifier NCT00432640). Of 241 patients with potentially resectable NSCLC, 123 were randomized to the endosonographic staging strategy and 118 to the surgical staging strategy in four tertiary referral centers in Leiden (the Netherlands), Ghent and Leuven (Belgium) and Cambridge (United Kingdom), between February 2007 and April 2009.[258] Surgical-pathological staging was the reference standard for mediastinal nodal assessment. At inclusion in ASTER, all participants provided written informed consent; the current analysis was either approved or waived by the involved ethical committees.

Between June 30 and October 15, 2015, survival data were obtained through patient records, death registers, or contact with general practitioners.

The proportion of survivors at five years for both staging strategies and odds ratios with 95%CI were calculated. Kaplan-Meier analysis was performed and hazard ratios were calculated to compare survival between the strategies, adjusting for mediastinal nodal metastases in a Cox model. Survival for patients with no date of death was censored on the date they were last known to be alive. The assumption of proportional hazard was tested and met. Subgroup analysis was performed for patients with nodal stages N2/N3 and N0/N1. Data were analyzed using SPSS version 22 (IBM, Armonk, NY, USA).

Results

Survival data were obtained for 237 of 241 patients (98%); two patients were lost to follow up in both groups. There were 182 males (77%) with a mean age at randomization of 65 years (SD 9). Detailed patient characteristics were previously reported.[258]

Survival at five years was 35% (42/121) for the endosonographic strategy versus 35% (41/116) for the surgical strategy (odds ratio 0.97 (95%CI 0.57 to 1.66)) (Table 1). The estimated median survival was 31 months (95%CI 21 to 41) versus 33 months (95%CI 23 to 43), respectively (adjusted hazard ratio 0.98 (95%CI 0.73 to 1.32) (Figure 1).

In the subgroup of patients with N2/N3 metastases, survival was 17% (11/64) in the endosonographic group versus 19% (10/52) in the surgical group (odds ratio 0.87 (95%CI 0.34 to 2.25)). In the subgroup of patients with N0/N1 metastases, survival was 54% (31/57) versus 48% (31/64), respectively (odds ratio 1.27 (95%CI 0.62 to 2.60)).

Table 1. Survival of the endosonographic versus the surgical staging strategy.

	Survival at five years n/N (%)	**Odds ratio for survival at five years (95% CI)**	
Overall			
Endosonographic staging strategy	42/121 (35)	0.97 (0.57-1.66)	
Surgical staging strategy	41/116 (35)		
N2/N3			
Endosonographic staging strategy	11/64 (17)	0.87 (0.34-2.25)	
Surgical staging strategy	10/52 (19)		
N0/N1			
Endosonographic staging strategy	31/57 (54)	1.27 (0.62-2.60)	
Surgical staging strategy	31/64 (48)		
	Estimated median survival in months (95%CI)	**Unadjusted hazard ratio for mortality (95%CI)**	**Adjusted hazard ratio for mortality[1] (95%CI)**
Overall			
Endosonographic staging strategy	31 (21-41)	1.04 (0.77-1.40)	0.98 (0.73-1.32)
Surgical staging strategy	33 (23-43)		
N2/N3			
Endosonographic staging strategy	21 (15-27)	1.04 (0.70-1.55)	-
Surgical staging strategy	22 (15-27)		
N0/N1			
Endosonographic staging strategy	72 (38-106)	0.91 (0.57-1.44)	-
Surgical staging strategy	57 (30-84)		

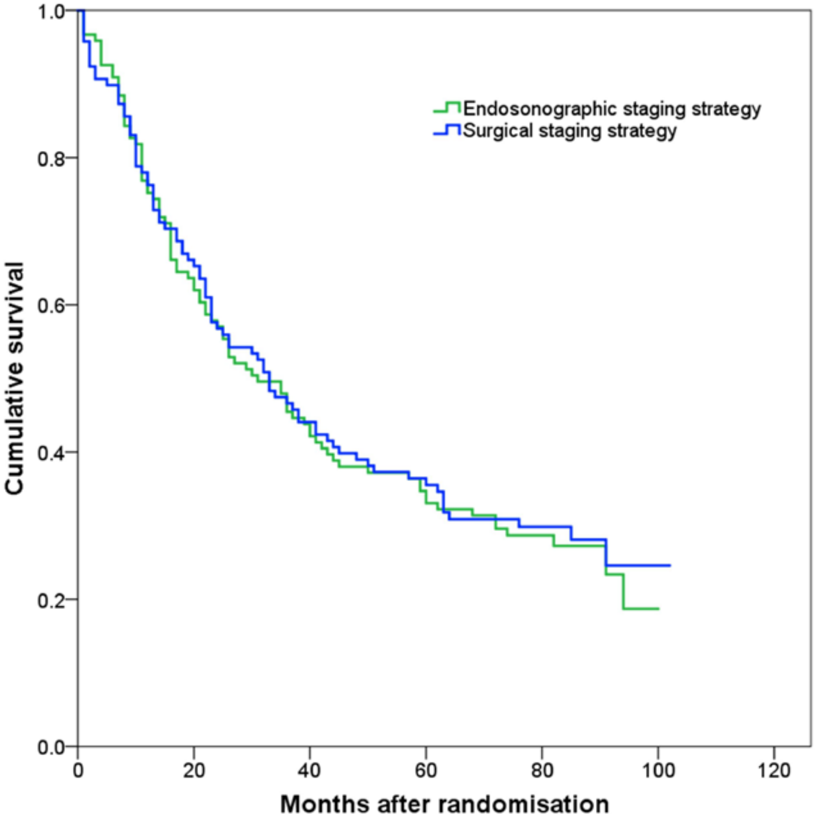
[1] Adjusted for mediastinal nodal metastases status (N0/1 versus N2/3).

Discussion

No survival difference was found five years following randomization to an endosonographic or surgical staging strategy of patients with NSCLC. Since the original results of ASTER were published, clinical guidelines on lung cancer management underwent major revisions, and now advocate the endosonographic strategy as the initial step for mediastinal nodal staging, instead of the surgical strategy.[248,250] The endosonographic strategy is more accurate, less invasive, reduces unnecessary thoracotomies[258] and has been shown to be cost-effective.[275]

Data from a recent randomized trial shows prolonged survival in patients who underwent endosonographic staging compared with conventional staging.[251] However, most patients in the latter group underwent bronchoscopy instead of mediastinoscopy. To our knowledge, ASTER is the first randomized trial to evaluate survival outcomes between endosonographic and surgical staging strategies.

Figure 1. Survival for the endosonographic versus the surgical staging strategy.



Why did improved mediastinal staging not lead to improved survival? Missing data occurred in less than 2%, and therefore are an unlikely source of bias. However, ASTER was powered to detect a difference in diagnostic sensitivity, not survival, as reflected by the wide confidence intervals. If a survival difference between the strategies exists, it is likely to be small and a larger sample size may be needed to detect it. However, randomized trials to detect a survival difference upon staging strategy are not likely to be conducted as the endosonographic strategy is now advised in clinical guidelines.[248,250]

Addendum

Summary

Introduction

Annually, more than a quarter of a trillion US Dollar is spent on biomedical research worldwide.[276] There have been increasing concerns that these funds are not allocated as efficiently as they could be. In 2009, Chalmers and Glasziou roughly estimated that a tremendous proportion - 85% - of research funding is avoidably wasted.[277] *Lancet* extended on this in 2014 by publishing a series of papers in which five major sources of research waste were identified, along with recommendations for their prevention.[278] Three of these sources relate to the planning and conduct of research: (1) irrelevant study objectives,[279] (2) inadequate study design or methods,[280] and (3) inefficient research regulation and management.[281] Two other sources of research waste have to do with the dissemination of research results: (4) inaccessible research[8] and (5) biased or unusable study reports.[9]

Evidence of a poor dissemination of findings from trials of therapeutic interventions started to emerge rapidly in the early 1990s, and the accumulated data are now extensive.[36,282] Such suboptimal dissemination is not only regrettable for financial reasons. Clinicians, nowadays trained to perform evidence-based medicine, rely on the published literature for making healthcare decisions and may not be able to act in patients' best interests if evidence is inaccessible or poorly presented.[10]

Because of these concerns, ethical principles for biomedical research, which historically mainly addressed the planning and conduct of research, have increasingly included the reporting and dissemination of study findings in their focus. The Declaration of Helsinki, for example, now states that "researchers, authors, sponsors, editors, and publishers all have ethical obligations with regard to the publication and dissemination of the results of research".[7]

The studies presented in this thesis aimed to uncover the extent of similar problems and deficiencies in the process of publishing and reporting diagnostic accuracy studies, with the ultimate goal of increasing value in diagnostic research.

Part A: Publication of full study reports

Many biomedical studies remain unpublished. Reporting bias lurks if promising results are more often and more rapidly published than less promising results.

With the studies presented in **Part A**, we wanted to map the extent to which failure to publish occurs among diagnostic accuracy studies, and what potential drivers of non- or delayed publication could be in this field of research.

In **Chapter 1**, we showed that non-publication is highly prevalent among diagnostic accuracy studies. In a sample of 418 diagnostic accuracy studies that were registered in ClinicalTrials.gov, only 54% were subsequently published in a peer-reviewed journal, at least 18 months after their completion. This percentage only increased to 59% when focusing on a subgroup of 302 studies that had been completed at least 30 months prior to our searches for corresponding publications. No less than 24% of published studies showed major discrepancies between the registered and published outcomes, such as registered primary outcomes that were omitted from the publication, or had become secondary outcomes. Unpublished studies may lead to the unnecessary duplication of research efforts, and can generally not be used to inform clinical practice.

In **Chapter 2**, we confirmed the results reported in the previous chapter in a sample of 399 diagnostic accuracy studies presented at the Annual Meeting of the Association for Research in Vision and Ophthalmology (ARVO). Of these, only 57% reached full-text publication in a peer-reviewed journal within five years after presentation. When evaluating the relationship between full-text publication and reported estimates of sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and diagnostic odds ratio, we found no statistically significant associations. These findings imply that conference abstracts may be a valuable source of unpublished diagnostic accuracy studies. Yet, we found no evidence that not including these in a systematic review will lead to reporting bias.

In **Chapter 3**, we demonstrated that reporting bias may certainly occur in the field of diagnostic testing. In a sample of 756 published diagnostic accuracy studies, we found that the time from completion of participant recruitment to publication was significantly associated with the reported estimates of sensitivity, specificity, and Youden's index: the beter the performance of the test, the faster the corresponding study was published. This resulted in a relative delay in publication of two months for studies reporting a sensitivity below the median, a relative delay of five months for studies reporting a specificity below the median, and a relative delay of five months for studies reporting a Youden's index below the median, compared to studies reporting estimates of these accuracy measures above the median. These delays occurred in the phase between study completion and submission of the study report to the publishing journal, and not in the phase between submission and publication. Delays in the publication of studies that show less promising results could lead to reporting bias in systematic reviews.

Part B: Prospective registration of study protocols

Registration of studies has been suggested as an important preventive measure against the negative consequences of failure to publish. For that reason, the International Committee of Medical Journal Editors (ICMJE) now requires registration of clinical trials before initiation of the study as a prerequisite for publication. The aim of the studies reported in **Part B** was to evaluate to which extent diagnostic accuracy studies are currently being registered, and to assess adherence of journals to ICMJE's trial registration policy.

In **Chapter 4**, we showed that diagnostic accuracy studies are rarely registered. In a sample of 351 published diagnostic accuracy studies, only 15% could be linked to a corresponding registered record. Of these, only 27% were registered before initiation of the study. This illustrates that, at this point, it is difficult to benefit from the advantages of prospective registration in diagnostic research.

In **Chapter 5**, we found that many journals poorly adhere to ICMJE's trial registration policy. In an analysis of the instructions to authors of 747 journals we found that only 51% included a statement of requiring trial registration. In a survey among journal editors only 50% of 232 responders indicated that trial registration was required at their journal. Only 18% cross-checked submitted papers against registered records to identify potential discrepancies, and 67% also considered retrospectively registered studies for publication. These results may provide an explanation to the findings of previous evaluations, which showed that registration rates among published trials of therapeutic interventions, although definitely increasing over time, remain too low.

Part C: Informative reporting of study reports

Readers of published reports of diagnostic accuracy studies often have difficulties in adequately interpreting the study findings, because key elements are incompletely or vaguely reported, or not reported at all. The Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement, introduced in 2003, offers authors and editors assistance in reporting this type of research. The aim of the studies reported in **Part C** was to assess the extent to which the completeness of reporting of diagnostic accuracy studies has improved over time, and to evaluate potential deficiencies in the current state of reporting of these studies.

In **Chapter 6**, we showed that reporting quality improved in the first few years after STARD's launch, but only to a modest extent. We performed a systematic review of evaluations of adherence to STARD. Of the 16 included evaluations, together analyzing the completeness of 1,496 study reports, all but one concluded that adherence to STARD was poor, medium, or suboptimal, or needed improvement. Across the evaluations, the overall mean number of items reported varied from 9.1 to 14.3 out of 25 STARD items. In a meta-analysis, we found that after the launch of STARD, on average, 1.41 more items were reported than before. However, we could not comment on the current state of reporting, nor whether this initial improvement in completeness of reporting had persisted over time, because the great majority of evaluated study reports had been published before 2007, which is roughly seven years before we performed our review.

In **Chapter 7**, we confirmed that the initial improvement in completeness of reporting after the launch of STARD identified in the previous chapter continued over time, but also concluded that substantial room for improvement remains. We evaluated adherence to STARD among 112 reports of diagnostic accuracy studies published in several high-impact factor journals. We also compared our findings with those of two previously published evaluations of study reports published in the same journals. Compared to studies published in 2000, on average 3.4 more items were reported in 2012, and compared to studies published in 2004, on average 1.7 more items were reported in 2012. Despite these improvements, the mean number of STARD items reported remained low in 2012, only 15.3 of 25 items. This illustrates that important study information is still often not provided.

In **Chapter 8** and **Chapter 9**, we found that the informativeness of journal and conference abstracts of diagnostic accuracy studies could improve as well. Using 21 items deemed potentially relevant to report in an abstract, we evaluated the information reported in 103 journal abstracts and in 126 conference abstracts presented at ARVO. Because abstracts are commonly limited to a couple of hundred words, and because guidelines for reporting abstracts of diagnostic accuracy studies were not available at the time of the analyses, it was interesting - though not surprising - to find that the information reported across abstracts was highly variable, and that some elements that seem critical to provide when summarizing a study were frequently absent.

In **Chapter 10**, we reported the methods used in the update of STARD, resulting in STARD 2015. The aim of the update was to (1) include new items, based on improved understanding of sources of bias and variability, and (2) to facilitate the use of the list, by rearranging and rephrasing existing items, and by improving consistency in wording with other major reporting guidelines. The updated STARD 2015 list now consists of 30 items. Compared to the previous version of STARD, three original items were each converted into two new items, four original items were incorporated into other items, and seven new items were added.

Part D: Diagnostic tests in respiratory medicine

In the clinical work-up of patients with asthma and lung cancer, clinicians rely on medical tests for diagnosis, but also for selecting patients for specific treatments, and for making statements about prognosis. The aim of the studies reported in **Part D** was to evaluate the diagnostic accuracy of a number of these tests.

In **Chapter 11**, we demonstrated that fraction of exhaled nitric oxide (FeNO), blood eosinophils, and total Immunoglobulin E (IgE) are insufficiently accurate to be used as single surrogate markers for airway eosinophilia in patients with asthma. We performed a systematic review of the diagnostic accuracy of minimally invasive markers in the detection of airway eosinophilia. These three markers had been evaluated most frequently across included studies, but their summary AUC after meta-analysis never exceeded 0.81. We recommended combining markers to arrive at a multivariable clinical prediction model with improved accuracy, and to establish thresholds of these markers for ruling-in and ruling-out airway eosinophilia. We were able to include a substantial amount of unpublished study results in this review, but found no evidence of reporting bias: summary accuracy estimates of published and unpublished data did not systematically differ.

In **Chapter 12**, we followed our own recommendations from the previous chapter. In a sample of 336 adult patients with asthma, we evaluated the diagnostic accuracy of FeNO, blood eosinophils, and total IgE in detecting sputum eosinophilia. Overall, AUC's were close to the summary estimates identified in the systematic review in the previous chapter. Accuracy could be improved, to an AUC of 0.89, when combining markers with several simple clinical parameters in a multivariable clinical prediction model. We also reported thresholds of these markers at a high sensitivity and at a high specificity, which, if applied, could rule-out or rule-in sputum eosinophilia with a high level of certainty in up to half of the patients.

In **Chapter 13**, we assessed the added value of using a combined approach of endobronchial endoscopy (EBUS) and esophageal endoscopy (EUS) for mediastinal nodal staging in patients with lung cancer. In a systematic review, we identified 13 relevant studies (2,395 patients), and found that adding EUS to EBUS increased sensitivity for the detection of mediastinal nodal metastases on average by 0.12, and adding EBUS to EUS increased sensitivity on average by 0.22. This led to an overall average sensitivity of the combined approach of 0.86, at a negative predictive value of 0.92, respectively. Most current clinical guidelines on lung cancer staging recommend performing endosonography for this purpose, but are often not specific about whether this should be through EBUS or EUS. The results of our review suggest that a combination of both tests should be considered.

In **Chapter 14**, we evaluated five-year survival in ASTER (Assessment of Surgical Staging versus Endosonographic Ultrasound in Lung Cancer: a Randomized Clinical Trial). This is a study in which patients with lung cancer had been randomized to receive a surgical mediastinal staging strategy with mediastinoscopy, or an endoscopic mediastinal staging strategy, which consisted of the combined approach of EBUS and EUS, followed by mediastinoscopy if endoscopy was negative. In ASTER, the endoscopic strategy was considerably more sensitive in detecting mediastinal nodal metastases: sensitivity was 94%, versus 79% for the surgical strategy. Because improved staging should lead to an improved treatment allocation, we assumed that this would also have a beneficial effect on survival. However, survival at five years was 35% for both strategies. These findings illustrate that improved test accuracy cannot always be translated to considerable improvements in patient-important outcomes.

Concluding remarks

The Declaration of Helsinki insists that researchers should register their studies involving human subjects, that they should make study results publically available, and that they should produce complete and accurate reports.[7] There is no reason to think that diagnostic accuracy studies are exempt from this ethical obligation.

Considerable improvements in the reporting and dissemination of the results of biomedical research have been observed over the past years. The proportion of registered trials has been growing strongly,[68] policies have been implemented that force researchers to publish results,[82] and journals have increasingly adopted reporting guidelines.[282]

Although diagnostic accuracy studies seem to be lagging behind a bit in most of these processes, some of the findings presented in this thesis show that there is room for modest optimism. Several authors already started to prospectively register their diagnostic accuracy studies, despite the fact that most journals do not yet require this. Reporting quality is slowly but visibly improving. And in our systematic reviews around tests in respiratory medicine, we were pleasantly surprised by the large number of researchers that were willing to share their unpublished data so that we could analyze and include them in our synthesis of all the available evidence.

It is crucial that additional steps will be made to further improve all of this in the coming years. All those professionally involved in biomedical research share a joint and therefore also a personal responsibility in making efforts to increase research value.[25]

Nederlandse samenvatting

Introductie

Jaarlijks wordt er wereldwijd meer dan 250 miljard US Dollar uitgegeven aan biomedisch onderzoek.[276] Er bestaan in toenemende mate zorgen dat deze financiële middelen niet zo efficiënt ingezet worden als zou kunnen. In 2009 schatten Chalmers en Glasziou ruwweg dat een enorm gedeelte - 85% - van de onderzoeksfinanciering vermijdbaar verspild wordt.[277] *Lancet* bouwde hier in 2014 op voort door een serie artikelen te publiceren waarin vijf belangrijke bronnen van verspilling geïdentificeerd werden, evenals aanbevelingen om ze te voorkómen.[278] Drie van deze bronnen verwijzen naar de planning en uitvoering van onderzoek: (1) irrelevante studiedoelen,[279] (2) inadequate onderzoeksopzetten of -methodes,[280] en (3) inefficiënte regelgeving en management van onderzoek.[281] Twee andere bronnen van verspilling van onderzoek hebben te maken met de verspreiding van onderzoeksresultaten: (4) ontoegankelijk onderzoek,[8] en (5) vertekende of onbruikbare studieverslagen.[9]

Bij trials van therapeutische interventies begon bewijs van een matige verspreiding van onderzoeksbevindingen in snel tempo op te duiken aan het begin van de jaren '90 van vorige eeuw, en de verzamelde bevindingen hieromtrent zijn nu imposant.[36,282] Clinici, tegenwoordig getraind om 'evidence-based medicine' toe te passen, vertrouwen op de literatuur bij het maken van klinische beslissingen; ze zijn wellicht niet in staat om in het beste belang van de patiënt te handelen als de onderbouwing van hun beslissingen niet toegankelijk is of slecht gepresenteerd wordt.[10]

Vanwege dit soort zorgen omvatten ethische richtlijnen voor biomedisch onderzoek, die zich van oudsher voornamelijk richtten op de planning en uitvoering van onderzoek, nu in toenemende mate ook de rapportage en verspreiding van onderzoeksresultaten. De Verklaring van Helsinki, bijvoorbeeld, vermeldt nu dat "onderzoekers, auteurs, sponsors, redacteuren en uitgevers allen ethische verplichtingen hebben met betrekking tot de publicatie en verspreiding van de resultaten van onderzoek".[7]

De onderzoeksprojecten die in dit proefschrift staan samengevat waren opgezet om de mate van vergelijkbare problemen en tekortkomingen in de publicatie en rapportage van onderzoek naar de diagnostische accuratesse van medische tests bloot te leggen, met als uiteindelijke doel om de waarde van diagnostisch onderzoek te vergroten.

Deel A: Publicatie van studierapporten

Veel biomedische studies blijven ongepubliceerd. 'Reporting bias' ligt op de loer als veelbelovende resultaten vaker en sneller gepubliceerd worden dan minder gunstige resultaten. Met het onderzoek dat wordt gepresenteerd in **Deel A** probeerden we in kaart te brengen in welke mate diagnostische accuratessestudies ongepubliceerd blijven, en wat de mogelijke oorzaken zijn van het niet of vertraagd publiceren van dergelijke studies.

In **Hoofdstuk 1** toonden we aan dat veel studies naar de diagnostische accuratesse van medische tests nooit worden gepubliceerd. In een steekproef van 418 diagnostische accuratessestudies die geregistreerd waren in ClinicalTrials.gov bleek dat slechts 54% gepubliceerd was in een 'peer reviewed' tijdschrift tenminste 18 maanden na hun afronding. Dit percentage steeg slechts tot 59% in de subgroep van 302 studies die 30 maanden of meer voorafgaand aan onze zoektocht naar bijpassende publicaties waren afgerond. Niet minder dan 24% van de gepubliceerde studies vertoonde duidelijke discrepanties tussen de vooraf geregistreerde en uiteindelijk gepubliceerde uitkomsten; het ging dan bijvoorbeeld om geregistreerde primaire uitkomsten die waren weggelaten uit de publicatie of die secundaire uitkomsten waren geworden. Ongepubliceerde studies kunnen leiden tot nodeloze duplicatie van onderzoeksinspanningen, en kunnen doorgaans niet gebruikt worden in de klinische praktijk.

In **Hoofdstuk 2** bevestigden we de resultaten die beschreven staan in het vorige hoofdstuk in een steekproef van 399 diagnostische accuratessestudies die waren gepresenteerd op de jaarlijkse bijeenkomst van de 'Association for Research in Vision and Ophthalmology' (ARVO). Hiervan bereikte slechts 57% publicatie in een 'peer reviewed' tijdschrift binnen vijf jaar na presentatie. We konden geen statistisch significante associaties vaststellen tussen de kans op publicatie en de gerapporteerde schattingen van sensitiviteit, specificiteit, 'area under the receiver operating characteristic curve' (AUC) en diagnostische odds ratio. Abstracts die gepresenteerd worden op wetenschappelijk congressen kunnen dus een aanvullende bron van ongepubliceerde diagnostische accuratessestudies zijn, maar we vonden geen bewijs dat het niet opnemen van ongepubliceerde studies in een systematisch literatuuroverzicht zal leiden tot 'reporting bias'.

In **Hoofdstuk 3** demonstreerden we dat 'reporting bias' wel degelijk voor zou kunnen komen in het veld van diagnostisch testen. In een steekproef van 756 gepubliceerde diagnostische accuratessestudies vonden we dat de tijd tussen voltooiing van de inclusie van proefpersonen en publicatie significant samenhing met de gerapporteerde schattingen van sensitiviteit, specificiteit en Youden's index: hoe beter de prestaties van de test, des te sneller raakte het onderzoek gepubliceerd. Dit resulteerde in een relatieve vertraging van twee maanden voor de publicatie van studies die een sensitiviteit onder de mediaan rapporteerden, een relatieve vertraging van vijf maanden voor studies die een specificiteit onder de mediaan rapporteerden, en een relatieve vertraging van vijf maanden voor studies met een Youden's index onder de mediaan, dat alles vergeleken met studies die schattingen van deze accuratessematen boven de mediaan naar buiten brachten. Deze vertragingen traden op in de fase tussen het afronden van de studie en het insturen van het studierapport naar het publicerende tijdschrift, en niet tussen insturen en publicatie. Vertragingen in de publicatie van studies met minder veelbelovende resultaten kan zo leiden tot 'reporting bias' in systematisch literatuuronderzoek.

Deel B: Prospectieve registratie van studieprotocollen

Registratie van studies wordt gezien als een belangrijke preventieve maatregel tegen de nadelige gevolgen van het niet publiceren van onderzoek. De 'International Committee of Medical Journal Editors' (ICMJE) heeft om die reden prospectieve registratie van klinische trials als een voorwaarde voor publicatie gesteld. Met het onderzoek dat wordt beschreven in **Deel B** van dit proefschrift wilden we evalueren in welke mate diagnostische accuratessestudies momenteel geregistreerd worden, en in hoeverre tijdschriften zich houden aan ICMJE's beleid betreffende trialregistratie.

In **Hoofdstuk 4** toonden we aan dat diagnostische accuratessestudies zelden worden geregistreerd. In een steekproef van 351 gepubliceerde diagnostische accuratessestudies konden we voor slechts 15% een bijpassend registratienummer vinden. Hiervan was ook nog eens slechts 27% geregistreerd vóór aanvang van de studie. Dit illustreert dat het op dit moment moeilijk is om te profiteren van de voordelen van prospectieve registratie in diagnostisch onderzoek.

In **Hoofdstuk 5** vonden we dat veel tijdschriften zich slecht houden aan ICMJE's beleid betreffende trialregistratie. In een analyse van de auteursinstructies van 747 tijdschriften vonden we dat slechts 51% expliciet vermelde dat trialregistratie verplicht was. En in een enquête onder tijdschriftredacteuren antwoorde 50% van de 232 respondenten dat trialregistratie verplicht was bij hun tijdschrift. Slechts 18% vergeleek ingestuurde artikelen met het geregistreerde protocol om eventuele discrepanties te identificeren, en 67% liet ook retrospectief geregistreerde studies in aanmerking komen voor publicatie. Deze resultaten kunnen een verklaring geven voor de bevindingen van eerdere studies, die aantoonden dat veel gepubliceerde trials van therapeutische interventies nog altijd niet geregistreerd worden, hoewel het aantal geregistreerde studies zeker toeneemt in de tijd.

Deel C: Informatieve rapportage van studierapporten

Lezers van gepubliceerde verslagen van diagnostische accuratessestudies hebben vaak moeite met het adequaat interpreteren van de studiebevindingen, omdat over belangrijke elementen incompleet, vaag, of helemaal geen verslag wordt uitgebracht. De STARD ('Standards for Reporting of Diagnostich Accuracy Studies') verklaring, die werd geïntroduceerd in 2003, biedt auteurs en redacteuren hulp bij het rapporteren van dit type onderzoek. Het doel van **Deel C** was om na te gaan in welke mate de compleetheid van rapportage van diagnostische accuratessestudies door de jaren heen is verbeterd, en om de eventuele tekortkomingen in de huidige staat van rapportage van deze studies te identificeren.

In **Hoofdstuk 6** lieten we zien dat de kwaliteit van rapportage verbeterde in de eerste paar jaar na de lancering van STARD, maar slechts in beperkte mate. We stelden een systematisch literatuuroverzicht samen van onderzoeksprojecten die hadden geëvalueerd hoe goed gepubliceerde onderzoeksverslagen voldeden aan de STARD criteria. Van de 16 geïncludeerde evaluaties, die samen de volledigheid van rapportage van 1.496 studierapporten analyseerden, concludeerden op één na allen dat het navolgen van de STARD-richtlijnen slecht, matig, of suboptimaal was, of hoe dan ook verbetering behoefte. Tussen de verschillende evaluaties varieerde het gemiddelde aantal gerapporteerde items van 9,1 tot 14,3, op een totaal van 25 STARD items. In een meta-analyse vonden we dat ná de lancering van STARD gemiddeld 1,41 méér items gerapporteerd werden dan voorheen. We konden echter geen uitspraken doen over de huidige staat van rapportage, noch of deze initiële verbetering in de volledigheid van rapportage zich in de tijd heeft doorgezet. Het overgrote deel van de geëvalueerde studierapporten was namelijk vóór 2007 gepubliceerd; dat is ruwweg 7 jaar voordat we ons overzicht samenstelden.

In **Hoofdstuk 7** bevestigden we dat de initiële verbetering in volledigheid van rapportage na de lancering van STARD, die beschreven wordt in het vorige hoofdstuk, zich in de tijd heeft doorgezet, maar we moesten ook concluderen dat er flinke ruimte voor verbetering blijft. In een steekproef van 112 verslagen van diagnostische accuratessestudies, gepubliceerd in diverse tijdschriften met een hoge impact factor, evalueerden we hoe goed deze voldeden aan STARD. Ook vergeleken we onze bevindingen met die van twee eerder gepubliceerde evaluaties van studierapporten gepubliceerd in dezelfde tijdschriften. Vergeleken met studies gepubliceerd in 2000 werden er in 2012 gemiddeld 3,4 items méér gerapporteerd;

vergeleken met studies gepubliceerd in 2000 werden er in 2012 gemiddeld 1,7 méér items gerapporteerd. Ondanks deze verbeteringen bleef het gemiddelde aantal gerapporteerde STARD items in 2012 laag: gemiddeld slechts 15,3 van de 25 items. Dit illustreert dat belangrijke informatie over de studie nog altijd vaak niet beschreven wordt.

In **Hoofdstuk 8** en **Hoofdstuk 9** vonden we dat de informativiteit van abstracts van diagnostisch accuratessestudies in tijdschriften en op congressen ook zou kunnen verbeteren. We stelden een lijst op van 21 items samen die we als potentiaal relevant bestempelden, belangrijk genoeg om op te nemen in een abstract. Vervolgens keken we naar de informatie die daadwerkelijk vermeld stond in 103 abstracts in wetenschappelijke tijdschriften en in 125 abstracts die waren gepresenteerd op ARVO. Omdat abstracts doorgaans slechts enkele honderden woorden mogen bevatten, en omdat richtlijnen voor het rapporteren van abstracts van diagnostische accuratessestudies niet beschikbaar waren op het moment van de analyses, was het interessant - maar niet verassend - om vast te stellen dat de gerapporteerde informatie sterk varieerde tussen de abstracts, en dat sommige elementen die wij als cruciaal bestempelden vaak gewoonweg ontbraken in de samenvatting van de studie.

In **Hoofdstuk 10** beschreven we de methode die is gebruikt bij het aanpassen van STARD, wat uiteindelijk resulteerde in STARD 2015. Het doel van de update was om (1) nieuwe items te includeren, gebaseerd op een verbeterd begrip van bronnen van vertekening en variabiliteit, en (2) het gebruik van de lijst te vergemakkelijken door bestaande items te herschikken of opnieuw te verwoorden, en door de consistentie in de gebruikte terminologie met andere veelgebruikte rapportagerichtlijnen te verbeteren. De nieuwe STARD 2015 lijst bestaat uit 30 items. Vergeleken met de vorige versie van STARD zijn er drie van de oorspronkelijke items vereenvoudigd tot twee nieuwe items, vier oorspronkelijke items zijn ondergebracht bij andere items, en zeven nieuwe items zijn toegevoegd.

Deel D: Diagnostische tests in longgeneeskunde

Bij de klinische behandeling van patiënten met astma en longkanker wordt er door clinici gebruik gemaakt van medische tests voor de diagnostiek, maar ook om patiënten te selecteren voor specifieke behandelingen, en om uitspraken te kunnen doen over prognose. Het onderzoek dat in **Deel D** staat gerapporteerd had als doel om de diagnostische accuratesse van een aantal van deze tests te evalueren.

In **Hoofdstuk 11** demonstreerden we dat ‘fraction of exhaled nitric oxide’ (FeNO), bloed eosinofielen en totaal Immunoglobuline E (IgE) onvoldoende accuraat zijn om gebruikt te kunnen worden als opzichzelfstaande surrogaatmarkers voor luchtwegeosinofilie in patiënten met astma. We stelden een systematisch literatuuroverzicht samen van onderzoek naar de diagnostische accuratesse van minimaal invasieve merkers voor het detecteren van luchtwegeosinofilie. Deze drie merkers waren het vaakst geëvalueerd in de geïncludeerde studies, maar hun gemiddelde AUC na meta-analyse was nooit hoger dan 0,81. We bevolen aan om merkers te combineren in een multivariabel klinische predictiemodel met een verbeterde accuratesse, en om afkapwaarden voor deze merkers vast te stellen voor het aantonen dan wel uitsluiten van luchtwegeosinofilie. Het lukte ons om een flinke hoeveelheid ongepubliceerde studieresultaten te includeren in dit review, maar we vonden geen aanwijzingen van 'reporting bias': de schattingen van de accuratesse in gepubliceerde en ongepubliceerde data verschilden niet systematisch.

In **Hoofdstuk 12** volgden we onze eigen aanbevelingen uit het vorige hoofdstuk op. In een groep van 336 volwassen patiënten met astma evalueerden we de diagnostische accuratesse van FeNO, bloed eosinofielen en totaal IgE voor het detecteren van eosinofilie in sputum. In de groep als geheel lagen de AUC's dicht bij de gemiddelde schattingen die we vonden in het systematische literatuuroverzicht in het vorige hoofdstuk. Als de merkers gecombineerd werden met een aantal simpele klinische parameters in een multivariabel klinisch predictiemodel kon de accuratesse worden verbeterd, tot een AUC van 0,89. We rapporteerden ook afkapwaarden van deze merkers voor een hoge sensitiviteit en een hoge specificiteit, welke, indien toegepast, eosinofilie in sputum met een hoge mate van zekerheid konden uitsluiten dan wel aantonen, in bijna de helft van de patiënten.

In **Hoofdstuk 13** onderzochten we de toegevoegde waarde van het gebruik van een gecombineerde benadering van endobronchiale endoscopie (EBUS) en oesofagale endoscopie (EUS) voor stadiëring van de mediastinale lymfeklieren bij patiënten met longkanker. We stelden opnieuw een systematisch literatuuroverzicht samen, en identificeerden 13 relevante studies (2.395 patiënten). We vonden dat het toevoegen van EUS aan EBUS de sensitiviteit voor de detectie van metastases in de mediastinale lymfeklieren met gemiddeld 0,12 verhoogde, en dat het toevoegen van EBUS aan EUS de sensitiviteit met gemiddeld 0,22 verhoogde. Dat leidde tot een totale gemiddelde sensitiviteit van 0,86, bij een negatief voorspellende waarde van 0,92. De meeste huidige klinische richtlijnen over de stadiëring van longkanker bevelen aan om endosonografie hiervoor te gebruiken, maar verduidelijken vaak niet of dit met EBUS of met EUS moet gebeuren. De resultaten van ons overzicht suggereren juist dat een combinatie van beide tests overwogen moet worden.

In **Hoofdstuk 14** evalueerden we de vijfjaarsoverleving in ASTER ('Assessment of Surgical Staging versus Endosonographic Ultrasound in Lung Cancer: a Randomized Clinical Trial'). Dit is een studie waarin patiënten met longkanker werden gerandomiseerd tussen twee strategieën om mediastinale stadiëring te ondergaan: een chirurgische strategie die bestond uit mediastinoscopie, of een endoscopische strategie die bestond uit de gecombineerde benadering van EBUS en EUS, gevolgd door mediastinoscopie als de endoscopie negatief was. In ASTER vond men dat de endoscopische strategie aanzienlijk sensitiever was voor het detecteren van metastases in de mediastinale lymfeklieren: de sensitiviteit was 94%, tegenover 79% voor de chirurgische strategie. Aangezien een betere stadiëring zou moeten leiden tot een betere behandelkeuze namen we aan dat de endoscopische strategie ook een gunstig effect had op overleving. De overleving na vijf jaar was echter 35%, en dat voor beide strategieën. Deze bevindingen tonen aan dat een betere accuratesse van een test niet per definitie vertaald kan worden naar een aanzienlijke verbetering in patiëntrelevante uitkomsten.

Slotopmerkingen

De Verklaring van Helsinki schrijft voor dat onderzoekers hun studies moeten registreren, dat zij hun studieresultaten publiekelijk beschikbaar moet maken, en dat zij complete en accurate studierapporten moeten produceren.[7] Er is geen reden om te denken dat diagnostische accuratessestudies uitgezonderd zijn van deze ethische verplichtingen.

De voorbije jaren zagen we aanzienlijke verbeteringen in de rapportage en verspreiding van de resultaten van biomedisch onderzoek. De proportie van geregistreerde trials is sterk gegroeid,[68] er is beleid geïmplementeerd om onderzoekers aan te zetten tot het publiceren van resultaten,[82] en tijdschriften zien in toenemende mate het belang van het verplichten van het gebruik van rapportagerichtlijnen.[282]

Hoewel het onderzoek naar de diagnostische accuratesse van tests in al deze processen wat achter lijkt te lopen, tonen sommige van de bevindingen uit dit proefschrift aan dat er ruimte is voor bescheiden optimisme. Verschillende auteurs zijn al begonnen met het registreren van hun diagnostische accuratessestudies, ondanks het feit dat de meeste tijdschriften dit nog niet verplicht stellen. De kwaliteit van de rapportage verbetert langzaam maar zeker. En bij het samenstellen van onze systematische literatuuroverzichten naar de accuratesse van tests in de longgeneeskunde waren we blij verrast door het grote aantal onderzoekers dat bereid was om ongepubliceerde data te delen zodat we die konden analyseren en opnemen in onze synthese van de beschikbare resultaten.

Het is cruciaal dat er de komende jaren aanvullende stappen genomen worden om dit alles verder te verbeteren. Allen die beroepsmatig betrokken zijn bij biomedisch onderzoek delen een gezamenlijke en dus ook persoonlijke verantwoordelijkheid om de waarde en de betekenis van wetenschap verder te vergroten.[25]

References

1. Graber ML. The incidence of diagnostic error in medicine. *BMJ quality & safety.* 2013;22 Suppl 2:ii21-ii27.
2. Cason B, Rostas J, Simmons J, Frotan MA, Brevard SB, Gonzalez RP. Thoracolumbar spine clearance: Clinical examination for patients with distracting injuries. *Journal of trauma and acute care surgery.* 2016;80(1):125-130.
3. Inaba K, DuBose JJ, Barmparas G, et al. Clinical examination is insufficient to rule out thoracolumbar spine injuries. *Journal of trauma.* 2011;70(1):174-179.
4. Martin E. *Concise medical dictionary (9 ed.).* Oxford University Press; 2015.
5. Clarke Jr. DS. Chapter two: The greek-medieval period. *Sources of semiotic: readings with commentary from antiquity to the present*: Southern Illinois University Press; 1990.
6. Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clinical chemistry.* 2012;58(9):1292-1301.
7. World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA.* 2013;310(20):2191-2194.
8. Chan AW, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet.* 2014;383(9913):257-266.
9. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet.* 2014;383(9913):267-276.
10. Smith R, Rennie D. Evidence-based medicine--an oral history. *JAMA.* 2014;311(4):365-367.
11. Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials. *The Cochrane database of systematic reviews.* 2007(2):MR000011.
12. Schmucker C, Schell LK, Portalupi S, et al. Extent of non-publication in cohorts of studies approved by research ethics committees or included in trial registries. *PLoS one.* 2014;9(12):e114023.
13. Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS one.* 2013;8(7):e66844.
14. Scherer RW, Langenberg P, von EE. Full publication of results initially presented in abstracts. *The Cochrane database of systematic reviews.* 2007(2):MR000005.
15. Song F, Parekh-Bhurke S, Hooper L, et al. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC medical research methodology.* 2009;9:79.
16. Korevaar DA, Hooft L. [The danger of unpublished trial results]. *Nederlands tijdschrift voor geneeskunde.* 2014;158:A7400.
17. Sutton AJ, Egger M, Moher D. Chapter 10: Addressing reporting bias. In: Higgins JP, Green S, eds. *Cochrane handbook for systematic reviews of interventions.* 5.1.0. ed: The Cochrane Collaboration; 2011.
18. Hooft L, Bossuyt PM. Prospective registration of marker evaluation studies: time to act. *Clinical chemistry.* 2011;57(12):1684-1686.
19. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clinical chemistry.* 2008;54(7):1101-1103.
20. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of clinical epidemiology.* 2005;58(9):882-893.
21. Dickersin K, Rennie D. The evolution of trial registries and their use to assess the clinical trial enterprise. *JAMA.* 2012;307(17):1861-1864.
22. Zarin DA, Tse T, Ide NC. Trial Registration at ClinicalTrials.gov between May and October 2005. *New England journal of medicine.* 2005;353(26):2779-2787.
23. Hooft L, Assendelft WJ, Hoeksema HL, Scholten RJ. [A national prospective trial register for randomised controlled trials: ethical and practical necessity]. *Nederlands tijdschrift voor geneeskunde.* 2004;148(38):1866-1869.
24. De Angelis CD, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *JAMA.* 2004;292(11):1363-1364.
25. Moher D, Glasziou P, Chalmers I, et al. Increasing value and reducing waste in biomedical research: who's listening? *Lancet.* 2016;387(10027):1573-1586.
26. Smidt N, Rutjes AW, van der Windt DA, et al. Quality of reporting of diagnostic accuracy studies. *Radiology.* 2005;235(2):347-353.
27. Whiting PF, Rutjes AW, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of clinical epidemiology.* 2013;66(10):1093-1104.
28. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine.* 2011;155(8):529-536.
29. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ.* 2003;326(7379):41-44.

30. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical chemistry.* 2003;49(1):7-18.

31. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ.* 2015;351:h5527.

32. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention 2014. *ginasthma.org (accessed Sept 23, 2015).* 2014.

33. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA: a cancer journal for clinicians.* 2015;65(2):87-108.

34. Hekking PP, Bel EH. Developing and emerging clinical asthma phenotypes. *Journal of allergy and clinical immunology: in practice.* 2014;2(6):671-680.

35. Vansteenkiste J, De Ruysscher D, Eberhardt WE, et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of oncology.* 2013;24 Suppl 6:vi89-98.

36. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA.* 1990;263(10):1385-1389.

37. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA.* 2004;291(20):2457-2465.

38. Sridharan L, Greenland P. Editorial policies and publication bias: the importance of negative studies. *Archives of internal medicine.* 2009;169(11):1022-1023.

39. Chalmers I. Underreporting research is scientific misconduct. *JAMA.* 1990;263(10):1405-1408.

40. McGauran N, Wieseler B, Kreis J, Schuler YB, Kolsch H, Kaiser T. Reporting bias in medical research - a narrative review. *Trials.* 2010;11:37.

41. Brazzelli M, Lewis SC, Deeks JJ, Sandercock PA. No evidence of bias in the process of publication of diagnostic accuracy studies in stroke submitted as abstracts. *Journal of clinical epidemiology.* 2009;62(4):425-430.

42. Bourgeois FT, Murthy S, Mandl KD. Outcome reporting among drug trials registered in ClinicalTrials.gov. *Annals of internal medicine.* 2010;153(3):158-166.

43. Ross JS, Mulvey GK, Hines EM, Nissen SE, Krumholz HM. Trial publication after registration in ClinicalTrials.Gov: a cross-sectional analysis. *PLoS medicine.* 2009;6(9):e1000144.

44. van de Wetering FT, Scholten RJ, Haring T, Clarke M, Hooft L. Trial registration numbers are underreported in biomedical publications. *PLoS one.* 2012;7(11):e49599.

45. Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. *BMJ.* 2012;344:d7292.

46. Gordon D, Taddei-Peters W, Mascette A, Antman M, Kaufmann PG, Lauer MS. Publication of trials funded by the National Heart, Lung, and Blood Institute. *New England journal of medicine.* 2013;369(20):1926-1934.

47. Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA.* 2009;302(9):977-984.

48. Hannink G, Gooszen HG, Rovers MM. Comparison of registered and published primary outcomes in randomized clinical trials of surgical interventions. *Annals of surgery.* 2013;257(5):818-823.

49. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology.* 2013;267(2):581-588.

50. Reitsma JB, Moons KG, Bossuyt PM, Linnet K. Systematic reviews of studies quantifying the accuracy of diagnostic tests and markers. *Clinical chemistry.* 2012;58(11):1534-1545.

51. Hua F, Walsh T, Glenny AM, Worthington H. Thirty percent of abstracts presented at dental conferences are published in full: a systematic review. *Journal of clinical epidemiology.* 2016.

52. Sterne JAC, Egger M, Moher D. Chapter 10: Addressing reporting biases. *Cochrane handbook for systematic reviews of interventions.* Version 5.1.0. ed: The Cochrane Collaboration; 2011.

53. Korevaar DA, Ochodo EA, Bossuyt PM, Hooft L. Publication and reporting of test accuracy studies registered in ClinicalTrials.gov. *Clinical chemistry.* 2014;60(4):651-659.

54. Wilson C, Kerr D, Noel-Storr A, Quinn TJ. Associations with publication and assessing publication bias in dementia diagnostic test accuracy studies. *International journal of geriatric psychiatry.* 2015;30(12):1250-1256.

55. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ.* 2006;332(7550):1127-1129.

56. van Enst WA, Ochodo E, Scholten RJ, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. *BMC medical research methodology.* 2014;14:70.

57. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology.* 2003;56(11):1129-1135.

58. Walter SD, Sinuff T. Studies reporting ROC curves of diagnostic and prediction data can be incorporated into meta-analyses using corresponding odds ratios. *Journal of clinical epidemiology.* 2007;60(5):530-534.

59. Kho ME, Brouwers MC. Conference abstracts of a new oncology drug do not always lead to full publication: proceed with caution. *Journal of clinical epidemiology.* 2009;62(7):752-758.

60. Juzych MS, Shin DH, Coffey J, Juzych L, Shin D. Whatever happened to abstracts from different sections of the association for research in vision and ophthalmology? *Investigative ophthalmology & visual science.* 1993;34(5):1879-1882.

61. Scherer RW, Dickersin K, Langenberg P. Full publication of results initially presented in abstracts. A meta-analysis. *JAMA.* 1994;272(2):158-162.

62. Saldanha IJ, Scherer RW, Rodriguez-Barraquer I, Jampel HD, Dickersin K. Dependability of results in conference abstracts of randomized controlled trials in ophthalmology and author financial conflicts of interest as a factor associated with full publication. *Trials.* 2016;17(1):213.

63. Korevaar DA, Cohen JF, Hooft L, Bossuyt PM. Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies. *Journal of clinical epidemiology.* 2015;68(6):708-715.

64. Scherer RW, Ugarte-Gil C, Schmucker C, Meerpohl JJ. Authors report lack of time as main reason for unpublished research presented at biomedical conferences: a systematic review. *Journal of clinical epidemiology.* 2015;68(7):803-810.

65. Wager E, Williams P. "Hardly worth the effort"? Medical journals' policies and their editors' and publishers' views on trial registration and publication bias: quantitative and qualitative study. *BMJ.* 2013;347:f5248.

66. Lumbreras B, Parker LA, Porta M, Pollan M, Ioannidis JP, Hernandez-Aguado I. Overinterpretation of clinical applicability in molecular diagnostic research. *Clinical chemistry.* 2009;55(4):786-794.

67. Hooft L, Korevaar DA, Molenaar N, Bossuyt PM, Scholten RJ. Endorsement of ICMJE's Clinical Trial Registration Policy: a survey among journal editors. *Netherlands journal of medicine.* 2014;72(7):349-355.

68. Viergever RF, Li K. Trends in global clinical trial registration: an analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013. *BMJ open.* 2015;5(9):e008932.

69. Korevaar DA, Bossuyt PM, Hooft L. Infrequent and incomplete registration of test accuracy studies: analysis of recent study reports. *BMJ open.* 2014;4(1):e004596.

70. Altman DG. The time has come to register diagnostic and prognostic research. *Clinical chemistry.* 2014;60(4):580-582.

71. Rifai N, Bossuyt PM, Ioannidis JP, et al. Registering diagnostic and prognostic trials of tests: is it the right thing to do? *Clinical chemistry.* 2014;60(9):1146-1152.

72. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA.* 1998;279(4):281-286.

73. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ.* 1997;315(7109):640-645.

74. Sune P, Sune JM, Montoro JB. Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. *PLoS one.* 2013;8(1):e54583.

75. de Vet HC, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Chapter 7: Searching for studies. In: Deeks JJ, Bossuyt PM, Gatsonis CA, eds. *Cochrane handbook for systematic reviews of diagnostic test accuracy*. Version 1.0.0. ed: The Cochrane Collaboration; 2008.

76. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32-35.

77. Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evidence based medicine.* 2014;19(2):47-54.

78. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Annals of internal medicine.* 2004;140(3):189-202.

79. Ross JS, Mocanu M, Lampropulos JF, Tse T, Krumholz HM. Time to publication among completed clinical trials. *JAMA internal medicine.* 2013;173(9):825-828.

80. Korevaar DA, Cohen JF, Spijker R, et al. Reported estimates of diagnostic accuracy in ophthalmology conference abstracts were not associated with full-text publication. *Journal of clinical epidemiology.* 2016: E-published ahead of print.

81. Cohen JF, Korevaar DA, Wang J, Leeflang MM, Bossuyt PM. Meta-epidemiological study showed frequent time trends in summary estimates from meta-analyses of diagnostic accuracy studies. *Journal of clinical epidemiology.* 2016: E-published ahead of print.

82. Prayle AP, Hurley MN, Smyth AR. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. *BMJ.* 2012;344:d7373.

83. De Angelis CD, Drazen JM, Frizelle FA, et al. Is this clinical trial fully registered?--A statement from the International Committee of Medical Journal Editors. *New England journal of medicine.* 2005;352(23):2436-2438.

84. Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PM. Quantifying the added value of a diagnostic test or marker. *Clinical chemistry.* 2012;58(10):1408-1417.

85. The registration of observational studies--when metaphors go bad. *Epidemiology.* 2010;21(5):607-609.

86. Lash TL. Preregistration of study protocols is unlikely to improve the yield from our science, but other strategies might. *Epidemiology.* 2010;21(5):612-613.

87. Vandenbroucke JP. Registering observational research: second thoughts. *Lancet.* 2010;375(9719):982-983.

88. Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: is it time? *Canadian Medical Association journal.* 2010;182(15):1638-1642.

89. Food and Drug Administration Amendments Act of 2007.

90. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *Journal of clinical epidemiology.* 2000;53(1):65-69.

91. 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(5):792-797.

92. Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. *Radiology.* 2008;248(3):817-823.

93. Should protocols for observational research be registered? *Lancet.* 2010;375(9712):348.

94. Loder E, Groves T, Macauley D. Registration of observational studies. *BMJ.* 2010;340:c950.

95. Pearce N. Registration of protocols for observational research is unnecessary and would do more harm than good. *Occupational and environmental medicine.* 2011;68(2):86-88.

96. Chavers S, Fife D, Wacholtz M, Stang P, Berlin J. Registration of Observational Studies: perspectives from an industry-based epidemiology group. *Pharmacoepidemiology and drug safety.* 2011;20(10):1009-1013.

97. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ.* 2005;330(7494):753.

98. Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association journal.* 2004;171(7):735-740.

99. Scherer RW, Huynh L, Ervin AM, Taylor J, Dickersin K. ClinicalTrials.gov registration can supplement information in abstracts for systematic reviews: a comparison study. *BMC medical research methodology.* 2013;13:79.

100. van Enst WA, Scholten RJ, Hooft L. Identification of additional trials in prospective trial registers for Cochrane systematic reviews. *PLoS one.* 2012;7(8):e42812.

101. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ.* 2010;340:c869.

102. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ.* 2010;340:c332.

103. Killeen S, Sourallous P, Hunter IA, Hartley JE, Grady HL. Registration rates, adequacy of registration, and a comparison of registered and published primary outcomes in randomized controlled trials published in surgery journals. *Annals of surgery.* 2014;259(1):193-196.

104. Viergever RF, Karam G, Reis A, Ghersi D. The quality of registration of clinical trials: still a problem. *PLoS one.* 2014;9(1):e84727.

105. Mathieu S, Chan AW, Ravaud P. Use of trial register information during the peer review process. *PLoS one.* 2013;8(4):e59910.

106. Altman DG. Endorsement of the CONSORT statement by high impact medical journals: survey of instructions for authors. *BMJ.* 2005;330(7499):1056-1057.

107. Hopewell S, Altman DG, Moher D, Schulz KF. Endorsement of the CONSORT Statement by high impact factor medical journals: a survey of journal editors and journal 'Instructions to Authors'. *Trials.* 2008;9:20.

108. Turner L, Shamseer L, Altman DG, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *The Cochrane database of systematic reviews.* 2012;11:MR000030.

109. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282(11):1061-1066.

110. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA.* 1995;274(8):645-651.

111. Plint AC, Moher D, Morrison A, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical journal of Australia.* 2006;185(5):263-267.

112. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA.* 2001;285(15):1992-1995.

113. Bossuyt PM. STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. *Radiology.* 2008;248(3):713-714.

114. Samaan Z, Mbuagbaw L, Kosa D, et al. A systematic scoping review of adherence to reporting guidelines in health care literature. *Journal of multidisciplinary healthcare.* 2013;6:169-188.

115. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC medical research methodology.* 2007;7:10.

116. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials.* 1986;7(3):177-188.

117. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557-560.

118. Selman TJ, Morris RK, Zamora J, Khan KS. The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: application of the STARD criteria. *BMC womens health.* 2011;11:8.

119. Areia M, Soares M, Dinis-Ribeiro M. Quality reporting of endoscopic diagnostic studies in gastrointestinal journals: where do we stand on the use of the STARD and CONSORT statements? *Endoscopy.* 2010;42(2):138-147.

120. Coppus SF, van d, V, Bossuyt PM, Mol BW. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. *Fertility and sterility.* 2006;86(5):1321-1329.

121. Fontela PS, Pant PN, Schiller I, Dendukuri N, Ramsay A, Pai M. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. *PLoS one.* 2009;4(11):e7753.

122. Freeman K, Szczepura A, Osipenko L. Non-invasive fetal RHD genotyping tests: a systematic review of the quality of reporting of diagnostic accuracy in published studies. *European journal of obstetrics & gynecology and reproductive biology.* 2009;142(2):91-98.

123. Gomez SN, Hernandez-Aguado I, Lumbreras B. [Observacional study: evaluation of the diagnostic research methodology in Spain after STARD publication]. *Medicina clínica (Barcelona).* 2009;133(8):302-310.

124. Johnson ZK, Siddiqui MA, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies of optical coherence tomography in glaucoma. *Ophthalmology.* 2007;114(9):1607-1612.

125. Lumbreras B, Jarrin I, Hernandez A, I. Evaluation of the research methodology in genetic, molecular and proteomic tests. *Gaceta sanitaria.* 2006;20(5):368-373.

126. Paranjothy B, Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using scanning laser polarimetry. *Journal of glaucoma.* 2007;16(8):670-675.

127. Rama KR, Poovali S, Apsingi S. Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal. *Clinical orthopaedics and related research.* 2006;447:237-246.

128. Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. *Investigative ophthalmology & visual science.* 2006;47(6):2317-2323.

129. Siddiqui MA, Azuara-Blanco A, Burr J. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. *British journal of ophthalmology.* 2005;89(3):261-265.

130. Zafar A, Khan GI, Siddiqui MA. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: a systematic review. *Clinical & experimental ophthalmology.* 2008;36(6):537-542.

131. Zintzaras E, Papathanasiou AA, Ziogas DC, Voulgarelis M. The reporting quality of studies investigating the diagnostic accuracy of anti-CCP antibody in rheumatoid arthritis and its impact on diagnostic estimates. *BMC musculoskelet disorders.* 2012;13:113.

132. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1(8476):307-310.

133. Ochodo EA, Bossuyt PM. Reporting the accuracy of diagnostic tests: the STARD initiative 10 years on. *Clinical chemistry.* 2013;59(6):917-919.

134. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clinical chemistry.* 2005;51(8):1335-1341.

135. Smidt N, Rutjes AW, van der Windt DA, et al. Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. *BMC medical research methodology.* 2006;6:12.

136. Hopewell S, Collins GS, Boutron I, et al. Impact of peer review on reports of randomised trials published in open peer review journals: retrospective before and after study. *BMJ.* 2014;349:g4145.

137. Cobo E, Cortes J, Ribera JM, et al. Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial. *BMJ.* 2011;343:d6783.

138. Hopewell S, Clarke M, Moher D, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS medicine.* 2008;5(1):e20.

139. Beller EM, Glasziou PP, Altman DG, et al. PRISMA for Abstracts: reporting systematic reviews in journal and conference abstracts. *PLoS medicine.* 2013;10(4):e1001419.

140. Soffer A. Abstracts of clinical investigations. A new and standardized format. *Chest.* 1987;92(3):389-390.

141. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. A proposal for more informative abstracts of clinical articles. *Annals of internal medicine.* 1987;106:598-604.

142. Comans ML, Overbeke AJ. [The structured summary: a tool for reader and author]. *Nederlands tijdschrift voor geneeskunde.* 1990;134(48):2338-2343.

143. Taddio A, Pain T, Fassos FF, Boon H, Ilersich AL, Einarson TR. Quality of nonstructured and structured abstracts of original research articles in the British Medical Journal, the Canadian Medical Association Journal and the Journal of the American Medical Association. *Canadian Medical Association journal.* 1994;150(10):1611-1615.

144. Berwanger O, Ribeiro RA, Finkelsztejn A, et al. The quality of reporting of trial abstracts is suboptimal: Survey of major general medical journals. *Journal of clinical epidemiology.* 2009;62(4):387-392.

145. Ghimire S, Kyung E, Kang W, Kim E. Assessment of adherence to the CONSORT statement for quality of reports on randomized controlled trial abstracts from four high-impact general medical journals. *Trials.* 2012;13:77.

146. Hopewell S, Ravaud P, Baron G, Boutron I. Effect of editors' implementation of CONSORT guidelines on the reporting of abstracts in high impact medical journals: interrupted time series analysis. *BMJ.* 2012;344:e4178.

147. Estrada CA, Bloch RM, Antonacci D, et al. Reporting and concordance of methodologic criteria between abstracts and articles in diagnostic test studies. *Journal of general internal medicine.* 2000;15(3):183-187.

148. Korevaar DA, Wang J, van Enst WA, et al. Reporting Diagnostic Accuracy Studies: Some Improvements after 10 Years of STARD. *Radiology.* 2015;274(3):781-789.

149. Deeks JJ, Altman DG. Inadequate reporting of controlled trials as short reports. *Lancet.* 1998;352(9144):1908.

150. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. *Annals of internal medicine.* 1990;113(1):69-76.

151. Kho ME, Eva KW, Cook DJ, Brouwers MC. The Completeness of Reporting (CORE) index identifies important deficiencies in observational study conference abstracts. *Journal of clinical epidemiology.* 2008;61(12):1241-1249.

152. Timmer A, Sutherland LR, Hilsden RJ. Development and evaluation of a quality score for abstracts. *BMC medical research methodology.* 2003;3:2.

153. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA.* 2010;303(20):2058-2064.

154. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association journal.* 2013;185(11):E537-E544.

155. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS medicine.* 2010;7(2):e1000217.

156. Can OS, Yilmaz AA, Hasdogan M, et al. Has the quality of abstracts for randomised controlled trials improved since the release of Consolidated Standards of Reporting Trial guideline for abstract reporting? A survey of four high-profile anaesthesia journals. *European journal of anaesthesiology.* 2011;28(7):485-492.

157. Ghimire S, Kyung E, Lee H, Kim E. Oncology trial abstracts showed suboptimal improvement in reporting: a comparative before-and-after evaluation using CONSORT for Abstract guidelines. *Journal of clinical epidemiology.* 2014;67(6):658-666.

158. Virgili G, Menchini F, Casazza G, et al. Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy. *The Cochrane database of systematic reviews.* 2015;1:CD008081.

159. Cohen JF, Korevaar DA, Hooft L, Reitsma JB, bossuyt PM. Development of STARD for Abstracts: essential items in reporting diagnostic accuracy studies in journal or conference abstracts. *equator-network.org/wp-content/uploads/2009/02/STARD-for-Abstracts-protocol.pdf (accessed July 1, 2016).*

160. Hirst A, Altman DG. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLoS one.* 2012;7(4):e35621.

161. Kunath F, Grobe HR, Rucker G, et al. Do journals publishing in the field of urology endorse reporting guidelines? A survey of author instructions. *Urologia Internationalis.* 2012;88(1):54-59.

162. Knuppel H, Metz C, Meerpohl JJ, Strech D. How psychiatry journals support the unbiased translation of clinical research. A cross-sectional study of editorial policies. *PLoS one.* 2013;8(10):e75995.

163. Meerpohl JJ, Wolff RF, Niemeyer CM, Antes G, von Elm E. Editorial policies of pediatric journals: survey of instructions for authors. *Archives of pediatrics and adolescent medicine.* 2010;164(3):268-272.

164. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. *Lancet.* 2008;371(9619):1149-1150.

165. Omoumi P, Bafort AC, Dubuc JE, Malghem J, Vande Berg BC, Lecouvet FE. Evaluation of rotator cuff tendon tears: comparison of multidetector CT arthrography and 1.5-T MR arthrography. *Radiology.* 2012;264(3):812-822.

166. Simel DL, Rennie D, Bossuyt PM. The STARD statement for reporting diagnostic accuracy studies: application to the history and physical examination. *Journal of general internal medicine.* 2008;23(6):768-774.

167. Noel-Storr AH, McCleery JM, Richard E, et al. Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem Initiative. *Neurology.* 2014;83(4):364-373.

168. Editors PM. From Checklists to Tools: Lowering the Barrier to Better Research Reporting. *PLoS medicine.* 2015;12(11):e1001910.

169. Barnes C, Boutron I, Giraudeau B, Porcher R, Altman DG, Ravaud P. Impact of an online writing aid tool for writing a randomized trial report: the COBWEB (Consort-based WEB tool) randomized controlled trial. *BMC medicine.* 2015;13:221.

170. McGrath KW, Icitovic N, Boushey HA, et al. A large subgroup of mild-to-moderate asthma is persistently noneosinophilic. *American journal of respiratory and critical care medicine.* 2012;185(6):612-619.

171. Petsky HL, Kynaston JA, Turner C, et al. Tailored interventions based on sputum eosinophils versus clinical symptoms for asthma in children and adults. *The Cochrane database of systematic reviews.* 2007(2):CD005603.

172. Chung KF, Wenzel SE, Brozek JL, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. *European respiratory journal.* 2014;43(2):343-373.

173. Ten Brinke A, Zwinderman AH, Sterk PJ, Rabe KF, Bel EH. Factors associated with persistent airflow limitation in severe asthma. *American journal of respiratory and critical care medicine.* 2001;164(5):744-748.

174. Lovett CJ, Whitehead BF, Gibson PG. Eosinophilic airway inflammation and the prognosis of childhood asthma. *Clinical & experimental allergy.* 2007;37(11):1594-1601.

175. Pavord ID, Korn S, Howarth P, et al. Mepolizumab for severe eosinophilic asthma (DREAM): a multicentre, double-blind, placebo-controlled trial. *Lancet.* 2012;380(9842):651-659.

176. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working G. Systematic reviews of diagnostic test accuracy. *Annals of internal medicine.* 2008;149(12):889-897.

177. Kim MA, Shin YS, Pham le D, Park HS. Adult asthma biomarkers. *Current opinion in allergy and clinical immunology.* 2014;14(1):49-54.

178. Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine.* New York City, NY: Wiley; 2002.

179. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine.* 2002;21(11):1539-1558.

180. Bacci E, Cianchetti S, Bartoli M, et al. Low sputum eosinophils predict the lack of response to beclomethasone in symptomatic asthmatic patients. *Chest.* 2006;129(3):565-572.

181. Berry MA, Shaw DE, Green RH, Brightling CE, Wardlaw AJ, Pavord ID. The use of exhaled nitric oxide concentration to identify eosinophilic airway inflammation: an observational study in adults with asthma. *Clinical & experimental allergy* 2005;35(9):1175-1179.

182. Amorim MM, Araruna A, Caetano LB, Cruz AC, Santoro LL, Fernandes AL. Nasal eosinophilia: an indicator of eosinophilic inflammation in asthma. *Clinical & experimental allergy.* 2010;40(6):867-874.

183. Choi JS, Jang AS, Park SW, et al. Role of neutrophils in persistent airway obstruction due to refractory asthma. *Respirology.* 2012;17(2):322-329.

184. De Carvalho-Pinto RM, Cukier A, Angelini L, et al. Clinical characteristics and possible phenotypes of an adult severe asthma population. *Respiratory medicine.* 2012;106(1):47-56.

185. Greulich T, Haldar P, Agbetile J, et al. FENO and blood eosinophil count as surrogate markers of eosinophilic airway inflammation in patients with severe asthma. *European respiratory journal.* 2012;36:1005S.

186. Hastie AT, Moore WC, Li H, et al. Biomarker surrogates do not accurately predict sputum eosinophil and neutrophil percentages in asthmatic subjects. *Journal of allergy and clinical immunology.* 2013;132(1):72-80.

187. Hillas G, Kostikas K, Mantzouranis K, et al. Exhaled nitric oxide and exhaled breath condensate pH as predictors of sputum cell counts in optimally treated asthmatic smokers. *Respirology.* 2011;16(5):811-818.

188. Ibrahim B, Marsden P, Smith JA, Custovic A, Nilsson M, Fowler SJ. Breath metabolomic profiling by nuclear magnetic resonance spectroscopy in asthma. *Allergy.* 2013;68(8):1050-1056.

189. Ibrahim B, Basanta M, Cadden P, et al. Non-invasive phenotyping using exhaled volatile organic compounds in asthma. *Thorax.* 2011;66(9):804-809.

190. Jia G, Erickson RW, Choy DF, et al. Periostin is a systemic biomarker of eosinophilic airway inflammation in asthmatic patients. *Journal of allergy and clinical immunology.* 2012;130(3):647-654.

191. Lemiere C, Ernst P, Olivenstein R, et al. Airway inflammation assessed by invasive and noninvasive means in severe asthma: eosinophilic and noneosinophilic phenotypes. *Journal of allergy and clinical immunology.* 2006;118:1033-1039.

192. Liang Z, Zhao H, Lv Y, et al. Moderate accuracy of peripheral eosinophil count for predicting eosinophilic phenotype in steroid-naive non-atopic adult asthmatics. *Internal medicine.* 2012;51(7):717-722.

193. Meijer RJ, Postma DS, Kauffman HF, Arends LR, Koeter GH, Kerstjens HAM. Accuracy of eosinophils and eosinophil cationic protein to predict steroid improvement in asthma. *Clinical & experimental allergy.* 2002;32(7):1096-1103.

194. Schleich FN, Manise M, Sele J, Henket M, Seidel L, Louis R. Distribution of sputum cellular phenotype in a large asthma cohort: predicting factors for eosinophilic vs neutrophilic inflammation. *BMC pulmonary medicine.* 2013;13:11.

195. Schneider A, Schwarzbach J, Faderl B, Welker L, Karsch-Volk M, Jorres RA. FENO measurement and sputum analysis for diagnosing asthma in clinical practice. *Respiratory medicine.* 2013;107(2):209-216.

196. Silkoff PE, Lent AM, Busacker AA, et al. Exhaled nitric oxide identifies the persistent eosinophilic phenotype in severe refractory asthma. *Journal of allergy and clinical immunology.* 2005;116(6):1249-1255.

197. Tseliou E, Bessa V, Hillas G, et al. Exhaled nitric oxide and exhaled breath condensate pH in severe refractory asthma. *Chest.* 2010;138(1):107-113.

198. Wagener A, de Nijs S, Lutter R, et al. External validation of blood eosinophils, FeNO, and serum periostin as surrogates for sputum eosinophils in asthma. *Thorax.* 2014;70(2):115-120.

199. Westerhof GA, Korevaar DA, Amelink M, et al. Biomarkers to identify sputum eosinophilia in different adult asthma phenotypes. *European respiratory journal.* 2015;46(3):688-696.

200. Yap E, Chua WM, Jayaram L, Zeng I, Vandal AC, Garrett J. Can we predict sputum eosinophilia from clinical assessment in patients referred to an adult asthma clinic? *Internal Medicine Journal.* 2013;43(1):46-52.

201. Zhang XY, Simpson JL, Powell H, et al. Full blood count parameters for the detection of asthma inflammatory phenotypes. *Clinical & experimental allergy.* 2014;44(9):1137-1145.

202. Fleming L, Tsartsali L, Wilson N, Regamey N, Bush A. Longitudinal relationship between sputum eosinophils and exhaled nitric oxide in children with asthma. *American journal of respiratory and critical care medicine.* 2013;188(3):400-402.

203. Lex C, Ferreira F, Zacharasiewicz A, et al. Airway eosinophilia in children with severe asthma: Predictive values of noninvasive tests. *American journal of respiratory and critical care medicine.* 2006;174(12):1286-1291.

204. Lex C, Payne DN, Zacharasiewicz A, et al. Sputum induction in children with difficult asthma: safety, feasibility, and inflammatory cell pattern. *Pediatric pulmonology.* 2005;39(4):318-324.

205. Shields MD, Brown V, Stevenson EC, et al. Serum eosinophilic cationic protein and blood eosinophil counts for the prediction of the presence of airways inflammation in children with wheezing. *Clinical & experimental allergy.* 1999;29(10):1382-1389.

206. Sivan Y, Gadish T, Fireman E, Soferman R. The Use of Exhaled Nitric Oxide in the Diagnosis of Asthma in School Children. *Journal of pediatrics.* 2009;155(2):211-216.

207. Toyran M, Bakirtas A, Dogruman-Al F, Turktas I. Airway inflammation and bronchial hyperreactivity in steroid naive children with intermittent and mild persistent asthma. *Pediatric pulmonology.* 2014;49:140-147.

208. Ullmann N, Bossley CJ, Fleming L, Silvestri M, Bush A, Saglani S. Blood eosinophil counts rarely reflect airway eosinophilia in children with severe asthma. *Allergy.* 2013;68(3):402-406.

209. Warke TJ, Fitch PS, Brown V, et al. Exhaled nitric oxide correlates with airway eosinophils in childhood asthma. *Thorax.* 2002;57(5):383-387.

210. Grootendorst DC, Sont JK, Willems LN, et al. Comparison of inflammatory cell counts in asthma: induced sputum vs bronchoalveolar lavage and bronchial biopsies. *Clinical & experimental allergy.* 1997;27(7):769-779.

211. van der Schee MP, Palmay R, Cowan JO, Taylor DR. Predicting steroid responsiveness in patients with asthma using exhaled breath profiling. *Clinical & experimental allergy.* 2013;43(11):1217-1225.

212. Bel EH, Wenzel SE, Thompson PJ, et al. Oral glucocorticoid-sparing effect of mepolizumab in eosinophilic asthma. *New England journal of medicine.* 2014;371(13):1189-1197.

213. Dweik RA, Boggs PB, Erzurum SC, et al. An official ATS clinical practice guideline: interpretation of exhaled nitric oxide levels (FENO) for clinical applications. *American journal of respiratory and critical care medicine.* 2011;184(5):602-615.

214. Ricciardolo FL, Sterk PJ, Gaston B, Folkerts G. Nitric oxide in health and disease of the respiratory system. *Physiological reviews.* 2004;84(3):731-765.

215. Spector SL, Tan RA. Is a single blood eosinophil count a reliable marker for "eosinophilic asthma?". *Journal of asthma.* 2012;49(8):807-810.

216. Mummadi SR, Hatipoglu US, Gupta M, Bossard MK, Xu M, Lang D. Clinically significant variability of serum IgE concentrations in patients with severe asthma. *Journal of asthma.* 2012;49(2):115-120.

217. Pijnenburg MW, Bakker EM, Lever S, Hop WC, De Jongste JC. High fractional concentration of nitric oxide in exhaled air despite steroid treatment in asthmatic children. *Clinical & experimental allergy.* 2005;35(7):920-925.

218. Cowan DC, Taylor DR, Peterson LE, et al. Biomarker-based asthma phenotypes of corticosteroid response. *Journal of allergy and clinical immunology.* 2015;135(4):877-883.

219. Taylor DR. Advances in the clinical applications of exhaled nitric oxide measurements. *Journal of breath research.* 2012;6(4):047102.

220. van Veen IH, Ten BA, Gauw SA, Sterk PJ, Rabe KF, Bel EH. Consistency of sputum eosinophilia in difficult-to-treat asthma: a 5-year follow-up study. *Journal of allergy and clinical immunology.* 2009;124(3):615-617.

221. Bacci E, Latorre M, Cianchetti S, et al. Transient sputum eosinophilia may occur over time in non-eosinophilic asthma and this is not prevented by salmeterol. *Respirology.* 2012;17(8):1199-1206.

222. D'silva L, Cook RJ, Allen CJ, Hargreave FE, Parameswaran K. Changing pattern of sputum cell counts during successive exacerbations of airway disease. *Respiratory medicine.* 2007;101(10):2217-2220.

223. Amelink M, de Groot JC, de Nijs SB, et al. Severe adult-onset asthma: A distinct phenotype. *Journal of allergy and clinical immunology.* 2013;132(2):336-341.

224. Simpson JL, Scott R, Boyle MJ, Gibson PG. Inflammatory subtypes in asthma: assessment and identification using induced sputum. *Respirology.* 2006;11(1):54-61.

225. Cowan DC, Cowan JO, Palmay R, Williamson A, Taylor DR. Effects of steroid therapy on inflammatory cell subtypes in asthma. *Thorax.* 2010;65(5):384-390.

226. Berry M, Morgan A, Shaw DE, et al. Pathological features and inhaled corticosteroid response of eosinophilic and non-eosinophilic asthma. *Thorax.* 2007;62(12):1043-1049.

227. Pavord ID, Bafadhel M. Exhaled nitric oxide and blood eosinophilia: independent markers of preventable risk. *Journal of allergy and clinical immunology.* 2013;132(4):828-829.

228. Schleich FN, Seidel L, Sele J, et al. Exhaled nitric oxide thresholds associated with a sputum eosinophil count >/=3% in a cohort of unselected patients with asthma. *Thorax.* 2010;65(12):1039-1044.

229. Schleich FN, Chevremont A, Paulus V, et al. Importance of concomitant local and systemic eosinophilia in uncontrolled asthma. *European respiratory journal.* 2014;44(1):97-108.

230. Cohn L, Woodruff PG. Update in asthma 2013. *American journal of respiratory and critical care medicine.* 2014;189(12):1487-1493.

231. Brusselle GG, Maes T, Bracke KR. Eosinophils in the spotlight: Eosinophilic airway inflammation in nonallergic asthma. *Nature medicine.* 2013;19(8):977-979.

232. Thomson NC, Chaudhuri R, Heaney LG, et al. Clinical outcomes and inflammatory biomarkers in current smokers and exsmokers with severe asthma. *Journal of allergy and clinical immunology.* 2013;131(4):1008-1016.

233. Pavord ID, Brightling CE, Woltmann G, Wardlaw AJ. Non-eosinophilic corticosteroid unresponsive asthma. *Lancet.* 1999;353(9171):2213-2214.

234. Amelink M, de Nijs SB, de Groot JC, et al. Three phenotypes of adult-onset asthma. *Allergy.* 2013;68(5):674-680.

235. Westerhof GA, Vollema EM, Weersink EJ, Reinartz SM, de Nijs SB, Bel EH. Predictors for the development of progressive severity in new-onset adult asthma. *Journal of allergy and clinical immunology.* 2014;134(5):1051-1056.

236. Bel EH, Sousa A, Fleming L, et al. Diagnosis and definition of severe refractory asthma: an international consensus statement from the Innovative Medicine Initiative (IMI). *Thorax.* 2011;66(10):910-917.

237. Juniper EF, O'Byrne PM, Guyatt GH, Ferrie PJ, King DR. Development and validation of a questionnaire to measure asthma control. *European respiratory journal.* 1999;14(4):902-907.

238. Juniper EF, Bousquet J, Abetz L, Bateman ED, Committee G. Identifying 'well-controlled' and 'not well-controlled' asthma using the Asthma Control Questionnaire. *Respiratory medicine.* 2006;100(4):616-621.

239. Paggiaro PL, Chanez P, Holz O, et al. Sputum induction. *European respiratory journal.* 2002;37:3s-8s.

240. Korevaar DA, Westerhof GA, Wang J, et al. Diagnostic accuracy of minimally invasive markers for detection of airway eosinophilia in asthma: a systematic review and meta-analysis. *Lancet respiratory medicine.* 2015;3(4):290-300.

241. van Veen IH, Ten Brinke A, Sterk PJ, Rabe KF, Bel EH. Airway inflammation in obese and nonobese patients with difficult-to-treat asthma. *Allergy.* 2008;63(5):570-574.

242. Haldar P, Brightling CE, Hargadon B, et al. Mepolizumab and exacerbations of refractory eosinophilic asthma. *New England journal of medicine.* 2009;360(10):973-984.

243. Polosa R, Thomson NC. Smoking and asthma: dangerous liaisons. *European respiratory journal.* 2013;41(3):716-726.

244. Bjermer L, Alving K, Diamant Z, et al. Current evidence and future research needs for FeNO measurement in respiratory diseases. *Respiratory medicine.* 2014;108(6):830-841.

245. Katz LE, Gleich GJ, Hartley BF, Yancey SW, Ortega HG. Blood eosinophil count is a useful biomarker to identify patients with severe eosinophilic asthma. *Annals of the American Thoracic Society.* 2014;11(4):531-536.

246. Tournoy KG, Keller SM, Annema JT. Mediastinal staging of lung cancer: novel concepts. *Lancet oncology.* 2012;13(5):e221-229.

247. Schmidt-Hansen M, Baldwin DR, Hasler E, Zamora J, Abraira V, Roque IFM. PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer. *The Cochrane database of systematic reviews.* 2014;11:CD009519.

248. Silvestri GA, Gonzalez AV, Jantz MA, et al. Methods for staging non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest.* 2013;143(5 Suppl):e211S-250S.

249. De Leyn P, Dooms C, Kuzdzal J, et al. Revised ESTS guidelines for preoperative mediastinal lymph node staging for non-small-cell lung cancer. *European journal of cardiothoracic surgery.* 2014;45(5):787-798.

250. Vilmann P, Clementsen PF, Colella S, et al. Combined endobronchial and oesophageal endosonography for the diagnosis and staging of lung cancer. European Society of Gastrointestinal Endoscopy (ESGE) Guideline, in cooperation with the European Respiratory Society (ERS) and the European Society of Thoracic Surgeons (ESTS). *European respiratory journal.* 2015;46(1):40-60.

251. Navani N, Nankivell M, Lawrence DR, et al. Lung cancer diagnosis and staging with endobronchial ultrasound-guided transbronchial needle aspiration compared with conventional approaches: an open-label, pragmatic, randomised controlled trial. *Lancet respiratory medicine.* 2015;3(4):282-289.

252. Dietrich CF, Annema JT, Clementsen P, Cui XW, Borst MM, Jenssen C. Ultrasound techniques in the evaluation of the mediastinum, part I: endoscopic ultrasound (EUS), endobronchial ultrasound (EBUS) and transcutaneous mediastinal ultrasound (TMUS), introduction into ultrasound techniques. *Journal of thoracic disease.* 2015;7(9):E311-325.

253. Micames CG, McCrory DC, Pavey DA, Jowell PS, Gress FG. Endoscopic ultrasound-guided fine-needle aspiration for non-small cell lung cancer staging: A systematic review and metaanalysis. *Chest.* 2007;131(2):539-548.

254. Gu P, Zhao YZ, Jiang LY, Zhang W, Xin Y, Han BH. Endobronchial ultrasound-guided transbronchial needle aspiration for staging of lung cancer: a systematic review and meta-analysis. *European journal of cancer.* 2009;45(8):1389-1396.

255. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ.* 2011;342:d549.

256. Vilmann P, Krasnik M, Larsen SS, Jacobsen GK, Clementsen P. Transesophageal endoscopic ultrasound-guided fine-needle aspiration (EUS-FNA) and endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) biopsy: A combined approach in the evaluation of mediastinal lesions. *Endoscopy.* 2005;37(9):833-839.

257. Wallace MB, Pascual JMS, Raimondo M, et al. Minimally invasive endoscopic staging of suspected lung cancer. *JAMA.* 2008;299(5):540-546.

258. Annema JT, van Meerbeeck JP, Rintoul RC, et al. Mediastinoscopy vs endosonography for mediastinal nodal staging of lung cancer: a randomized trial. *JAMA.* 2010;304(20):2245-2252.

259. Herth FJ, Krasnik M, Kahn N, Eberhardt R, Ernst A. Combined endoscopic-endobronchial ultrasound-guided fine-needle aspiration of mediastinal lymph nodes through a single bronchoscope in 150 patients with suspected lung cancer. *Chest.* 2010;138(4):790-794.

260. Hwangbo B, Lee GK, Lim KY, et al. Transbronchial and transesophageal fine-needle aspiration using an ultrasound bronchoscope in mediastinal staging of potentially operable lung cancer. *Chest.* 2010;138(4):795-802.

261. Szlubowski A, Zielinski M, Soja J, et al. A combined approach of endobronchial and endoscopic ultrasound-guided needle aspiration in the radiologically normal mediastinum in non-small-cell lung cancer staging--a prospective trial. *European journal of cardio-thoracic surgery.* 2010;37(5):1175-1179.

262. Ohnishi R, Yasuda I, Kato T, et al. Combined endobronchial and endoscopic ultrasound-guided fine needle aspiration for mediastinal nodal staging of lung cancer. *Endoscopy.* 2011;43(12):1082-1089.

263. Szlubowski A, Soja J, Kocon P, et al. A comparison of the combined ultrasound of the mediastinum by use of a single ultrasound bronchoscope versus ultrasound bronchoscope plus ultrasound gastroscope in lung cancer staging: a prospective trial. *Interactive cardiovascular and thoracic surgery.* 2012;15(3):442-446; discussion 446.

264. Kang HJ, Hwangbo B, Lee GK, et al. EBUS-centred versus EUS-centred mediastinal staging in lung cancer: a randomised controlled trial. *Thorax.* 2014;69(3):261-268.

265. Liberman M, Sampalis J, Duranceau A, Thiffault V, Hadjeres R, Ferraro P. Endosonographic mediastinal lymph node staging of lung cancer. *Chest.* 2014;146(2):389-397.

266. Oki M, Saka H, Ando M, Kitagawa C, Kogure Y, Seki Y. Endoscopic ultrasound-guided fine needle aspiration and endobronchial ultrasound-guided transbronchial needle aspiration: Are two better than one in mediastinal staging of non-small cell lung cancer? *Journal of thoracic and cardiovascular surgery.* 2014;148(4):1169-1177.

267. Crombag L, Stigt J, Dooms C, et al. Added value of EUS-B to EBUS for lung cancer staging. *European respiratory journal.* 2015;46(s59).

268. Hauer J, Szlubowski A, Zanowska K, et al. Minimally invasive strategy for mediastinal staging of patients with lung cancer. *Polskie archiwum medycyny wewnetrzne.* 2015;125(12):910-913.

269. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of internal medicine.* 2013;158(7):544-554.

270. Naaktgeboren CA, van Enst WA, Ochodo EA, et al. Systematic overview finds variation in approaches to investigating and reporting on sources of heterogeneity in systematic reviews of diagnostic studies. *Journal of clinical epidemiology.* 2014;67(11):1200-1209.

271. Dooms C, Tournoy KG, Schuurbiers O, et al. Endosonography for mediastinal nodal staging of clinical N1 non-small cell lung cancer: a prospective multicenter study. *Chest.* 2015;147(1):209-215.

272. Von Bartheld MB, Van Breda A, Annema JT. Complication rate of endosonography (endobronchial and endoscopic ultrasound): A systematic review. *Respiration.* 2014;87(4):343-351.

273. Konge L, Colella S, Vilmann P, Clementsen PF. How to learn and to perform endoscopic ultrasound and endobronchial ultrasound for lung cancer staging: A structured guide and review. *Endoscopic ultrasound.* 2015;4(1):4-9.

274. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ.* 2008;336(7653):1106-1110.

275. Rintoul RC, Glover MJ, Jackson C, et al. Cost effectiveness of endosonography versus surgical staging in potentially resectable lung cancer: a health economics analysis of the ASTER trial from a European perspective. *Thorax.* 2014;69(7):679-681.

276. Chakma J, Sun GH, Steinberg JD, Sammut SM, Jagsi R. Asia's ascent--global trends in biomedical R&D expenditures. *New England journal of medicine.* 2014;370(1):3-6.

277. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet.* 2009;374(9683):86-89.

278. Macleod MR, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. *Lancet.* 2014;383(9912):101-104.

279. Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *Lancet.* 2014;383(9912):156-165.

280. Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet.* 2014;383(9912):166-175.

281. Al-Shahi Salman R, Beller E, Kagan J, et al. Increasing value and reducing waste in biomedical research regulation and management. *Lancet.* 2014;383(9912):176-185.

282. Altman DG, Simera I. A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. *Journal of the Royal Society of Medicine.* 2016;109(2):67-77.

Abbreviations

95%CI	95% Confidence interval
95%PI	95% Prediction interval
ACQ	Asthma control questionnaire
AMD	Adjusted mean difference
AMSTAR	Assessment of multiple systematic reviews
ANOVA	Analysis of variance
ARVO	Association for research in vision and ophthalmology
ASTER	Assessment of surgical staging versus endosonographic ultrasound in lung cancer: a randomized clinical trial
AUC	Area under the receiver operating characteristic curve
BMI	Body mass index
CONSORT	Consolidated standards of reporting trials
COPD	Chronic obstructive pulmonary disease
CT	Computed tomography
DOR	Diagnostic odds ratio
EBUS	Endobronchial ultrasound
EQUATOR	Enhancing the quality and transparency of health research
EUS	Transesophageal endoscopic ultrasound
EUS-B	Transesophageal endoscopic ultrasound with a bronchoscope
FDA	Food and drug administration
FeNO	Fraction of exhaled nitric oxide
FEV_1	Forced expiratory volume in 1 s
FVC	Forced vital capacity
FN	False negative
FP	False positive
HIV	Human immunodeficiency virus

HR	Hazard ratio
ICMJE	International committee of medical journal editors
ICS	Inhaled corticosteroid
IgE	Immunoglobulin E
IQR	Interquartile range
IRB	Institutional review board
NIH	National institues of health
NPV	Negative predictive value
NSCLC	Non-small cell lung cancer
PET-CT	Positron emission tomography-computed tomography
ppb	Parts per billion
PPV	Positive predictive value
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
QUADAS	Quality assessment of diagnostic accuracy studies
RCT	Randomised controlled trial
ROC	Receiver operating characteristic curve
SD	Standard deviation
STARD	Standards for reporting of diagnostic accuracy studies
Th	T-helper
TN	True negative
TP	True positive

Contributing authors

Douglas G. Altman
Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.

Marijke Amelink
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Jouke T. Annema
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Elisabeth H. Bel
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Patrick M. Bossuyt
Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Anneke ten Brinke
Department of Respiratory Medicine, Medical Center Leeuwarden, Leeuwarden, the Netherlands.

David E. Bruns
Department of Pathology, University of Virginia School of Medicine, Charlottesville, Virginia, USA.

Jérémie F. Cohen
Inserm UMR 1153, Obstetrical, Perinatal and Pediatric Epidemiology Research Team, Center for Epidemiology and Statistics Sorbonne Paris Cité, Paris Descartes University, Paris, France.

Laurence Crombach
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Kay Dickersin
Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.

Christophe Dooms
Department of Respiratory Medicine, Leuven University Hospitals, Leuven, Belgium.

W. Annefloor van Enst
Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Nick van Es
Department of Vascular Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Constantine A. Gatsonis
Center for Statistical Sciences, Brown University School of Public Health, Providence, Rhode Island, USA.

Paul P. Glasziou
Centre for Research in Evidence-Based Practice, Faculty of Health Sciences and Medicine, Bond University, Gold Coast, Queensland, Australia.

Jantina C. de Groot
Department of Respiratory Medicine, Medical Center Leeuwarden, Leeuwarden, the Netherlands.

Lotty Hooft
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.

Les Irwig
Screening and Diagnostic Test Evaluation Program, School of Public Health, University of Sydney, Sydney, New South Wales, Australia.

Jolanda C. Kuijvenhoven
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Mariska M. Leeflang
Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Thomas L. Malfait
Department of Respiratory Medicine, University Hospital Ghent, Ghent, Belgium.

David Moher
Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada.

Nina Molenaar
Department of Psychiatry, Erasmus University, Rotterdam, the Netherlands.

Selma B. de Nijs
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Eleanor A. Ochodo
Centre for Evidence-based Health Care, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

Robert C. Rintoul
Department of Thoracic Oncology, Papworth Hospital, Cambridge, United Kingdom.

Johannes B. Reitsma
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.

Maurice W. de Ronde
Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Ian J. Saldanha
Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.

Rob J. Scholten
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.

Nynke Smidt
Department of Epidemiology, University Medical Center Groningen, Groningen, the Netherlands.

René Spijker
Medical Library, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Peter J. Sterk
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Kurt G. Tournoy
Department of Respiratory Medicine, Onze-Lieve-Vrouw Hospital, Aalst, Belgium.

Henrica C. de Vet
Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands.

Gianni Virgili
Department of Translational Surgery and Medicine, Eye Clinic, University of Florence, Florence, Italy.

Junfeng Wang
Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Els J. Weersink
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Guus A. Westerhof
Department of Respiratory Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Aeilko H. Zwinderman
Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

PhD portfolio

Courses at AMC Graduate School	Year
Advanced topics in biostatistics	2014
Genetic epidemiology	2014
Randomized clinical trials	2014
Computing in R	2013
Practical biostatistics	2013
Advanced topics in clinical epidemiology	2013
Clinical epidemiology	2013
Systematic reviews	2013
Expert management of medical literature	2013
Clinical data management	2013

Courses at other institutions	Year
Advanced topics in decision making in medicine *Erasmus Winter Programme, Erasmus University, Rotterdam*	2016
Basic course organization clinical scientific research *Netherlands Federation of University Medical Centres, Amsterdam*	2016
Interventional endosonography for lung diseases *European Respiratory Society, Amsterdam*	2015
Advanced analysis of prognostic studies *Erasmus Winter Programme, Erasmus University, Rotterdam*	2015
Advanced diagnostic research *Julius Center, University Utrecht*	2014
Grading scientific evidence and recommendations *GRADE Working Group, Utrecht*	2014
Reporting guidelines *EQUATOR Network, Chicago*	2013
Statistical methods for diagnostic accuracy reviews *UK Cochrane Centre, Birmingham*	2013
Evidence-based medicine in clinical practice *Dutch Cochrane Centre, Amsterdam*	2013

Courses at Graduate Summer Institute of Epidemiology and Biostatistics, Johns Hopkins University, Baltimore	Year
Advanced methods in observational studies: inference	2014
Advanced methods in observational studies: design	2014
Clinical trials: issues and controversies	2014
Critical reading of epidemiologic literature	2014
Conducting epidemiological research	2014
Methods for clinical and translational research	2014

(Inter)national conferences and symposia	Year
Cochrane Colloquium *Seoul*	2016
Netherlands Research Integrity Network Conference *Amsterdam*	2016
Cochrane Colloquium *Vienna*	2015
European Respiratory Society International Congress *Amsterdam*	2015
European Congress of Epidemiology *Maastricht*	2015
Respiratory Society Young Investigator Symposium *Amsterdam*	2014
Improving Scientific Practice: Dealing with the Human Factors *Amsterdam*	2014
European Respiratory Society International Congress *Munich*	2014
Improving Reporting to Decrease the Waste of Research *Paris*	2014
International Congress on Peer Review and Biomedical Publication *Chicago*	2013
Methods for Evaluating Medical Tests and Biomarkers *Birmingham*	2013

Visiting researcher fellowships	Year
Knowledge Synthesis Group, Ottawa Hospital Research Institute *University of Ottawa, Ottawa*	2015
Department of Epidemiology, Bloomberg School of Public Health *Johns Hopkins University, Baltimore*	2014

Teaching	Year
Critical appraisal of published medical literature *Amsterdam Medical Student convention, Amsterdam*	2016
Critical appraisal of a randomised controlled trial *Academic Medical Center, Amsterdam*	2014-2016
Clinical and scientific methodology *Academic Medical Center, Amsterdam*	2013-2016

Peer-reviewing	Year
At least one review for the following journals:	2014-2016
BMC Medical Research Methodology	
BMJ	
Clinical and Experimental Allergy	
Current Medical Research and Opinion	
European Congress of Epidemiology	
JAMA	
Journal of Clinical Epidemiology	
Journal of Clinical Monitoring and Computing	
PLoS Neglected Tropical Diseases	
PLoS One	
Research and Reports in Tropical Medicine	
Respirology	
Scientific Reports	
Systematic Reviews	
Ultrasound in Obstetrics and Gynecology	

Awards	Year
Louise Gunning public health study fund (AMC young talent fund)	2014

Presentations

Korevaar DA, van Es N, Zwinderman AH, Cohen JF, Bossuyt PM. Time to publication among completed diagnostic accuracy studies: associated with reported accuracy estimates. *Cochrane Colloquium*, Seoul, October 24, 2016. *Oral presentation*.

McGrath TA, McInnes MD, Langer FW, Hong J, **Korevaar DA**, Bossuyt PM. Treatment of multiple test readers in diagnostic accuracy systematic reviews of imaging studies. Cochrane Colloquium, Seoul, October 23, 2016. *Poster presentation*.

Korevaar DA, Cohen JF, Askie LM, Faure H, Gatsonis CA, Hunter KE, Kressel HY, McInnes MD, Moher D, Rifai N, Hooft L, Bossuyt PM. STARD for Registration: establishing guidance on where and how to prospectively register diagnostic accuracy studies. *Cochrane Colloquium*, Seoul, October 23, 2016. *Poster presentation*.

Korevaar DA, Cohen JF, Spijker R, Saldanha IJ, Dickersin K, Virgili G, Hooft L, Bossuyt PM. Reported estimates of diagnostic accuracy in ophthalmology conference abstracts are not associated with full-text publication. *Cochrane Colloquium*, Seoul, October 23, 2016. *Poster presentation*.

Korevaar DA. The impact of 10 years of STARD on the reporting of diagnostic accuracy studies. *Netherlands Research Integrity Network Conference*, Amsterdam, May 25, 2016. *Oral presentation*.

Korevaar DA. Pearl of laboratory medicine: optimal reporting of diagnostic accuracy studies. *American Association for Clinical Chemistry Trainee Council*, February 11, 2016. *Online oral presentation*.

Korevaar DA, Westerhof GA, Wang J, Cohen JF, Spijker R, Sterk PJ, Bel EH, Bossuyt PM. Including unpublished data in a systematic review of diagnostic accuracy studies: impact on summary estimates (long oral presentation). *Cochrane Colloquium*, Vienna, October 7, 2015. *Oral presentation*.

Cohen JC, **Korevaar DA**, Wang J, Leeflang MM, Geskes R, Bossuyt PM. Time trends in summary estimates from meta-analyses of diagnostic accuracy studies. *Cochrane Colloquium*, Vienna, October 7, 2015. *Oral presentation*.

Korevaar DA, Cohen JF, de Ronde MW, Virgili G, Dickersin K, Bossuyt PM. Journal and conference abstracts of diagnostic accuracy studies: sufficiently informative? *Cochrane Colloquium*, Vienna, October 4, 2015. *Poster presentation*.

Korevaar DA, Crombag L, Cohen JF, Spijker R, Bossuyt PM, Annema JT. Combined endobronchial and esophageal endosonography for mediastinal staging in lung cancer: systematic review and meta-analysis. *European Respiratory Society*, Amsterdam, September 28, 2015. *Poster discussion*.

Korevaar DA, Hooft L, Cohen JF, Bossuyt PM. Complete and accurate reporting of studies of diagnostic accuracy: updating the STARD statement. *European Congress of Epidemiology*, Maastricht, June 27, 2015. *Oral presentation*.

Korevaar DA, Westerhof GA, Wang J, Cohen JF, Spijker R, Sterk PJ, Bel EH, Bossuyt PM. Including unpublished results in a systematic review of diagnostic accuracy studies: impact on summary estimates? *European Congress of Epidemiology*, Maastricht, June 27, 2015. *Poster presentation*.

Korevaar DA, Cohen JF, Hooft L, Bossuyt PM. Reporting weaknesses in journal abstracts of diagnostic accuracy studies. *European Congress of Epidemiology*, Maastricht, June 26, 2015. *Poster presentation*.

Korevaar DA, Westerhof GA, Wang J, Spijker R, Sterk PJ, Bel EH, Bossuyt PM. Diagnostic accuracy of markers for detection of airway eosinophilia in asthma: a systematic review and meta-analysis. *Netherlands Respiratory Society young investigator symposium*, Amsterdam, November 14, 2014. *Oral presentation*.

Korevaar DA, Ochodo EA, Bossuyt PM, Hooft L. Failure to publish and selective reporting: also prevalent among test accuracy studies? *Improving scientific practice: dealing with the human factors*, Amsterdam, September 11, 2014. *Poster presentation*.

Korevaar DA, Westerhof GA, Spijker R, Sterk PJ, Bel EH, Bossuyt PM. Diagnostic accuracy of markers for detection of airway eosinophilia in asthma: a systematic review. *European Respiratory Society*, Munich, September 6, 2014. *Poster presentation*.

Korevaar DA, Hooft L, Bossuyt PM. Registration of studies quantifying the accuracy of diagnostic tests and markers. *Methods for evaluating medical tests and biomarkers*, Birmingham, July 15, 2013. *Poster presentation*.

Korevaar DA, Hooft L, ter Riet G. Systematic reviews of laboratory animal experiments. *New paradigms in laboratory animal science*, Helsinki, June 15, 2010. *Oral presentation*.

Publications

Publications included in this thesis

Korevaar DA, Crombag LM, Cohen JF, Spijker R, Bossuyt PM, Annema JT. Added value of combined endobronchial and esophageal endosonography for mediastinal nodal staging in lung cancer: a systematic review and meta-analysis. *Lancet Respiratory Medicine* 2016; In press.

Korevaar DA, Cohen JF, Spijker R, Saldanha IJ, Dickersin K, Virgili G, Hooft L, Bossuyt PM. Full publication of conference abstracts describing diagnostic accuracy studies in ophthalmology: no observed association with reported accuracy estimates. *Journal of Clinical Epidemiology* 2016; E-published ahead of print.

Kuijvenhoven JC, **Korevaar DA**, Tournoy KG, Malfait TL, Dooms C, Rintoul RC, Annema JT. Five-year survival after endosonography versus mediastinoscopy for mediastinal nodal staging of lung cancer. *JAMA* 2016;316(10):1110-2.

Korevaar DA, Cohen JF, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Moher D, de Vet HC, Altman DG, Hooft L, Bossuyt PM. Updating standards for reporting diagnostic accuracy: the development of STARD 2015. *Research Integrity and Peer Review* 2016;1:7.

Korevaar DA, van Es N, Zwinderman AH, Cohen JF, Bossuyt PM. Time to publication among completed diagnostic accuracy studies: associated with reported accuracy estimates. *BMC Medical Research Methodology* 2016;16(1):68.

Korevaar DA, Cohen JF, de Ronde MW, Virgili G, Dickersin K, Bossuyt PM. Reporting weaknesses in conference abstracts of diagnostic accuracy studies in ophthalmology. *JAMA Ophthalmology* 2015;133(12):1464-7.

Korevaar DA, Cohen JF, Hooft L, Bossuyt PM. Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies. *Journal of Clinical Epidemiology* 2015;68(6):708-15.

Korevaar DA, Westerhof GA, Wang J, Cohen JF, Spijker R, Sterk PJ, Bel EH, Bossuyt PM. Diagnostic accuracy of minimally invasive markers for detection of airway eosinophilia in asthma: a systematic review and meta-analysis. *Lancet Respiratory Medicine* 2015;3(4):290-300.

Westerhof GA, **Korevaar DA**, Amelink M, de Nijs SB, de Groot JC, Wang J, Weersink EJ, ten Brinke A, Bossuyt PM, Bel EH. Biomarkers to identify sputum eosinophilia in different adult asthma phenotypes. *European Respiratory Journal* 2015;46(3):688-96.

Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, Bossuyt PM. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology* 2015;274(3):781-9.

Hooft L, **Korevaar DA**, Molenaar N, Bossuyt PM, Scholten RJ. Endorsement of ICMJE's clinical trial registration policy: a survey among journal editors. *Netherlands Journal of Medicine* 2014;72(7):349-55.

Korevaar DA, Ochodo EA, Bossuyt PM, Hooft L. Publication and reporting of test accuracy studies registered in ClinicalTrials.gov. *Clinical Chemistry* 2014;60(4):651-9.

Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evidence-Based Medicine* 2014;19(2):47-54*.*

Korevaar DA, Bossuyt PM, Hooft L. Infrequent and incomplete registration of test accuracy studies: analysis of recent study reports. *BMJ Open* 2014;4(1):e004596.

Publications not included in this thesis

Korevaar DA, Cohen JF, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HC, Bossuyt PM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016; In press.

Cohen JF, **Korevaar DA**, Wang J, Leeflang MM, Bossuyt PM. Meta-epidemiological study showed frequent time trends in summary estimates from meta-analyses of diagnostic accuracy studies. *Journal of Clinical Epidemiology* 2016; E-published ahead of print.

McGrath TA, McInnes MD, **Korevaar DA**, Bossuyt PM. Meta-analyses of diagnostic accuracy in imaging journals: analysis of pooling techniques and their effect on summary estimates of diagnostic accuracy. *Radiology* 2016;281(1):78-85.

Korevaar DA, Bossuyt PM. STARD 2015 voor de evaluatie van diagnostische tests / STARD 2015 for the evaluation of diagnostic tests. *Nederlands Tijdschrift voor Geneeskunde* 2016;160(0):D113 (invited commentary).

Korevaar DA, Westerhof GA, Bel, EH. Biomarkers for diagnosing asthma: a smoking gun? *Clinical and Experimental Allergy* 2016;46(4):516-8 (invited editorial).

Bossuyt PM, Cohen JF, Gatsonis CA, **Korevaar DA**. STARD 2015: updated reporting guidelines for all diagnostic accuracy studies. *Annals of Translational Medicine* 2016;4(4):85 (letter to the editor).

Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PM, **Korevaar DA**, Graham ID, Ravaud P, Boutron I. Increasing value and reducing waste in biomedical research: who's listening? *Lancet* 2016;387(10027):1573-86.

Roth JM, **Korevaar DA**, Leeflang MM, Mens PF. Molecular malaria diagnostics: a systematic review and meta-analysis. *Critical Reviews in Clinical Laboratory Sciences* 2016;53(2):87-105.

Dilauro M, McInnes MD, **Korevaar DA**, van der Pol CB, Petrcich W, Walther S, Quon J, Kurowecki D, Bossuyt PM. Is there an association between STARD statement adherence and citation rate? *Radiology* 2016;280(1):62-7.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HC, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, **Korevaar DA**, Cohen JF. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h552, *Clinical Chemistry* 2015;61(12):1446-52, and *Radiology* 2015;277(3):826-32.

Cohen JF, Chalumeau M, Cohen R, **Korevaar DA**, Khoshnood B, Bossuyt PM. Cochran's *Q* test was useful to assess heterogeneity in likelihood ratios in studies of diagnostic accuracy. *Journal of Clinical Epidemiology* 2015;68(3):299-306.

Vilmann P, Clementsen PF, Colella S, Siemsen M, De Leyn P, Dumonceau JM, Herth FJ, Larghi A, Vasquez-Sequeiros E, Hassan C, Crombag L, **Korevaar DA**, Kong L, Annema JT. Combined endobronchial and oesophageal endosonography for the diagnosis and staging of lung canger: European Society of Gastrointestinal Endoscopy (ESGE) Guideline, in cooperation with the European Respiratory Society (ERS) and the European Society of Thoracic Surgeons (ESTS). *European Journal of Cardiothoracic Surgery* 2015;48(1):1-15, *European Respiratory Journal* 2015;46(1):40-60, and *Endoscopy* 2015;47(6):545-59.

Cohen JF, **Korevaar DA**, Wang J, Spijker R, Bossuyt PM. Should we search Chinese biomedical databases when performing systematic reviews? *Systematic Reviews* 2015;4(1):23.

Korevaar DA, Hooft, L. Het gevaar van ongepubliceerd onderzoeksresultaten / The danger of unpublished trial results. *Nederlands Tijdschrift voor Geneeskunde* 2014;158:A7400.

Visser BJ, **Korevaar DA**, Nolan T. Mobile medical apps: dangers and potential solutions. *Journal of Telemedicine and Telecare* 2013;19(4):229-230.

Korevaar DA, Visser BJ. Reviewing the evidence on nodding syndrome, a mysterious tropical disorder. *International Journal of Infectious Diseases* 2013;17(3):e149-52.

Visser BJ, **Korevaar DA**. Clinical involvement and transparency in medical apps. *Colorectal Disease* 2013;15(1):121-2 (letter to the editor).

Ter Riet G, **Korevaar DA**, Leenaars M, Sterk PJ, Van Noorden CJF, Bouter LM, Lutter R, Elferink RP, Hooft L. Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. *PLoS One* 2012;7(9):e43404.

Korevaar DA, Visser BJ. A worm emerging from the foot. *Netherlands Journal of Medicine* 2012;70(8):375-76.

Korevaar DA, Visser BJ. Podoconiosis, a neglected tropical disease. *Netherlands Journal of Medicine* 2012;70(5):210-4.

Visser BJ, Huiskens F, **Korevaar DA.** A social media self-evaluation checklist for medical practitioners. *Indian Journal of Medical Ethics* 2012;9(4):245-8.

Visser BJ, **Korevaar DA**, van der Zee J. A 24-year old Ethiopian farmer with burning feet. *American Journal of Tropcial Medicine and Hygiene* 2012;87(4):583.

Korevaar DA, Hooft L, Ter Riet G. Systematic reviews and meta-analyses of preclinical studies: publication bias in laboratory animal experiments. *Laboratory Animals* 2011;45(4):225-30.

Acknowledgments

Graag wil ik de volgende personen hartelijk bedanken voor hun begeleiding, samenwerking, hulp of steun bij het tot stand komen van dit proefschrift.

Patrick Bossuyt en Lotty Hooft. Beste Patrick, je bood me veel vrijheid om mijn eigen ideeën uit te werken, maar gaf indien nodig ook duidelijk richting aan. Dit schiep voor mij een ideaal werkklimaat. Ik keek altijd erg uit naar onze wekelijkse afspraken, waarin we op ontspannen en efficiënte wijze alle lopende projecten in hoog tempo doornamen. Ik bewonder de zorgvuldigheid, snelheid en creativiteit waarmee je werkt. Beste Lotty, de enthousiaste en hartelijke manier waarmee je me vanaf mijn wetenschappelijke stage begeleid hebt was een belangrijke reden om voor dit promotietraject te solliciteren. Je zag zelden beren op de weg als ik weer eens een nieuw project voorstelde, en voor eventuele problemen onderweg had je altijd onmiddellijk een praktische oplossing. Beste promotoren, ik heb genoten van onze samenwerking en hoop dat we in de toekomst mogelijkheden zullen vinden om deze voort te zetten.

Jérémie Cohen. Dear Jérémie, as co-author on half of the chapters, you made an extremely important contribution to the development of this thesis. It was exciting and motivating to collaborate so closely and intensively with you during the two years you spent in Amsterdam. You have been a great colleague and friend.

René Spijker. Beste René, met veel plezier ging ik regelmatig bij je langs om onze projecten te bespreken. Meestal duurde deze afspraken veel langer dan gepland, omdat ze uitmondden in een uitwisseling van interessante artikelen en ideeën voor onderzoek. Bedankt voor je bijdrage aan een flink aantal hoofdstukken.

Collega's van de Afdeling Longziekten, Laurence Crombag, Jolanda Kuijvenhoven, Guus Westerhof, Liesbeth Bel, Jouke Annema en Peter Sterk. Vanuit mijn interesse voor longziekten en wens om longarts te worden zocht ik al in een vroeg stadium van dit promotietraject de samenwerking met jullie op. Zonder aarzelen zijn jullie hierop ingegaan. Ik ben trots op de mooie projecten die we hebben uitgevoerd en kijk uit naar onze toekomstige samenwerking, zowel klinisch als wetenschappelijk.

Leden van de BiTE group. Wekelijks bespraken we onze onderzoeksprojecten. Ik vond deze ontmoetingen erg plezierig en leerzaam, en zal ze missen.

Mijn kamergenoten op de KEBB, Mareen Datema, Annefloor van Enst, Nina Steutel en Maurice de Ronde. Naast fijne collega's zijn we ook goede vrienden geworden. Hopelijk heeft de toekomst nog veel mooie gezamenlijke borrels, reisjes en jeu de bouleswedstrijden voor ons in het verschiet.

Corien Meijer. Beste Corien, je was altijd bereid om direct en op heel ontspannen wijze een helpende hand te bieden. Heel hartelijk bedankt hiervoor.

Colleagues at Johns Hopkins University, Kay Dickersin, Kristina Lindsley, Tianjing Li, and Ian Saldanha. I enjoyed my stay in Baltimore a lot. I am impressed by your thorough way of working, and the warm atmosphere at your department. The highlight was obviously being introduced to the American way of camping!

Colleagues at the University of Ottawa, David Moher and Matthew McInnes. Dear David, your devotion to improving research made my stay at the Knowledge Synthesis Group very inspiring. Dear Matt, not checking my email for one day means missing, on average, 10 messages from you or your students. Thank you for introducing me to this energetic and productive way of collaborating; I love it!

Koos Zwinderman. Beste Koos, de vriendelijke en ontspannen manier waarop je de afdeling leidt, en de laagdrempeligheid waarmee je iedereen altijd bereid bent te woord te staan, maken het een feest om naar het werk te gaan.

Gerben ter Riet. Beste Gerben, je hebt me tijdens mijn wetenschappelijke stage enthousiast gemaakt voor onderzoek. Ik denk met veel plezier terug aan deze inspirerende periode. Ik vond het erg leuk om het de afgelopen jaren met enige regelmaat met je te hebben over bepaalde problemen in de wetenschap, en hoop dat onze wegen elkaar in de toekomst zullen blijven kruisen

Mijn vrienden. Zonder al die hardloop-, cineville-, barbecue-, risk-, koehandel-, pubquiz- en wetenschapsmiddagen en -avonden was het uiteraard niks geworden. Vormt voldoende ontspanning immers niet de basis voor succes?

Tim Badir en Jelle Visser. De paranimfen dienen het proefschrift te verdedigen in het geval de promovendus hier zelf niet toe in staat is. Bedankt dat jullie bereid zijn mij, zoals altijd, ook nu weer in woord en daad bij te staan.

Jan Korevaar, Carla Schakel, Niels Korevaar, Yara Korevaar en Jon Thissen. Lieve familie, jullie zijn met Tessa mijn voornaamste raadgevers als het gaat om belangrijke of moeilijke beslissingen. Bedankt dat jullie me altijd steunen in mijn - soms wat doorgeschoten - passies en ambities. Jullie enthousiaste betrokkenheid is erg belangrijk voor me.

Tessa van Ligten. Lieve Tessa, regelmatig heb ik tijdens dit promotieproject een beroep gedaan op je heldere manier van denken en formuleren, en je haarfijne analyses van sociale contexten en interacties. Ik ben blij en trots dat je vlak bij me zit, tijdens de verdediging van mijn proefschrift. Ik kijk uit naar de nieuwe avonturen die we tegemoet gaan. Bedankt dat je er altijd voor me bent.

Curriculum vitae

Daniël A. Korevaar was born in Amsterdam on November 8, 1986. After obtaining his pre-university education diploma at the Sint Ignatius Gymnasium in 2005, he went to study Medicine at the Academic Medical Center, University of Amsterdam. He combined his studies with road race cycling at Elite level, and participated in (semi-)professional races all over the world. In the context of his Master's degree (cum laude), he did his scientific internship at the Department of General Practice, in which he aimed to map the extent of publication bias in preclinical animal research. After finishing his thesis in 2009, he worked as a research officer at the same department for one year. During this period he developed a special interest for the critical appraisal of scientific research, and for identifying hurdles in the translation of scientific research to clinical applications. After obtaining his Medical Doctor degree in 2012, he started working on the studies described in this PhD thesis at the Department of Clinical Epidemiology, Biostatistics and Bioinformatics at the University of Amsterdam. During his PhD, he was a member of the four-member Project Team that coordinated the update of the Standards for Reporting of Diagnostic Accuracy Studies (STARD), a process that involved 85 scientists from all over the world. He also did visiting researcher fellowships at the Bloomberg School of Public Health at Johns Hopkins University in Baltimore in the USA in 2014, and at the Ottawa Hospital Research Institute at the University of Ottawa in Canada in 2015, and followed a training program as Scientific Researcher Epidemiologist with the Netherlands Epidemiological Society. He currently is a member of the Young Dutch Medicines Evaluation Board. In December 2016, he will start his residency in respiratory medicine.